Regularized deep learning models for multi-anatomy segmentation in pediatric imaging

Arnaud Boutillon

DOCTORAT BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPÉRIEURE
MINES-TÉLÉCOM ATLANTIQUE BRETAGNE
PAYS-DE-LA-LOIRE - IMT ATLANTIQUE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, Image, Vision*

Par
Arnaud BOUTILLON

Regularized deep learning models for multi-anatomy segmentation in pediatric imaging

Thèse présentée et soutenue à Brest, le 14 novembre 2022
Unité de recherche : LaTIM, UMR 1101, Inserm
Thèse N° : 2022IMTA0311

Rapporteurs avant soutenance :

Hervé DELINGETTE	Directeur de Recherche INRIA, Projet EPIONE
Carole LARTIZIEN	Directeur de Recherche CNRS, CREATIS Lyon

Composition du Jury :

Président :	Hervé DELINGETTE	Directeur de Recherche INRIA, Projet EPIONE
Examinateurs :	Carole LARTIZIEN	Directeur de Recherche CNRS, CREATIS Lyon
	Maria DEPREZ	Senior Lecturer, King's College London
Dir. de thèse :	Valérie BURDIN	Professeure, IMT Atlantique
Co-enc. de thèse :	Pierre-Henri CONZE	Maître de Conférences, IMT Atlantique
	Bhushan BOROTIKAR	Associate Professor, SCMIA, Symbiosis International University

Invité :

Douraied BEN SALEM PU-PH, CHU Brest, Université de Bretagne Occidentale

ACKNOWLEDGMENTS

Remerciements

During the three years of this thesis, I had the chance to be surrounded and supported by numerous people to whom I would like to express my sincere gratitude.

First, I am deeply indebted to my thesis director, Prof. Valérie Burdin, for offering me the opportunity to conduct this research project and for the confidence she has placed in me. I appreciate her consistent support and encouragement throughout my thesis. I am also thankful to my thesis co-supervisors, Prof. Pierre-Henri Conze and Prof. Bhushan Borotikar, for the quality of their supervision, which allowed me to work in excellent conditions. I could not have undertaken this journey without their remarkable guidance, their great availability, their insightful corrections on the manuscript, and all our enriching discussions.

I would like to offer my special thanks to the members of my thesis committee for agreeing to participate in the evaluation of my work. In particular, I thank the reviewers, Prof. Carole Lartizien, and Prof. Hervé Delingette, for their comments on the manuscript, and I thank Prof. Maria Deprez for taking part in the thesis committee as an examiner.

My gratitude extends to all the people involved in the pediatric image acquisition process, especially Dr. Christelle Pons from the Ildys Foundation, for her involvement in shoulder data collection and the enrollment of OBPP patients. I also thank the patients and their families who agreed to participate in the clinical studies and without whom the research conducted in this thesis would not have been possible.

I would like to express my genuine appreciation to my colleagues and friends at IMT Atlantique and LaTIM for their professional support and promotion of a stimulating work environment. I also feel indebted to the communities behind the multiple open-source software packages on which this research project is built, as well as the anonymous reviewers whose critical comments helped improve the publications resulting from this thesis.

Lastly, I would like to mention that the research work conducted in this thesis was funded by IMT, Fondation Mines-Télécom, and Institut Carnot TSN through the Futur

& Ruptures program. I am therefore thankful to the generous sponsors (companies or individuals) for their financial contributions, without which the Futur & Ruptures program could not have been funded.

Enfin, je tiens à exprimer mes sincères remerciements à mes parents. Ce travail n'aurait pas été possible sans leur soutien et leur confiance indéfectibles durant toutes ces années d'études. Un grand merci également à toute ma famille, mes grands-mères, oncles et tantes, cousins et cousines.

RÉSUMÉ EN FRANÇAIS

Modèles d'apprentissage profond régularisés pour la segmentation multi-anatomie en imagerie pédiatrique

---

Contexte

Dans la pratique clinique, l'imagerie médicale est un outil précieux pour aider les cliniciens à diagnostiquer des pathologies, évaluer le suivi des traitements thérapeutiques et planifier des interventions chirurgicales. Pour la gestion des troubles musculo-squelettiques pédiatriques, l'analyse d'images médicales fournit des informations morphologiques et fonctionnelles essentielles pour estimer la gravité du handicap, guider la chirurgie et optimiser les programmes de rééducation. Dans la chaîne de traitement des images médicales, la segmentation est une technologie clef qui permet d'identifier et de localiser les structures anatomiques en délimitant leurs contours [1], [2]. La segmentation permet ainsi de générer des modèles tridimensionnels (3D) solides ou surfaciques des muscles, os, cartilages et ligaments à partir d'images par résonance magnétique (RM) de l'appareil musculo-squelettique pédiatrique. En retour, **ces modèles 3D permettent une compréhension plus précise de l'anatomie pédiatrique, qui est d'autant plus nécessaire car le verdict clinique des pathologies musculo-squelettiques exige une connaissance exacte de la déformation anatomique et du dysfonctionnement articulaire associé** [3]–[5]. De plus, les informations morphologiques et physiologiques ainsi extraites permettent de concevoir des stratégies de rééducation plus efficaces et durables [6]. Les approches de segmentation sont donc primordiales pour la population pédiatrique, où les troubles musculo-squelettiques peuvent gravement entraver la croissance et le développement de l'enfant.

Cependant, la segmentation des images RM repose généralement sur un processus de délimitation manuelle, qui est fastidieux, chronophage et souffre de la variabilité intra- et inter-observateur [7], [8]. En outre, la segmentation de l'appareil musculo-squelettique pédiatrique peut s'avérer plus difficile que celui adulte en raison de la finesse des struc-

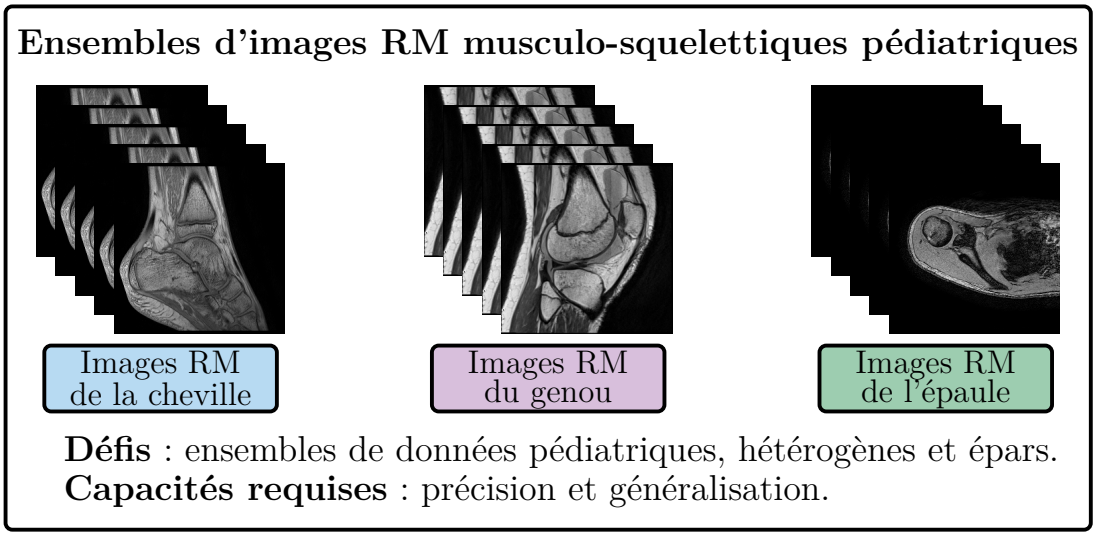


Figure A – Le contexte général de cette thèse est défini par : 1) les ensembles d'images RM musculo-squelettiques pédiatriques disponibles, 2) les défis liés à ces données et 3) les capacités requises pour la segmentation automatique.

tures anatomiques, du processus d'ossification progressive et de la plus grande variabilité morphologique au sein des classes d'âge de l'enfant [7]–[9]. De leur côté, les images pathologiques pédiatriques présentent également des structures anormales, irrégulières et complexes qui sont difficiles à délimiter en raison des altérations de formes et d'apparences [8], [9]. Le développement de techniques de segmentation robustes et entièrement automatisées devient donc nécessaire pour améliorer la fiabilité et la robustesse des délimitations générées, tout en réduisant le besoin d'intervention humaine dans les tâches de traitement d'images médicales [3]–[5]. Dans ce contexte, **cette thèse vise à développer des méthodes de segmentation intégralement automatiques basées sur l'apprentissage profond pour des ensembles de données d'images RM pédiatriques de trois articulations musculo-squelettiques : la cheville, le genou et l'épaule** (Figure A). En particulier, nous ciblons la segmentation de multiples structures osseuses, et les ensembles de données pédiatriques considérés sont caractérisés comme étant hétérogènes, non appariés (c'est-à-dire provenant de différentes cohortes de patients) et épars.

Au cours de la dernière décennie, les approches d'apprentissage profond ont atteint des résultats prometteurs pour la résolution de tâches liées à l'imagerie médicale et ont notamment surpassé les techniques traditionnels d'apprentissage automatique (par exemple, les méthodes variationnelles, les modèles de contour actif ou les approches par partitionnement de graphe) [1], [2], [10]–[12]. Plus précisément, les réseaux de neurones

convolutifs (en anglais CNN pour *convolutional neural networks*) sont devenus des méthodes de pointe dans de nombreuses applications d'imagerie médicale, en raison de leur capacité à apprendre des représentations hiérarchiques encodant les caractéristiques des images d'une manière purement fondée sur les données [13], [14]. Parmi les exemples d'applications basées sur les CNN, citons la détection des lésions de Covid-19 dans les radiographies thoraciques [15] ou le diagnostic de la rétinopathie diabétique à partir de photographies du fond de la rétine [16]. Néanmoins, les techniques d'apprentissage profond sont encore à un stade précoce de déploiement dans la pratique clinique [17]. En 2021, une cinquantaine de dispositifs médicaux et d'algorithmes basés sur l'intelligence artificielle ont été approuvés par la *Food and Drug Administration* états-unienne [1], mais aucun en pédiatrie [18], [19]. En effet, les modèles profonds nécessitent généralement une grande quantité de données annotées pour être entraînés de manière supervisée. Cependant, la complexité du processus d'acquisition et d'annotation des images médicales rend difficile la construction d'ensemble de données à grande échelle. En fin de compte, **les modèles d'apprentissage profond entraînés sur des ensembles de données médicales épars peuvent présenter des performances médiocres sur les images rencontrées lors du déploiement clinique, en raison de capacités de généralisation limitées**.

Pour atténuer ces problèmes, de nouveaux paradigmes d'apprentissage profond ont vu le jour, notamment des approches exploitant des données faiblement annotées ou non annotées [20], [21]. On peut également citer l'apprentissage multi-domaine qui tire parti des caractéristiques partagées entre des ensembles de données acquis à des fins différentes [21] et les techniques de régularisation qui visent à éviter le sur-apprentissage [1]. Il a été démontré que ces méthodologies avancées permettent d'améliorer les performances des réseaux neuronaux standards et sont donc prometteuses pour rendre possible le déploiement généralisé de solutions d'apprentissage profond dans la pratique clinique [21]. **Ainsi, cette thèse porte sur l'analyse automatique d'images pédiatriques qui se révèle encore plus difficile que pour les cohortes adultes, principalement due à la rareté inhérente des ressources d'imagerie pédiatrique.** En effet, l'un des principaux défis associés à l'analyse d'images pédiatriques réside dans la création de grandes bases de données, car l'acquisition d'examens pédiatriques est entravée par le besoin de personnel de santé spécialisé, de considérations éthiques plus strictes et, si nécessaire, de protocoles d'acquisition dédiés [22]–[27]. **Ainsi, la disponibilité limitée des**

1. `https://www.fda.gov/`

ressources d'imagerie pédiatrique rend d'avantage ardu le développement de modèles généralisables qui pourraient être intégrés dans la pratique clinique. Il apparaît donc d'autant plus essentiel de suivre ces nouveaux paradigmes d'apprentissage profond lorsqu'on considère la population pédiatrique.

Objectifs de recherche

Motivé par les problèmes mis en évidence ci-dessus, **l'objectif global de cette thèse est de résoudre les problèmes d'erreurs de généralisation et de rareté des données, rencontrés lors du développement de modèles d'apprentissage profond pour la segmentation d'images musculo-squelettiques pédiatriques. Dans cette direction, cette thèse a proposé de tirer parti des nouvelles méthodologies avancées d'apprentissage profond.** En particulier, nous avons ciblé l'incorporation de régularisations pendant l'optimisation afin d'éviter le sur-apprentissage et l'adoption de schéma multi-anatomie pour bénéficier des caractéristiques partagées entre les ensembles de données d'imagerie musculo-squelettique. Par conséquent, cette thèse vise à développer et à valider des modèles d'apprentissage profond régularisés pour la segmentation multi-anatomie en imagerie pédiatrique. Ce but a été divisé en deux objectifs de recherche décrits ci-dessous :

- **Objectif de recherche 1.** *Développer et valider une approche de segmentation automatique multi-structure avec une régularisation combinée issue d'a priori de formes et de réseaux antagonistes.*
  Cet objectif de recherche vise à développer un pipeline de segmentation osseuse multi-structure pour les images RM pédiatriques. Le modèle a tiré parti d'une combinaison de régularisations issue d'*a priori* de formes et d'un réseau antagoniste pour atténuer les problèmes d'erreurs de généralisation et de rareté des données. En outre, le modèle a exploité une architecture de l'état de l'art et un schéma de d'apprentissage par transfert pour améliorer les performances de segmentation. L'approche est validée et comparée à plusieurs stratégies de segmentation multi-structure et divers architectures CNN sur deux ensembles de données d'imagerie musculo-squelettique pédiatrique des articulations de la cheville et de l'épaule.
- **Objectif de recherche 2.** *Développer et valider un pipeline de segmentation multi-tâche et multi-domaine généralisable avec a priori de formes multi-articulation et une régularisation contrastive multi-échelle.*

Cet objectif de recherche visait à développer et valider une méthode de segmentation multi-tâche et multi-domaine pour l'imagerie musculo-squelettique pédiatrique. Contrairement à l'objectif de recherche précédent, le modèle a simultanément appris à segmenter plusieurs régions anatomiques afin d'atténuer le problème de rareté des ressources pédiatriques. La généralisation du réseau neuronal a également été améliorée grâce à l'intégration d'un encodeur pré-entraîné, d'*a priori* de formes multi-articulation et d'une régularisation contrastive multi-échelle. Le pipeline est validé et comparé à plusieurs approches de segmentation multi-tâche et multi-domaine et divers réseaux convolutifs pour la segmentation de trois ensembles de données d'imagerie pédiatrique des articulations de la cheville, du genou et de l'épaule.

Contenu de la thèse

Étant donné les différentes contributions proposées dans cette thèse, ce manuscrit est divisé en trois parties et la structure de cette thèse est la suivante :

- La **Partie I** présente le contexte et les défis de l'analyse d'images médicales basée sur l'apprentissage profond, en mettant un accent particulier sur les difficultés spécifiques à l'imagerie pédiatrique. Cette partie introduit également les motivations cliniques de l'étude du système musculo-squelettique pédiatrique et le cadre mathématique de l'apprentissage profond pour la segmentation d'images médicales. Les éléments établis dans cette partie fournissent les motivations générales et servent de cadre pour la suite de cette thèse.
  - Le **Chapitre 1** présente le contexte général et les tendances récentes dans le domaine de l'analyse d'images médicales et introduit les défis spécifiques aux applications d'imagerie pédiatrique. En outre, ce chapitre positionne également les méthodes de segmentation proposées dans les parties II et III par rapport aux nouveaux paradigmes d'apprentissage profond développés pour l'analyse d'images médicales.
  - Le **Chapitre 2** présente les motivations cliniques de l'étude du système musculo-squelettique pédiatrique. Ce chapitre décrit les défis associés à l'acquisition et à l'analyse des images musculo-squelettiques pédiatriques. En particulier, le chapitre présente les pathologies visées et les ressources d'imagerie pédiatrique employées dans les Parties II et III.

- Le **Chapitre 3** fournit un cadre mathématique général pour la segmentation d'images basée sur l'apprentissage profond. Ce chapitre vise à fournir des informations de base sur l'apprentissage profond afin de construire des architectures et des schémas d'apprentissage plus avancés. En particulier, ce chapitre présente l'architecture de référence et les détails d'implémentation utilisés dans les expériences réalisées dans les Parties II et III.

- La **Partie II** porte sur le développement et la validation d'un pipeline de segmentation multi-structure incluant une régularisation combinée issue d'*a priori* de formes et de réseaux antagonistes. (**Objectif de recherche 1**).
  - Le **Chapitre 4** présente une méthode de segmentation automatique et multi-structure des os pédiatriques à partir d'images RM. Le pipeline exploite des *a priori* de formes préalablement appris par un auto-encodeur afin de guider le réseau de segmentation à produire des prédictions anatomiquement cohérentes avec des ressources d'imagerie limitées. Ce chapitre montre que l'approche proposée peut être facilement intégrée dans diverses stratégies de segmentation osseuse et démontre l'efficacité de l'utilisation d'un schéma d'apprentissage multi-structure.
  - Le **Chapitre 5** étend l'approche du chapitre 4 en intégrant un encodeur pré-entraîné et une régularisation antagoniste. Le modèle exploite simultanément une combinaison d'*a priori* de formes et un cadre d'apprentissage antagoniste pour réduire le problème de rareté des données tout en améliorant les capacités de généralisation. Enfin, ce chapitre illustre la pertinence d'utiliser des modèles pré-entraînés et de combiner différents schémas de régularisation pour la segmentation d'images médicales basée sur l'apprentissage profond.

- La **Partie III** porte sur le développement et la validation d'un pipeline de segmentation multi-tâche et multi-domaine généralisable avec *a priori* de formes multi-articulation et une régularisation contrastive multi-échelle (**Objectif de recherche 2**).
  - Le **Chapitre 6** présente un cadre d'apprentissage multi-tâche et multi-domaine qui comprend un unique réseau de segmentation optimisé sur l'union de plusieurs ensembles de données d'imagerie. Contrairement aux méthodes précédentes de la Partie II, le modèle ici proposé apprend à segmenter simultanément plusieurs articulations anatomiques afin d'éviter un sur-apprentissage dû à la rareté des données pédiatriques. Ce chapitre présente également des *a priori* de formes

multi-articulaire qui encodent les caractéristiques anatomiques de plusieurs articulations. Enfin, ce chapitre illustre l'adéquation de la mise en œuvre d'un schéma d'apprentissage multi-anatomie pour des ensembles de données d'images musculo-squelettiques pédiatriques.

- Le **Chapitre 7** étend le cadre d'apprentissage multi-anatomie du Chapitre 6 en intégrant une régularisation contrastive multi-échelle qui améliore les capacités de généralisation des modèles de segmentation. En outre, ce chapitre s'appuie sur des schémas d'apprentissage par transfert pour réduire davantage les limitations liées à la rareté des données. Finalement, ce chapitre fournit une évaluation approfondie du cadre d'apprentissage multi-tâche et multi-domaine proposé.

Les différentes contributions proposées au cours de cette thèse ont été valorisées au travers de conférences et revues internationales, et les publications issues de ces projets de recherche sont listées à la fin du manuscrit.

Méthodes proposées

Nous commençons par présenter brièvement les éléments méthodologiques et expérimentaux sur lesquels reposent le développement et la validation des deux pipelines proposés durant cette thèse. **Dans le cadre de l'apprentissage profond, la tâche de segmentation visée dans cette thèse est formulée comme un problème d'approximation de fonction dans lequel le but est d'établir une correspondance entre le domaine des images RM pédiatriques et l'espace des segmentations osseuses.** On peut approximer cette fonction par un réseau de neurones dont les paramètres doivent être appris lors d'une étape d'optimisation basée sur un ensemble d'entraînement incluant les images RM et les segmentations vérités-terrains associées (produites manuellement). En pratique, l'étape d'optimisation des modèles profonds s'appuie sur une fonction de perte (ou coût) permettant d'apprendre les paramètres (ou neurones) du réseau. Dans le contexte de l'apprentissage supervisé, la fonction de perte mesure généralement l'erreur entre les prédictions du modèle et les annotations vérités-terrains. En pratique, il est conseillé de suivre le principe du maximum de vraisemblance et d'utiliser une fonction de perte d'entropie croisée qui permet d'obtenir le meilleur modèle selon les exemples d'apprentissage. La procédure d'apprentissage vise donc à trouver les paramètres qui minimisent cette fonction de perte et l'algorithme de descente de gradient est un outil

d'optimisation standard pour trouver un minimum local d'une telle fonction [13], [14]. Il convient de noter que cette procédure d'apprentissage est entièrement générique et ne se limite pas au cadre de la segmentation d'images médicales. Néanmoins, l'architecture des réseaux de neurones utilisés est quant à elle spécifique à leurs applications et les réseaux convolutifs sont ainsi les plus appropriés pour résoudre des tâches de traitement de l'image [13], [14], [28].

Pour la segmentation d'images médicales, la plupart des modèles d'apprentissage profond sont conçus sur la base de UNet [29] en raison de ses performances surpassant les autres réseaux convolutifs. **Le modèle UNet est un CNN comportant un encodeur qui extrait les caractéristiques de l'image et un décodeur qui prédit une segmentation à partir de la représentation encodée.** Par ailleurs, l'ajout de connexions par sauts (en anglais *skip connections*) entre l'encodeur et le décodeur permet à UNet d'extraire les détails fins de l'image et de générer des segmentations plus précises [29]. Ainsi, UNet et son homologue 3D VNet [30] ont déjà été appliqués à la segmentation de structures musculo-squelettiques dans des images RM d'adultes, notamment les os, les muscles, les cartilages et les ligaments du genou [31]–[36], les os de l'épaule [37], les cartilages du poignet [38] et les muscles de la cuisse [39]. Cependant, **les études consacrées à la segmentation d'images RM musculo-squelettiques pédiatriques restent rares dans la littérature**, à l'exception des travaux de Conze et al. ciblant les muscles de l'épaule [40]. Ainsi, dans le contexte de la segmentation osseuse pédiatrique, la question reste ouverte de savoir si des réseaux spécialisés pour chaque structure osseuse offrent de meilleures performances qu'un unique modèle exploitant les caractéristiques partagées entre les os. En parallèle, de nombreuses extensions du modèle UNet ont été récemment proposées, notamment des modèles intégrant des couches de normalisation par lots (en anglais *batch normalization* [41]) qui permettent d'améliorer la stabilité de l'optimisation. On peut également mentionner l'Attention UNet (Att-UNet [42]) qui intègre le concept d'attention (en anglais *attention gate*) aux *skip connections* favorisant une focalisation sur les zones les plus pertinentes de l'image.

Les expériences menées dans cette thèse ont été réalisées sur des ensembles de données d'imagerie RM pédiatrique de trois articulations musculo-squelettiques : la cheville, le genou et l'épaule (Figure A). Les ensembles d'images de la cheville et de l'épaule ont été acquis au Centre Hospitalier Régional Universitaire (CHRU) La Cavale Blanche, Brest, France, à l'aide d'un scanner Achieva 3.0T (Philips

Healthcare, Best, Pays-Bas) [2] tandis que les données d'imagerie du genou ont été obtenues au *Children's Mercy Hospital*, Kansas City, États-Unis [3]. Les images du genou ont été acquises à l'aide d'un scanner 3.0T MAGNETOM Skyra, Siemens Healthineers, Siemens AG). L'acquisition des données d'imagerie par résonance magnétique (IRM) a été réalisée conformément aux principes de la Déclaration d'Helsinki. Les autorisations éthiques ont été respectivement accordées par le Comité de Protection de Personnes Ouest VI du CHRU de Brest (2015-A01409-40) et par le comité d'éthique de la recherche du *Children's Mercy Hospital*, Kansas City, États-Unis. Des informations supplémentaires sur les cohortes de patients et les structures osseuses ciblées sont fournies pour chaque ensemble de données, comme suit :

- **Ensemble de données de la cheville.** L'ensemble de données d'images de la cheville contient 20 examens RM acquis sur des individus pédiatriques âgés de 7 à 13 ans (âge moyen : $10,1 \pm 2,1$ ans). Toutes les images ont été annotées par un expert médical (15 ans d'expérience) afin d'obtenir les segmentations vérités-terrains du calcanéus, du talus et du tibia.
- **Ensemble de données du genou.** L'ensemble de données d'imagerie du genou est constitué de 17 examens RM extraits d'une cohorte pédiatrique composée de patients âgés de 13 à 18 ans (âge moyen : $15,4 \pm 1,6$ ans). Les masques de segmentation des os du fémur, de la fibula, de la patella et du tibia ont été produits manuellement.
- **Ensemble de données de l'épaule.** Des images RM de 15 articulations de l'épaule ont été obtenues chez des enfants âgés de 5 à 17 ans (âge moyen : $11,6 \pm 4,4$ ans). Les segmentations vérités-terrains des os de l'humérus et de la scapula ont été réalisées en suivant le même protocole que pour les ensembles de données de la cheville et du genou.

Au cours des expériences réalisées durant cette thèse, nous avons utilisé le modèle Att-UNet avec couches de *batch normalization* comme modèle de référence pour comparer les performances des stratégies de segmentation proposées. **Les capacités de généralisation des modèles proposés ont été évaluées sur des données de test non vues durant l'apprentissage et en utilisant des schémas de validation**

2. Les données ont été acquises avec le soutien de la Fondation motrice (2015/7), la Fondation de l'Avenir (AP-RM-16-041), le PHRC 2015 (POPB 282), et le programme Innoveo du CHRU Brest.

3. Nous tenons à remercier le Dr Antonis Stylianou de *University of Missouri-Kansas City*, Kansas City, États-Unis, et le Dr Donna Pacicca du *Children's Mercy Hospital*, Kansas City, États-Unis, pour avoir partagé les images anonymisées de l'articulation du genou.

croisée "un-contre-tous" (en anglais *leave-one-out*). Cette évaluation a été basée sur six métriques calculées à partir des segmentations 3D vérités-terrains produites manuellement par l'expert. Ces mesures ont inclus le coefficient de Dice, la sensibilité, la spécificité, la distance symétrique surfacique maximale, la distance symétrique surfacique moyenne et la différence absolue de volume relatif. Les scores obtenus sont des indicateurs de la similarité entre la vérité-terrain et la segmentation prédite, et permettent ainsi d'évaluer la capacité du modèle à générer automatiquement les mêmes segmentations que celles produites manuellement. Par ailleurs, un système de classement regroupant toutes les mesures de segmentation en un score unique a été créé afin de faciliter la comparaison quantitative de toutes les méthodes implémentées. Nous avons également calculé des tests statistiques à partir des scores de segmentation pour estimer la différence statistique entre les performances atteintes par les approches proposées et les modèles de référence. Finalement, une validation qualitative approfondie de nos méthodes a été effectuée grâce à une comparaison visuelle entre les segmentations générées.

Amélioration de la segmentation multi-structure par une régularisation combinée issue d'*a priori* de formes et de réseaux antagonistes

Pour atténuer les problèmes d'erreurs de généralisation et de rareté des données, des travaux récents visent à intégrer des régularisations dans les modèles de segmentation profonds. **En apprentissage profond, le concept de régularisation couvre des techniques variées qui peuvent affecter l'architecture du réseau, les paramètres appris, les données d'apprentissage ou la fonction de perte [43].** L'architecture UNet [29] contient déjà de nombreuses régularisations de part la présence de couches convolutionnelles qui contraignent le réseau à n'intégrer que des transformations équivariantes avec des interactions locales [13], [14], [28]. Par ailleurs, la *batch normalization* peut être considérée comme une technique de régularisation basée sur les données qui renforce la robustesse des paramètres appris de part ses propriétés inhérentes de stochasticité [41]. Concernant les schémas de régularisation affectant les poids du réseau, on peut également mentionner le transfert d'apprentissage qui fait référence à l'utilisation de poids pré-entraînés sur un domaine d'images similaire [44]. Ainsi, le transfert d'apprentissage à partir de grands ensembles de données d'images naturelles, en particulier ImageNet [45], s'est révélé être une approche efficace pour l'analyse d'images médicales. En effet,

celui-ci permet d'exploiter les caractéristiques de bas niveau (par exemple, les contours) généralement partagées entre différents types d'images et les poids préalablement appris fournissent une initialisation robuste pour l'optimisation [40], [44], [46].

Des schémas de régularisation spécifiques à l'imagerie médicale ont également émergés et ceux-ci peuvent provenir de différentes informations *a priori* liées aux structures anatomiques d'intérêt, telles que leur contour, leur forme ou leur topologie [47]–[49]. L'exploitation de ces connaissances *a priori* s'est avérée efficace pour obtenir des résultats plus précis et plus cohérents dans les techniques de segmentation d'images médicales traditionnelles (c'est-à-dire par apprentissage automatique) [50]. Plus précisément, les techniques de régularisation permettent notamment d'atténuer la présence d'artefacts qui sont intégrés à une image pendant son acquisition [50]. Suite à cela, des travaux récents visent à incorporer des contraintes de régularisation similaires dans des modèles profonds de segmentation. Dans ce contexte, deux approches de régularisation basées sur des fonctions de perte ont montré des résultats prometteurs : la régularisation issue d'*a priori* de formes (en anglais *shape priors*) [51]–[55] et la régularisation antagoniste (en anglais *adversarial*) [46], [56]–[58]. Ces approches reposent sur l'ajout d'un terme de pénalité à fonction de coût qui permet de régulariser, contraindre et guider le modèle durant l'apprentissage. Plus important encore, **ces techniques de régularisation apparaissent comme des stratégies clés pour améliorer les résultats de segmentation et les capacités de généralisation des modèles entraînés sur des ensembles de données épars**.

En effet, des contributions récentes ont proposé d'utiliser un auto-encodeur convolutif pour apprendre une représentation de l'anatomie à partir des annotations vérités-terrains. En raison de la nature contrainte des structures anatomiques (par exemple, la position, la taille et la forme globale des os), les modèles axés sur les données tels que les auto-encodeurs sont particulièrement adaptés pour apprendre des *a priori* de formes [51]–[55]. La représentation de la forme anatomique ainsi apprise peut ensuite être intégrée dans le réseau de segmentation pendant l'étape d'optimisation. Cette intégration s'effectue via un terme de régularisation spécialement conçu qui oblige les segmentation prédites à être proches des vérités-terrains dans l'espace de forme. Par conséquent, une telle régularisation encourage des prédictions avec des formes anatomiques globalement cohérentes [53], [55]. En parallèle, des chercheurs en imagerie médicale ont également proposé d'utiliser des réseaux antagonistes pour affiner les résultats de segmentation [46], [56]–[58]. Dans ces pipelines, un réseau de segmentation et un discriminateur sont entraînés simultanément et de manière compétitive. Le premier apprend à produire des segmentations valides tandis

que le second apprend à discriminer entre les annotations générées et réelles. Un terme antagoniste calculé par le discriminateur est ainsi ajouté pendant l'optimisation du réseau de segmentation, ce qui encourage ce dernier à tromper le discriminateur et à produire des segmentation de plus en plus réalistes [46], [56]–[58].

Pour le premier pipeline de cette thèse, nous avons proposé un encodeur-décodeur convolutif partiellement pré-entraîné intégrant une régularisation combinée issue d'*a priori* de formes et d'un réseau antagoniste ($\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$). Notre approche a permis d'améliorer les performances de segmentation multi-structure sur des ensembles de données d'imagerie pédiatrique du système musculo-squelettique. **Contrairement aux méthodes précédentes [46], [52]–[58], le modèle exploite simultanément les deux régularisations pour réduire le problème de rareté des données tout en améliorant les capacités de généralisation.** En particulier, la régularisation basée sur les *a priori* de formes, dérivés d'une représentation non linéaire de la forme osseuse, a guidé le réseau de segmentation à prédire des segmentations anatomiquement cohérentes. De son côté, la régularisation antagoniste calculée par un réseau discriminateur a encouragé des délimitations plus précises avec des ressources d'imagerie limitées. En outre, le pipeline tire parti d'un encodeur `ResNet50` [59] issu de l'état de l'art et d'un schéma d'apprentissage par transfert à partir de la base de données ImageNet [45] pour atténuer davantage les limitations liées à la rareté des données. Finalement, la méthode proposée exploite également des annotations multi-classes afin d'apprendre des caractéristiques spécifiques et partagées entre les structures osseuses et ainsi améliorer les performances de segmentation.

Pour la première expérience, nous avons employé l'Att-UNet comme réseau de référence pour étudier trois stratégies de segmentation osseuse : individuelle, globale et multiple. Dans le schéma par classe individuelle, des réseaux distincts ont été optimisés sur chaque classe d'intérêt, et les poids appris ont donc été spécifiques à un seul os. Pour l'approche classe globale, les différentes structures osseuses ont été considérées comme une unique classe globale sans distinction entre les os, et les poids appris ont été spécifiques à cette classe. Enfin, dans la stratégie multi-classe, les réseaux ont été entraînés sur des segmentations contenant plusieurs classes, et les poids appris ont ainsi été partagés entre toutes les structures anatomiques. Ainsi, l'approche individuelle a nécessité de segmenter chaque os de manière séquentielle à l'aide de réseaux distincts, tandis que les schémas global et multiple ont chacun reposé sur un CNN unique, générant soit une classe osseuse globale, soit des segmentations spécifiques à chaque os. En outre, pour chaque stratégie de segmentation osseuse (individuelle, globale et multiple), nous avons réalisé une étude par ablation

afin d'évaluer les contributions des termes de régularisation. Nous avons ainsi comparé Att-UNet [42], Att-UNet avec régularisation issue d'*a priori* de formes [53], Att-UNet avec régularisation antagoniste [57] et Att-UNet avec la régularisation combinée proposée. Les résultats obtenus ont montré que la méthode de segmentation basée sur un modèle multi-classe avec régularisation combinée a obtenu les meilleurs performances. En particulier, nous avons observé que, pour un schéma de régularisation fixé, la stratégie multi-classe a surclassé la stratégie globale, qui à son tour a surpassé la stratégie individuelle. De son côté, **la régularisation combinée a constamment surpassé les autres méthodes de régularisation, démontrant ainsi l'efficacité de l'approche proposée**.

Dans un second temps, nous avons évalué les performances de notre méthode basée sur un réseau Res-UNet pré-entraîné incorporant une stratégie de segmentation multi-classe et la régularisation combinée proposée ($\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$). En particulier, nous l'avons comparée à deux autres architectures avec des encodeurs pré-entraînés issus de l'état de l'art (VGG-UNet [60] et Dense-UNet [61]). Nous avons uniquement utilisé la stratégie multi-classe avec régularisation combinée car celle-ci a obtenu les meilleures performances durant l'expérience précédente. La méthode $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ a surpassé toutes les autres approches avec encodeur pré-entraîné sur quasiment toutes les métriques des deux ensemble de données de la cheville et de l'épaule. **Le pipeline proposé a notamment obtenu d'excellents scores de Dice :** $94,1\%$ **pour les os de la cheville et** $89,5\%$ **pour les structures de l'épaule.** $\mathbf{CombReg^{Multi}_{Res\text{-}UNet}}$ **s'est ainsi classé premier en termes de performance pour les deux ensembles de données.** En outre, les tests statistiques réalisés ont indiqué que le modèle proposé a produit des améliorations significatives sur chaque métrique. Enfin, les comparaisons visuelles ont fourni des preuves qualitatives de l'amélioration progressive des performances de segmentation. L'approche multi-classe a permis au réseau d'apprendre simultanément des caractéristiques osseuses spécifiques (par exemple la position) et partagées (par exemple l'intensité et la forme), et a ainsi produit des délimitations précises tout en évitant les problèmes d'os fusionnés rencontrés avec le schéma par classe globale. Parallèlement, la régularisation combinée proposée a exploité les avantages des deux régularisations précédentes et a permis une extraction des os plus régulière et plus précise.

Bien que les résultats obtenus soient satisfaisants et apportent de nouvelles perspectives pour la gestion des troubles musculo-squelettiques dans la population pédiatrique, le développement de modèles d'apprentissage profond généralisables reste néanmoins un défi. Dans cette direction, la prochaine méthode proposée durant cette thèse a pour but de

formaliser et d'implémenter un cadre d'apprentissage multi-tâche et multi-domaine. **En effet, si le pipeline ici proposé a illustré les avantages de la segmentation multi-structure pour tirer profit des caractéristiques partagées entre les os d'une même articulation anatomique, on peut facilement étendre cette réflexion et supposer que les os de régions anatomiques distinctes présentent également des caractéristiques communes qui pourraient être exploitées.**

Segmentation multi-tâche et multi-domaine généralisable avec *a priori* de formes multi-articulation et une régularisation contrastive multi-échelle

Comme évoqué précédemment, le concept de régularisation, qui englobe toutes les méthodes visant à réduire le sur-apprentissage, ne se limite pas à l'ajout de termes de pénalité à la fonction de perte. Récemment, les approches d'apprentissage multi-tâche [62]–[65] et multi-domaine [66]–[71] ont suscité l'intérêt des chercheurs en analyse d'images médicales. **Intuitivement, les modèles multi-tâche et multi-domaine bénéficient du partage de paramètres entre tâches ou domaines pour apprendre des représentations plus robustes et plus génériques que leurs homologues individuels [72]–[74].** Ces approches sont particulièrement dignes d'intérêt dans le cadre de la segmentation de multiples ensembles de données pédiatriques de régions musculo-squelettiques distinctes. En effet, on peut facilement supposer que les ensembles de données pédiatriques RM provenant de différentes articulations anatomiques (par exemple, la cheville, le genou et l'épaule) présentent des caractéristiques communes, en termes de forme, de pose et d'intensité. En parallèle, il pourrait également être bénéfique de concevoir des termes de régularisation spécifiques à l'apprentissage multi-tâche et multi-domaine afin d'améliorer les performances de segmentation et d'obtenir des modèles avec de plus grandes capacités de généralisation. Par exemple, **les études portant sur les *a priori* de formes n'ont jamais proposé d'encoder simultanément plusieurs régions anatomiques afin d'exploiter les corrélations de position, d'orientation, de taille et de forme entre des objets anatomiques similaires**, tels que des os pédiatriques dans des articulations musculo-squelettiques séparées.

Les premiers cadres d'apprentissage multi-tâches et multi-domaines ont été développés pour l'analyse d'images naturelles. Dans le contexte de la segmentation, Fourure et al. [75] ont proposé d'entraîner un seul réseau sur l'union de plusieurs ensembles de

données pour faire face à la quantité limitée de données annotées. Dans leur approche, chaque ensemble de données est caractérisé par sa propre tâche de segmentation et son propre domaine d'images. Par conséquent, ce cadre est plus générique que les approches multi-tâches traditionnelles qui se concentrent généralement sur plusieurs tâches dans le même domaine ou que les techniques multi-domaines traditionnelles qui considèrent des domaines contenant le même ensemble d'objets. Par la suite, **des études ultérieures ont proposé d'employer un modèle unique avec des couches convolutionnelles agnostiques, puisque les primitives visuelles peuvent être partagées entre les tâches et les domaines, et des couches spécifiques à chaque ensemble de données qui permettent une spécialisation pour chaque tâche et chaque domaine [76]–[78]**. Ces approches, basées sur des représentations partagées, ont permis d'atteindre des performances égales ou supérieures aux modèles individuels traditionnels. A notre connaissance, l'apprentissage multi-tâche et multi-domaine a cependant rarement été appliqué à l'analyse d'images médicales, à l'exception des travaux de Moeskops et al. [79] qui ont démontré qu'un seul réseau de neurones peut segmenter plusieurs anatomies (cerveau, poitrine et cœur) simultanément. Néanmoins, de part l'absence de couches spécifiques à chaque domaine, leur modèle n'a pu tenir compte de la différence de distribution d'intensité entre les domaines.

Même si les modèles multi-tâches et multi-domaines peuvent intégrer des informations spécifiques à chaque tâche et domaine par le biais de couches spécialisées, les connaissances préalables (ou *a priori*) pourraient être davantage exploitées pour améliorer les capacités de généralisation des représentations partagées. Dans cette direction, Zhu et al. [80] ont imposé une distribution de mélange gaussien sur la représentation partagée de leur réseau afin de préserver les informations de bas niveau entre les domaines. Cependant, une telle hypothèse peut se révéler trop restrictive en pratique. En effet, dans le cadre de l'apprentissage de représentations, une bonne représentation peut être caractérisée par la présence de *clusters* correspondant aux classes du problème (c'est-à-dire une représentation démêlée ou *disentangled representation* en anglais) [81]. Par conséquent, un certain nombre de techniques d'apprentissage auto-supervisée proposent d'utiliser une métrique contrastive afin de regrouper les données de la même classe et séparer celles de classes différentes dans les représentations cachées (ou espaces latents) [82]–[84]. Une contribution récente a étendu cette idée au cadre de la classification supervisée en tirant parti de l'information sur les classes des images [85]. Ainsi, la régularisation contrastive maximise la performance du classificateur en imposant une cohésion intra-classe et une séparation

inter-classe dans les couches latentes. Contrairement à l'approche de Zhu et al. [80], il n'est pas nécessaire de définir, au préalable, une distribution pour les variables latentes. Par conséquent, les techniques de régularisation contrastive semblent plus génériques et appropriées pour imposer des *clusters* spécifiques à chaque domaine dans les représentations partagées des modèles profonds multi-tâche et multi-domaine.

Pour le second pipeline de cette thèse, nous avons implémenté et évalué un cadre d'apprentissage multi-tâche et multi-domaine pour la segmentation d'os pédiatriques à partir d'images RM. **Contrairement au pipeline précédent, cette approche multi-anatomie bénéficie de représentations partagées apprises à partir d'articulations anatomiques distinctes, afin d'atténuer le problème de rareté des ressources pédiatriques.** En particulier, le modèle de segmentation multi-tâche et multi-domaine a intégré un encodeur pré-entraîné de la famille EfficientNet, des filtres convolutifs partagés, des *attention gates* multi-domaine, des couches de *batch normalisation* spécifiques à chaque domaine (en anglais DSBN pour *domain-specific batch normalisation*) et une couche de segmentation spécifique à chaque tâche. En effet, dans le cadre de l'apprentissage multi-domaine, sachant que les statistiques individuelles des domaines peuvent être très différentes les unes des autres, une unique couche de *batch normalization* partagée pourrait conduire à négliger certaines caractéristiques et à apprendre des poids nuls. Pour calibrer plus minutieusement les caractéristiques internes du modèle, nous avons utilisé des DSBN [66]–[69]. **Ainsi, les filtres convolutifs partagés exploitent les caractéristiques partagées entre les tâches et les domaines pour être plus robustes que leurs homologues individuels, tandis que les DSBN permettent de meilleures capacités de généralisation grâce à un calibrage spécifique à chaque domaine.** D'autre part, une couche de segmentation agnostique pouvant prédire les classes de chaque articulation est contre-productive puisque les tâches ciblées sont disjointes (par exemple, prédire des os de cheville à partir d'une image d'épaule) [75]. Il a donc été essentiel d'utiliser une couche finale dédiée pour chaque tâche de segmentation. En outre, les *attention gates* multi-domaine ont permis d'améliorer l'interprétabilité du modèle de part leurs capacités à produire des cartes d'attention qui mettent en évidence les régions d'intérêt dans chaque domaine. Finalement, nous étendons le cadre d'apprentissage multi-tâche et multi-domaine en définissant une régularisation contrastive multi-échelle (MSC en anglais pour *multi-scale contrastive*) et des *a priori* de formes multi-articulation (MJSP en anglais pour *multi-joint shape priors*) qui permettent d'améliorer les capacités de généralisation du modèle de segmentation. Plus précisément, **la régularisation contrastive multi-**

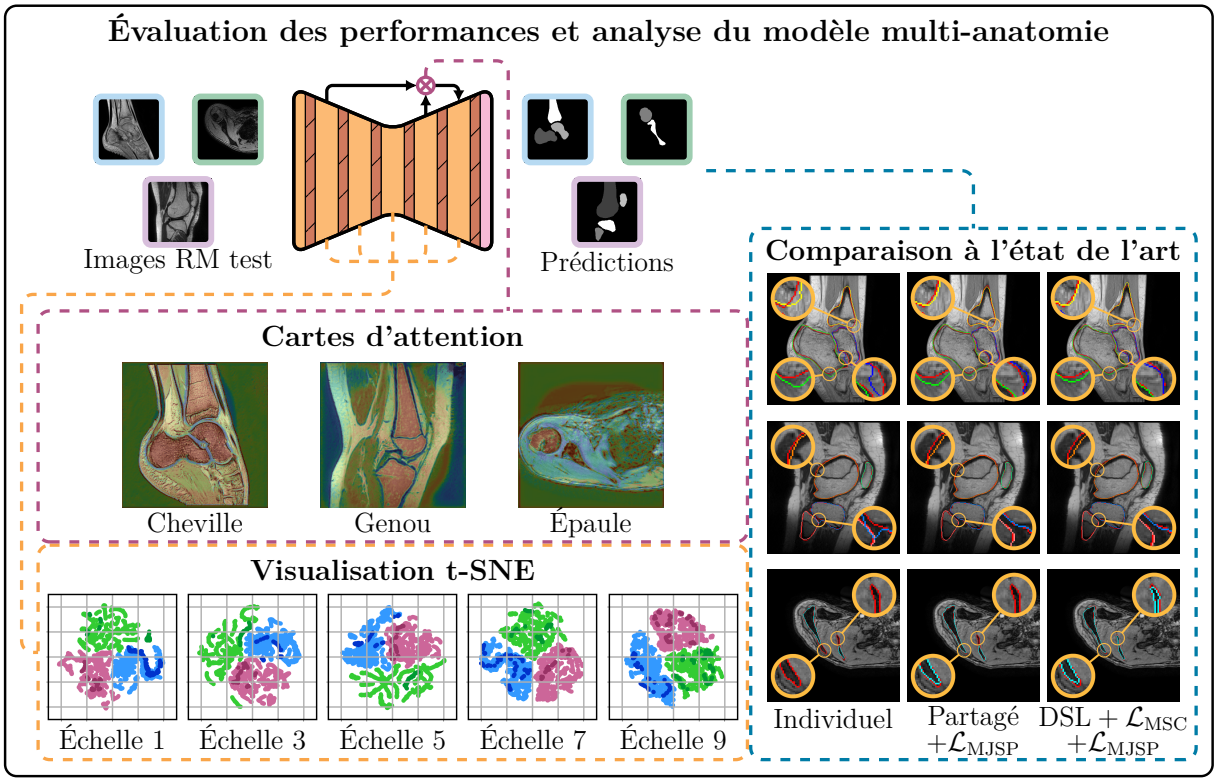


Figure B – Évaluation des performances et analyse du modèle de segmentation multi-tâche et multi-domaine généralisable avec *a priori* de formes multi-articulation et régularisation contrastive multi-échelle.

échelle vise à améliorer la similarité intra-domaine et à imposer des marges inter-domaines dans les représentations latentes du réseau; tandis que les *a priori* de formes multi-articulation encodent les caractéristiques anatomiques de plusieurs articulations pour contraindre la tâche de segmentation.

Pour la première expérience, nous avons comparé différentes stratégies de segmentation multi-tâche et multi-domaine avec Att-UNet comme architecture de référence. Les quatre approches de segmentation proposées sont les suivantes : individuelle (entraînée sur des domaines individuels), transfert (pré-entraînée sur un domaine et affinée sur les autres), partagée (entraînée sur tous les domaines à la fois, avec tous les paramètres partagés entre les domaines) et DSL (pour *domain-specific layers* en anglais, entraînée sur tous les domaines à la fois, avec des paramètres partagés et des couches spécifiques à chaque domaine). L'approche partagée diffère du schéma DSL par ses couches de *batch normalization* et de segmentation partagées. En outre, nous avons réalisé une étude par ablation pour évaluer les contributions des *a priori* de formes multi-articulation et de la régularisa-

tion contrastive multi-échelle. Plus précisément, les *a priori* de formes multi-articulation ont été incorporés dans les approches partagées (en utilisant un auto-encodeur multi-articulation avec tous les paramètres partagés) et DSL (en utilisant un auto-encodeur multi-articulation avec des paramètres partagés et spécifiques à chaque domaine). Pour sa part, la régularisation contrastive multi-échelle n'a pu être intégrée que dans le schéma DSL, puisque les domaines n'ont pas été différenciés dans l'approche partagée. D'après les résultats obtenus, **notre approche comprenant des représentations partagées, des couches spécialisées, des *a priori* de formes multi-articulation et une régularisation contrastive multi-échelle a permis d'obtenir des améliorations de performance par rapport aux modèles indépendants, par transfert et partagé sur tous les ensembles de données**. Par ailleurs, les schémas partagé et DSL offrent un avantage supplémentaire de part leurs capacités à apprendre toutes les paires de tâches et de domaines simultanément plutôt que de manière séquentielle, source d'oublis catastrophiques. Enfin, nous avons observé que les *a priori* de formes multi-articulation ont amélioré les performances de segmentation dans les deux schémas partagés et DSL, et ce pour chaque articulation anatomique. Cela a davantage illustré l'efficacité des *a priori* de formes qui s'étaient déjà avérés bénéfiques dans le pipeline précédent, mais pour chaque articulation anatomique séparée.

Dans un second temps, nous avons comparé les performances de notre méthode (Efficient-UNet DSL+$\mathcal{L}_{\text{MSC}}$+$\mathcal{L}_{\text{MJSP}}$) basée sur Efficient-UNet avec encodeur pré-entraîné, DSL, régularisation contrastive multi-échelle et *a priori* de formes multi-articulation face à deux autres architectures avec des encodeurs pré-entraînés issus de l'état de l'art (Inception-UNet [86] et Dense-UNet [61]). Plus précisément, les modèles pré-entraînés Inception-UNet, Dense-UNet et Efficient-UNet ont été comparés en utilisant des schémas individuels, partagés avec *a priori* de formes multi-articulation, et DSL avec *a priori* de formes multi-articulation et une régularisation contrastive multi-échelle. Pour chacune des stratégies partagées et DSL, nous n'avons retenu que la meilleure approche observée lors des expériences précédentes basées sur Att-UNet. Enfin, le schéma de transfert a été écarté dans ce dispositif expérimental car les réseaux ont déjà été tous partiellement pré-entraînés sur la base de données ImageNet. Le pipeline Efficient-UNet DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ a surpassé toutes les autres approches avec encodeur pré-entraîné sur quasiment toutes les métriques des trois ensembles de données de la cheville, du genou et de l'épaule. **La méthode proposée a notamment obtenu d'excellents scores de Dice :** $93,8\%$**,** $95,4\%$ **et** $87,9\%$ **respectivement sur les ensembles de données de la cheville, du genou et de**

l'épaule. Efficient-UNet DSL + $\mathcal{L}_{\text{MSC}} + \mathcal{L}_{\text{MJSP}}$ s'est ainsi classé premier en termes de performances pour la tâche de segmentation multi-anatomie. Les tests statistiques ont confirmé que notre pipeline a conduit à des améliorations significatives sur chaque métrique par rapport aux autres approches. Enfin, les comparaisons visuelles (Figure B) ont permis de souligner de manière qualitative les améliorations progressives de la précision des segmentations. Nous avons notamment observé que les *a priori* de formes ont imposé des délimitations globalement plus anatomiquement cohérentes pour toutes les structures osseuses ciblées, tandis que la régularisation contrastive a encouragé une extraction plus précise des os dans tous les domaines grâce à des représentations partagées plus robustes avec des *clusters* spécifiques à chaque domaine. En outre, nous avons fourni une visualisation des cartes d'attention calculées par les *attention gates* multi-domaine (Figure B). Cette visualisation a confirmé que les modèles de segmentation ont exploité les informations spatiales et contextuelles pour se concentrer sur les os ciblés dans chaque articulation anatomique. Finalement, nous avons utilisé l'algorithme de visualisation t-SNE [87] pour évaluer les effets de la régularisation contrastive multi-échelle sur les représentations internes des réseaux multi-domaine (Figure B). Nous avons observé que les représentations apprises à l'aide des schémas partagé et DSL n'ont pas présenté de marges entre les domaines. Au contraire, l'ajout de la régularisation contrastive a conduit à des *clusters* distincts spécifiques à chaque domaine. Par conséquent, le domaine de l'image d'entrée a été conservé au travers des différentes représentations partagées des réseaux proposés, ce qui s'est traduit par de meilleures performances sur les images non vues.

Conclusion

Les travaux de recherche menés dans le cadre de cette thèse visaient à résoudre les problèmes d'erreurs de généralisation et de rareté des données rencontrés lors du développement de méthodes d'apprentissage profond pour la segmentation d'images musculo-squelettiques pédiatriques. **Nous avons proposé et évalué des méthodes basées sur de nouveaux paradigmes d'apprentissage profond qui ont atteint des performances prometteuses pour la tâche de segmentation osseuse sur des ensembles de données d'imagerie RM épars et hétérogènes des articulations de la cheville, du genou et de l'épaule.** En particulier, les performances de généralisation des modèles de segmentation ont été améliorées en exploitant des architectures de l'état de l'art, des

schémas d'apprentissage par transfert, des approches multi-anatomie et des techniques de régularisation.

Dans le cadre de cette thèse, nous n'avons pu évalué si les performances de généralisation obtenues par les deux pipelines proposés ($\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ et Efficient-UNet DSL + $\mathcal{L}_{\text{MSC}} + \mathcal{L}_{\text{MJSP}}$) sont suffisantes pour permettre le déploiement dans la pratique clinique de méthodes de segmentation osseuse d'images RM pédiatriques entièrement automatiques. Cependant, **les résultats obtenus illustrent de façon significative que l'utilisation collaborative des ressources pédiatriques et la conception intelligente des modèles d'apprentissage profond peuvent améliorer les performances de segmentation sur de petits ensembles de données d'imagerie musculo-squelettique.** Enfin, nos pipelines fournissent actuellement une description incomplète du système musculo-squelettique pédiatrique qui englobe uniquement les tissus osseux. Par conséquent, les travaux futurs visent à compléter nos modèles afin d'extraire d'autres structures anatomiques (par exemple, les cartilages de la cheville, les ligaments du genou ou les muscles de l'épaule). Ainsi, l'analyse morphologique et fonctionnelle reposera sur une modélisation plus complète du système musculo-squelettique, en vue d'une meilleure gestion des troubles pédiatriques.

La suite de ce manuscrit est rédigée en anglais (États-Unis).

Table of Contents

List of Figures

List of Tables

INTRODUCTION

Context

In clinical practice, medical imaging is a valuable aid for diagnosis, treatment planning, surgery assessment, and post-surgical monitoring. For the management of pediatric musculoskeletal disorders, medical image analysis delivers morphological and functional information essential for assessing the patient's level of impairment, guiding surgery, and optimizing rehabilitation programs. In the medical image analysis workflow, semantic segmentation is a critical technology that allows identifying and localizing meaningful anatomical structures by extracting their boundaries [1], [2]. Hence, segmentation enables the generation of three-dimensional (3D) solid or surface models of muscles, bones, cartilages, and ligaments from pediatric musculoskeletal magnetic resonance (MR) images. In turn, these **3D anatomical models provide an accurate understanding of the pediatric anatomy, which is especially needed as the clinical verdict of musculoskeletal pathologies requires precise knowledge of anatomical deformity and associated joint dysfunction** (Figure C) [3]–[5]. Additionally, the extracted morphological and physiological information allows the design of more efficient and sustainable rehabilitation strategies [6]. Such approaches are paramount in the pediatric population, where musculoskeletal disorders may seriously impede a child's growth and development.

However, the process of segmenting MR images typically relies on manual delineation, which is tedious, time-consuming, and suffers from the lack of intra- and inter-observer reproducibility [7], [8]. Moreover, the pediatric musculoskeletal system may be more challenging to segment than its adult counterpart due to thinner structures, the ongoing bone ossification process, and higher anatomical variability between age groups [7]–[9]. For their part, pediatric pathological imaging examinations also exhibit irregular and complex pathological structures which are difficult to delineate due to alterations in shape and appearance [8], [9]. Developing robust and fully-automated segmentation techniques becomes necessary to improve the generated delineations' reliability and robustness while reducing the need for human intervention in image processing tasks [3]–[5]. **In this thesis, we aim to develop deep learning-based fully-automatic segmentation methods**

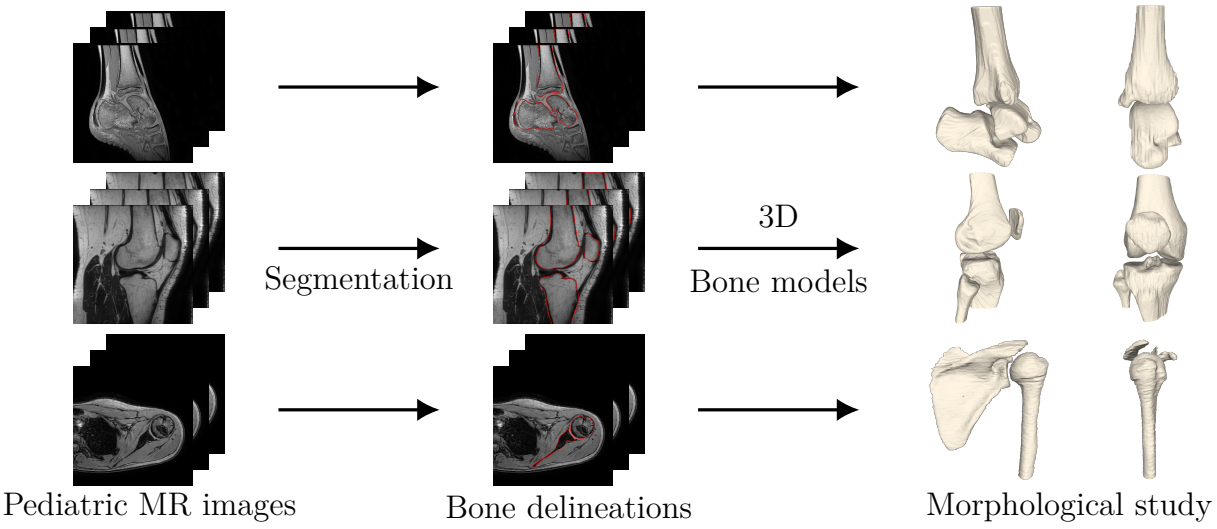


Figure C – Pediatric MR images segmentation workflow and 3D bone models generation for the morphological study of the three musculoskeletal joints of interest: ankle, knee, and shoulder.

for pediatric MR image datasets of three musculoskeletal joints: ankle, knee, and shoulder. In particular, we target the segmentation of multiple bone structures (Figure C), and the available pediatric datasets are characterized as heterogeneous, unpaired (i.e., from different patient cohorts), and sparse.

In the last decade, deep learning approaches have achieved promising results for solving medical imaging-based tasks compared to traditional variational, model-based, or graph-partitioning learning schemes [1], [2], [10]–[12]. Specifically, convolutional neural networks (CNNs) have become state-of-the-art methods in numerous medical imaging-based applications due to their ability to learn hierarchical representations of image features in a purely data-driven manner [13], [14]. Examples of CNNs-based medical image analysis applications include the detection of Covid-19 lesions in chest radiographs [15] or the diagnosis of diabetic retinopathy from retinal fundus photographs [16]. Nevertheless, deep learning techniques are still at an early stage of deployment in clinical practice [17]. In 2021, about fifty artificial intelligence-based medical devices and algorithms have been approved by the United States Food and Drug Administration [4], of which none in pediatrics [18], [19]. Indeed, deep learning models usually require a large amount of annotated data to be trained in a supervised manner. However, the complexity of the medical image acquisition and annotation process makes it challenging to develop large-scale datasets. Ul-

4. `https://www.fda.gov/`

timately, **deep learning models trained on sparse medical datasets may present poor performance on unseen images encountered in real-world deployment due to limited generalization capabilities**.

To mitigate these issues, novel deep learning paradigms have emerged, including annotation-efficient approaches to leverage weakly labeled or unlabeled data [20], [21], multi-domain learning to benefit from features shared across datasets acquired for different purposes [21], and regularization techniques to prevent over-fitting [1]. These advanced methodologies have been demonstrated to improve performance over standard deep learning models and thus show promise in enabling the widespread deployment of deep learning solutions in clinical practice [21]. **Nevertheless, in this thesis, we target the automatic analysis of pediatric images, which is even more difficult than for adult cohorts, primarily due to the inherent scarcity of pediatric imaging resources.** Indeed, one of the major challenges associated with pediatric image analysis resides in creating large-scale imaging databases, as the acquisition of pediatric examinations is hindered by the need for specialized healthcare personnel, dedicated acquisition protocols, and stricter ethical considerations [22]–[27]. **The limited availability of pediatric imaging resources makes it even more difficult to develop generalizable models that could be integrated into clinical practice. It thus appears all the more essential to follow novel deep learning paradigms when considering the pediatric population.**

Research objectives

Motivated by the problems highlighted above, **the global aim of this thesis was to address the generalization gap and data scarcity issues encountered when developing deep learning models for pediatric musculoskeletal image segmentation. In this direction, we proposed to leverage emerging advanced deep learning methodologies.** In particular, we targeted the incorporation of regularization during optimization to prevent over-fitting and the adoption of multi-anatomy learning to benefit from shared features across musculoskeletal imaging datasets. Hence, this thesis aimed at developing and validating regularized deep learning models for multi-anatomy segmentation in pediatric imaging. This aim was factorized into two research objectives outlined below:

- **Research objective 1.** *Develop and validate an automatic multi-structure seg-*

mentation framework with combined regularization from shape priors and adversarial networks.

This research objective aimed to develop a multi-structure bone segmentation pipeline for pediatric MR images. To mitigate the generalization gap and data scarcity issues, we employed a combination of regularization from shape priors and an adversarial network. In addition, the framework leveraged a state-of-the-art architecture and transfer learning scheme to further improve segmentation performance. We validated the framework using several multi-structure segmentation strategies and backbone architectures on two pediatric musculoskeletal imaging datasets of the ankle and shoulder joints.

- **Research objective 2.** *Develop and validate a generalizable multi-task, multi-domain segmentation pipeline with multi-joint shape priors and multi-scale contrastive regularization.*

  This research objective targeted the development and validation of a multi-task, multi-domain segmentation framework for pediatric musculoskeletal imaging. As opposed to the previous research objective, the framework simultaneously learned to segment multiple anatomical regions to mitigate the scarcity issue of pediatric resources. We further improved model generalizability by integrating a pre-trained encoder, multi-joint shape priors, and multi-scale contrastive regularization. The framework is validated and compared with several multi-task, multi-domain segmentation approaches and backbone networks for the segmentation of three pediatric imaging datasets of the ankle, knee, and shoulder joints.

Thesis outline

Considering the multiple research elements included in this thesis, the remainder of this manuscript is divided into three parts, as summarized in the diagram of Figure D. The structure of this thesis is as follows:

- **Part I** provides the context and challenges of medical image analysis using deep learning, with a focus on pediatric imaging. This part also introduces the clinical motivations for the analysis of the pediatric musculoskeletal system and the mathematical framework for deep learning-based medical image segmentation. The elements established in this part provide the general motivations and serve as the backbone for the remainder of this thesis.

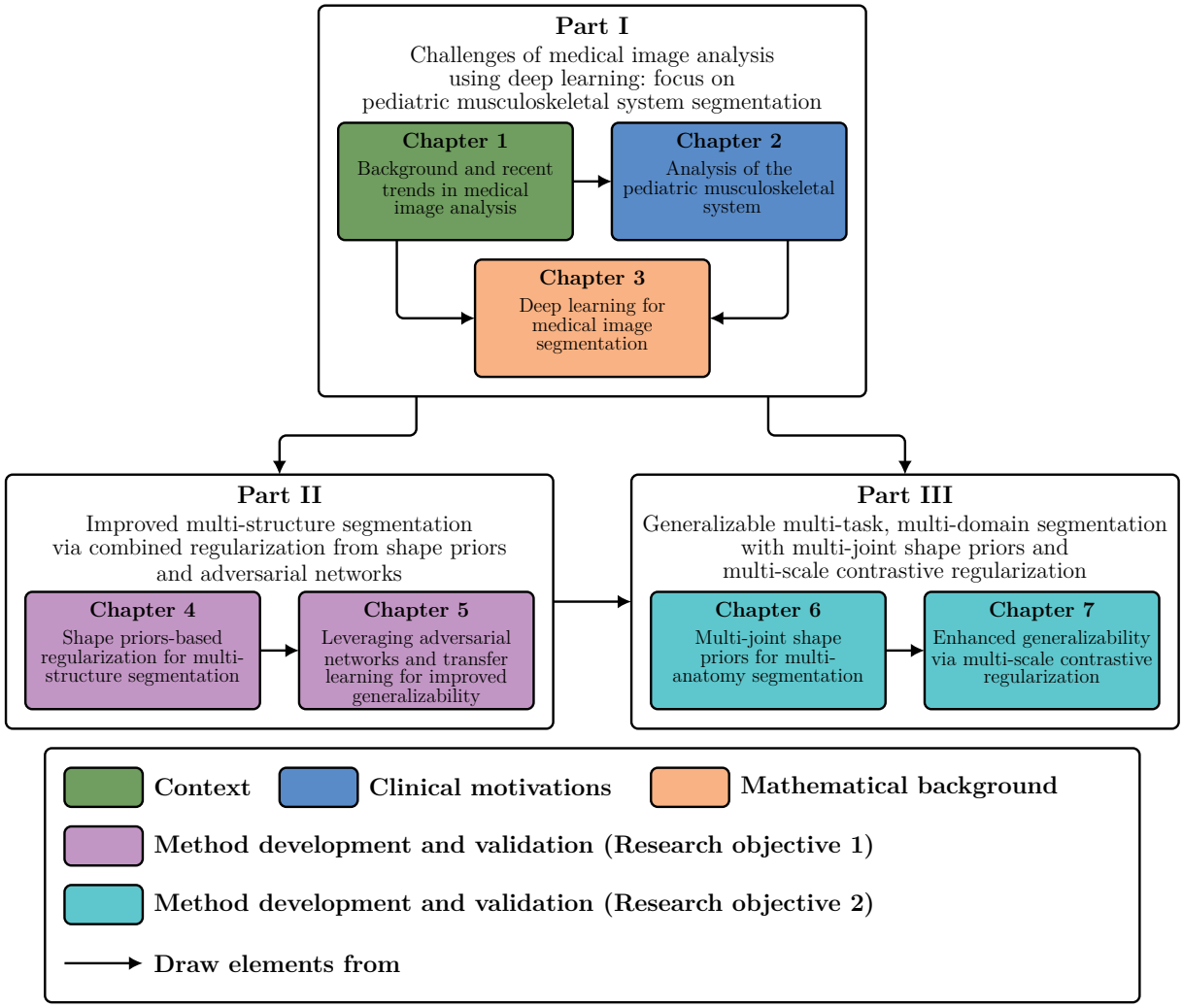


Figure D – Diagram of the thesis structure. The thesis is organized into three parts, in which each chapter draws elements from the previous chapters.

- **Chapter 1** presents general background information and recent trends in the field of medical image analysis, with a final emphasis placed on challenges specific to pediatric imaging applications. In addition, this chapter also positions the segmentation methodologies of Parts II and III in relation to the novel and emerging deep learning paradigms developed for medical image analysis.
- **Chapter 2** introduces the clinical motivations for the study of the pediatric musculoskeletal system. This chapter describes the challenges associated with the acquisition and analysis of pediatric musculoskeletal images. In particular, the chapter presents the pathologies targeted and the pediatric imaging resources employed in Parts II and III.

  - **Chapter 3** provides a general mathematical framework for deep learning-based image segmentation. This chapter aims to provide background on deep learning to build more advanced architectures and training schemes. In particular, this chapter introduces the baseline architecture and the implementation details used in the experiments performed in Parts II and III.
- **Part II** targets the development and validation of the improved multi-structure segmentation framework with combined regularization from shape priors and adversarial networks (**Research objective 1**).
  - **Chapter 4** introduces an automatic and multi-structure pediatric bone segmentation method. The framework leverages auto-encoder based shape priors to guide the segmentation network to make anatomically consistent predictions with restricted imaging resources. This chapter illustrates that the proposed approach can be easily integrated into various bone segmentation strategies, and demonstrates the effectiveness of employing a multi-structure learning scheme.
  - **Chapter 5** extends the framework of Chapter 4 by integrating a pre-trained encoder and an adversarial regularization. The framework simultaneously leverages a combination of shape priors and an adversarial regularizer to reduce the data scarcity issue while improving model generalizability. Finally, this chapter demonstrates the usefulness of employing pre-trained models along with combining different regularization schemes for deep learning-based medical image segmentation.
- **Part III** targets the development and validation of the generalizable multi-task, multi-domain segmentation pipeline with multi-joint shape priors and multi-scale contrastive regularization (**Research objective 2**).
  - **Chapter 6** proposes a multi-task, multi-domain learning framework. Unlike the previous methods of Part II, the framework simultaneously learns to segment multiple anatomical joints to overcome the inherent scarcity of pediatric data. This chapter also presents multi-joint shape priors which encode the anatomical characteristics of multiple joints to further avoid over-fitting. Ultimately, this chapter illustrates the effectiveness of employing a multi-anatomy learning scheme.
  - **Chapter 7** extends the multi-anatomy learning framework of Chapter 6 by integrating a multi-scale contrastive regularization to improve the generalization capabilities of segmentation models. In addition, this chapter leverages transfer

learning scheme to further reduce data scarcity limitations. Finally, this chapter provides an in-depth evaluation of the proposed multi-task, multi-domain learning framework.

Although the thesis is structured so that each chapter uses elements from the previous ones (Figure D), each chapter can also be read independently. Publications resulting from research projects performed during this thesis and for which the author is the primary author or a collaborator are listed at the end of the dissertation.

Part I

Challenges of medical image analysis using deep learning: focus on pediatric musculoskeletal system segmentation

Chapter 1

Background and recent trends in medical image analysis

1.1 Introduction

Imaging of the human anatomy and function serves as an essential step in the medical workflow, as the acquired images can significantly assist clinicians in diagnosing pathologies, assessing morphological evolution over time, and optimally guiding surgeries [11]. **Medical imaging has revolutionized medicine over the past few decades and continues to progress rapidly, especially with the continuous improvements in spatio-temporal resolution and signal-to-noise ratio [10].** However, the analysis of the obtained medical images is traditionally performed by radiologists through visual inspection which is a time-consuming process sensitive to human subjectivity and variability among interpreters (e.g., experience, fatigue) [11]. Most importantly, human-based analysis of medical images can be extremely expensive as it requires strong expert knowledge.

Therefore, as discussed by Rueckert and Schnabel [11], **the automatic extraction and analysis of quantitative information from medical images is of crucial importance and has become an explosive research area in recent years**. More precisely, robust and automatic derivation of anatomical and physiological information could enable the generation of patient-specific modeling to support diagnostic and treatment approaches appropriately tailored to each individual patient, leading to more personalized medicine. In parallel, the analysis of large-scale population studies could also benefit from such an automatic analysis pipeline to extract previously unknown biological patterns and trends present within a specific population and to design new customized strategies for early detection, prediction, and primary prevention of major diseases. In particular, the standard inspection of medical images conducted by human experts is notably unable to cope with such studies involving thousands of patients, contrary to computerized medical

image analysis tools [11].

The work of Litjens et al. [2] provides background on the rich history of computerized medical image analysis that spans from the 1970s to today and reflects the trends within the fields of computer vision and artificial intelligence [14]. Indeed, healthcare and medical image analysis are just one of the many application areas of artificial intelligence, which is a very vast field of research aimed at developing intelligent systems that exhibit behavior similar to human perception and reasoning (e.g., image perception and understanding). The first computerized medical image analysis techniques developed from the 1970s to the 1990s involved low-level pixel processing (e.g., edge filters, region growing) and simple mathematical primitives (e.g., line, ellipsoid, or circle fitting) to derive rule-based systems built upon conditional statements. However, these expert systems based on handcrafted features and rules required extensive knowledge to solve a particular task and proved difficult to transfer to new problems involving unseen images [2], [14].

Thus, a new kind of approach exploiting the knowledge of previously labeled training images emerged in the late 1990s. One can notably mention active shape models which iteratively deform to extract an instance of an object in a new image, with shape deformations remaining consistent with the variability observed in the set of labeled examples. Additionally, statistical classifiers (e.g., support vector machines, k-nearest neighbor, random forests) trained using feature vectors extracted from the example data have been employed for computer-aided detection and diagnosis. Contrary to previous rule-based systems completely designed by practitioners, these machine learning approaches directly learned from the data to identify patterns and provide predictions with minimal human interventions. More precisely, the process of extracting discriminant features from the images was still performed by human researchers (i.e., handcrafted features) while the machine learning algorithms were employed to determine the optimal decision boundary in the high-dimensional feature space [2], [14].

Following this, the next logical step was therefore based on the idea of letting the algorithm directly learn the features that optimally represent the imaging data and the considered task [2], [14]. This concept is at the core of deep learning models which were developed and popularized by Yann Le Cun, Yoshua Bengio, and Geoffrey Hinton [13]. These neural networks are composed of a succession of layers that transform input data (e.g., image, video, sound, text) to output predictions while learning hierarchical representations with increasingly abstract features [13], [14]. Although initial works on deep learning started in the first half of the 20^{th} century, deep models only gained momentum in

the 2010s when technological advancement allowed them to be efficiently developed and optimized. Consequently, **the medical image analysis community has gradually adopted these methodologies and deep networks have now become state-of-the-art methods in almost all medical imaging-based applications** [2], [13], [14].

Indeed, an ever-increasing number of research studies have illustrated the numerous applications of deep learning in healthcare, including diagnosing Covid-19 from chest radiographs [15], detecting breast cancer in mammograms [88], predicting the development of neurodegenerative diseases from brain positron emission tomography (PET) [89], analyzing cardiovascular risk from ultrasound acquisition [90], assessing the morphology of abdominal organs from computed tomography (CT) scans [91], or quantifying musculoskeletal disorders from magnetic resonance imaging (MRI) [92]. Applications have also been demonstrated in identifying skin cancer lesions [93], interpreting retinal fundus images for diabetic retinopathy [94], and detecting tumor tissues in histopathological images [95]. This rapid overview of medical imaging applications illustrates **the diversity of medical image content originating from the multiplicity of imaging modalities, acquisition protocols, anatomical structures of interest, studied pathologies, and patient populations** [1], [2], [17].

Furthermore, it is worth mentioning that medical applications employing non-image data input have also emerged, such as: identifying heart disorders from electrocardiograms [96], extracting semantic information from clinical transcripts [97], summarizing doctor-patient consultations [98], or predicting 3D protein structures from amino acid sequences [99]. To solve these clinical challenges, several key technologies have been developed with the most predominant being: classification, detection, enhancement, reconstruction, registration, regression, and segmentation [1], [2]. **Deep learning enables the definition of all these technologies in a unified mathematical formalism (i.e., function approximation), which partly explains its rapid and widespread success among very diverse medical applications [1], [2]. However, despite this wide array of studies demonstrating its great potential, actual deployments of deep learning in clinical practice remains rare and its applications are still at an early stage of development [17]–[19].** Indeed, deep learning technology requires large amount of imaging resources that are inherently scare in the clinical setting due to the complexity of the medical image acquisition and annotation process. Moreover, neural networks trained on available imaging data may present poor performance on new unseen images encountered in real-world deployment, due to potential differences in acquisition settings and

image characteristics (i.e., generalization gap).

Regarding the automatic analysis of the pediatric population, the literature remains scarce due to the limited availability of pediatric imaging resources which makes it even more difficult to develop deep learning models. Indeed, one of the major challenges in pediatrics resides in the creation of large-scale imaging databases, and consequently the development of generalizable tools that could be integrated into clinical practice. In particular, the acquisition of pediatric images is hindered by the need for specialized healthcare personnel and devices, while the amount of associated labels available for training deep learning models is even more limited due to the demand for expert pediatric radiologists. Pediatric image analysis also presents unique challenges associated with the rapidly growing anatomy of children, which differentiates it from that of adults [100], [101]. **These pediatric characteristics further aggravate the challenges associated with deep learning-based medical image analysis and lead to the necessity to develop specific computerized methods, which represents the goal of this thesis. In particular, this thesis focuses on the analysis of pediatric musculoskeletal disorders** which will be introduced in the next Chapter 2 and we aim at developing novel deep learning segmentation models, whose mathematical framework will be introduced in Chapter 3. For its part, this chapter presents general background and recent trends in medical image analysis, with a final focus on challenges specific to pediatric imaging applications. In addition, this chapter also positions the deep learning methodologies developed in Parts II and III in relation to the global ongoing trends in medical image analysis.

The remainder of this chapter is structured as follows. Section 1.2 provides a rapid overview of medical image acquisition technologies, from X-ray and ultrasound (Section 1.2.1), to nuclear medicine (Section 1.2.2) and magnetic resonance imaging (Section 1.2.3). Next, Section 1.3 presents the recent trends in medical image applications and focus on the diversity of medical image domains (Section 1.3.1), multiplicity of medical image analysis tasks (Section 1.3.2) and the resulting challenges for deep learning-based analysis (Section 1.3.3). In addition, more theoretical challenges with impact on healthcare are presented in Section 1.3.4. Finally, Section 1.4 introduces the technical challenges specific to pediatric image acquisition (Section 1.4.1) and analysis (Section 1.4.2), as well as advanced deep learning techniques addressing these issues (Section 1.4.3).

1.2 Background on medical image acquisition technologies

The history of medical imaging can be traced back to the discovery of X-rays by Wilhelm Röntgen in 1895 and his experiments to image the bones in his wife's hand [10], [102]. Thereafter, European and North American physicists quickly started replicating Röntgen's research and improving his technique for visualizing the human body. At the same time, physicians began to exploit the clinical potential of radiographs which allows the study of the interior of the human body in a non-invasive manner. For instance, it was possible to assess skeletal trauma using radiographs, a well-known example being Marie Curie aiding doctors by visualizing shattered bones during World War I. One can also mention the invention of mammography in 1913 and the first cerebral angiogram in 1927 [10], [103].

However, modern medical imaging did not begin to take shape until decades later, with the development and creation of PET (in the 1950s) and ultrasound imaging devices (in the 1960s). The field was then firmly established and clearly defined during the 1970s with the beginning of new computational medical imaging techniques, namely: CT and MRI. **Nowadays, these technologies (i.e., X-ray, PET, ultrasound, CT, and MRI) represent the most commonly used imaging modalities in daily clinical routine and medical applications [10], [103].** Furthermore, the definition of medical imaging also includes other type of images such as visible light images captured with simple digital cameras that are notably employed in dermatology [93], ophthalmology [94], or even during minimally invasive surgical procedures such as endoscopic surgery [104]. However, in this chapter we only consider the most common imaging modalities that are able to capture a visual representation of the internal anatomy (i.e., X-ray, PET, ultrasound, CT, and MRI).

From a very general perspective, these medical imaging devices exploit physical phenomena such as radioactivity, electromagnetic radiation, nuclear magnetic resonance, sound waves, and their respective interaction with the internal tissues of the human body in order to generate non-invasive visual representations of both anatomy and physiology. The following brief summary strives to provide essential background information for each modality and to present the diversity within the medical imaging field. We focus here on the physical phenomena and the image generation process that characterize each modality, and provide some key clinical applications. Chapter 2 will provide clinical details on

the applications of these modalities for the analysis of pediatric musculoskeletal disorders.

1.2.1 Earliest medical image modalities: X-ray and ultrasound

The process of exposing an object to X-rays and capturing the resulting remnant beam allows to reveal its internal structures as part of the penetrating high-energy electromagnetic radiation is absorbed in a process known as attenuation [105]. The absorption of X-ray photons by denser structures such as bone will result in less exposed areas on the image receptor, which can then be easily distinguished from low-density tissues, hence the interest of X-rays in assessing bone fractures [102], [105]. Conventional radiography thus corresponds to a projection of the 3D anatomy on a 2D image, and it is sometimes necessary to acquire a supplementary X-ray from a different angle to obtain additional spatial information [105].

To address this problem, CT devices use a rotating radiation source and a row of detectors to measure X-ray attenuation from multiple angles [106]. These measurements are then processed using reconstruction algorithms to produce tomographic (i.e., cross-sectional) images of the human body and obtain a 3D representation of the anatomy [102], [106]. Because of their effectiveness, CT scans have been widely employed to screen diseases in several anatomical regions, including the head, the lungs, or the abdomen. In addition, one can incorporate temporal information by using fluoroscopic imaging, which employs a continuous source of radiation to obtain real-time moving images, which in turn, make it possible to analyze the function of the anatomical structures of interest [102]. Real-time CT techniques (or CT fluoroscopy) have thus emerged to obtain complete 4-dimensional imaging data, and have found applications in the guidance of interventional procedures [106]. It is also possible to employ radio-contrast agents which absorb radiation and consequently decrease the exposure on the detector to enhance the visibility of targeted internal structures such as blood vessels (i.e., angiogram). Nevertheless, **radiation exposure remains one of the major concern when employing X-rays and the required radiation dose must thus be kept as low as reasonably achievable (i.e., ALARA guideline), especially in pediatrics** [102].

For its part, the concept of medical ultrasound differs completely, as it operates on high-frequency sound waves to image the human body by analyzing the echo of transmitted signals over time. More precisely, sound waves propagating through the body will be attenuated and reflected at separate intervals depending on the composition of the different tissues encountered. To produce an image, the ultrasound device must determine

the propagation time and amplitude of the received echo signals in order to compute the intensity of each pixel in the image. The different reflection and transmission properties of the anatomical structures will hence result in distinct echo amplitude and pixel brightness, making it easy to distinguish the relative tissues in the image [107], [108]. Because of its effectiveness in soft tissue imaging, key applications of ultrasonography include gastroenterology, gynecology, pulmonology, or cardiology.

While ultrasound images are typically two-dimensional (2D), one can capture a series of ultrasound images and record the transducer orientation for each slice to create a 3D representation. It is also possible to employ the Doppler effect to accurately assess the direction and velocity of blood flow in arteries and veins, as in angiology. This information is usually represented as a color scale displayed on ultrasound images for simultaneous visualization of both anatomy and function. Finally, **ultrasound imaging presents several additional advantages for clinical practice: real-time acquisition to study the function of moving structures, easily flexible scanning equipment, and most importantly (especially in pediatrics), no emission of ionizing radiation**. However, compared to other imaging modalities, ultrasound may provide less anatomical detail as the depth propagation of sound waves is limited, and some structures (i.e., bone, air) reveal to be challenging to penetrate [107], [108].

1.2.2 Nuclear medicine for functional imaging

Unlike X-ray and ultrasound imaging modalities which rely on external sources to emit a physical signal interacting with human tissues, PET imaging and, more generally, nuclear medicine involves injecting radioisotopes (i.e., radionuclides) directly into the patient and recording the resulting internal radiation with gamma cameras to provide a visualization of the human body. Radionuclides are usually incorporated into molecules whose properties cause them to bind and aggregate on specific types of tissues and act as markers of metabolic use. These labeled chemical compounds, known as radiotracers, are typically administered into the body by intravenous injection or aerosol inhalation. Hence, differing from most other imaging modalities, nuclear medicine primarily focuses on studying the physiology of the system being investigated [109]. Following the administration of radiotracers, gamma cameras are then employed to capture the electromagnetic radiation emitted by the radioisotopes and generate 2D images by extracting position and intensity information from the interaction between the gamma ray and the external detectors, in a process known as scintigraphy [110].

Similar to CT, single-photon emission computed tomography (SPECT) techniques employ a rotating gamma camera to acquire distinct 2D images from multiple angles and then apply a tomographic reconstruction algorithm, yielding a 3D image [111]. Furthermore, in contrast with SPECT methods which directly measured the gamma radiation emitted by the radioisotopes, the radiotracers used in PET emit positrons particles (i.e., by positive beta decay) that annihilate with neighboring electrons causing two gamma photons to be emitted in opposite directions. The PET scanner thus collects a list of simultaneous detection of pairs of photons (i.e., coincidence events) which can be grouped into projection images referred to as sinograms. In turn, these 2D sinogram images can be converted into a 3D physiological image by employing analytic techniques similar to CT and SPECT image reconstruction algorithms [109], [110]. One key application of these functional images provided by SPECT and PET devices is to help investigating tumor or infection in multiple anatomical structures such as the lungs, heart, or bones.

Although nuclear medicine emphasizes on imaging the human function, hybrid scanning systems (e.g., SPECT/CT or PET/MR scanners) were developed to simultaneously provide anatomical and physiological information that would otherwise be unavailable or require a more invasive procedure or surgery [111], [112]. While different imaging modalities collected at separate scanning sessions can be superimposed thanks to image registration algorithms, a simultaneous acquisition using hybrid cameras offers better alignment of images and direct correlation. Nevertheless, as for radiography, the primary concern of nuclear medicine remains radiation exposure which should be kept as low as reasonably practicable, with additional vigilance for pediatric patients.

1.2.3 Magnetic resonance imaging

As opposed to CT and PET imaging modalities relying on X-rays or ionizing radiation, MRI scanners use magnetic fields and electromagnetic signals to visualize the anatomy and the physiological processes of organs inside the human body [113], [114]. Nuclear magnetic resonance-based imaging techniques rely on the physical properties of hydrogen atoms abundantly present in the human body, especially in water and fat, that are capable of absorbing and emitting radio frequency energy when placed in an external magnetic field.

More precisely, the magnetic moments of protons, subjected to the strong and uniform magnetic field of the scanner, align to be either parallel or anti-parallel to the direction

of the field, with the former corresponding to a lower energy state. Protons in parallel alignment are then excited to an anti-parallel state using an external radio frequency pulse, and a resultant electromagnetic signal is emitted as the protons return to the lower energy state by the relaxation process. The spatial positions of the emitting protons are subsequently encoded thanks to additional local gradient magnetic fields, and the 3D intensity image is finally obtained by employing the Fourier transform on the signals measured and sampled in k-space (i.e., spatial frequency domain) during scanning [113], [114].

In addition, the rate at which excited atoms return to their equilibrium state varies widely between different tissues due to distinct magnetic susceptibility, which in turn determines the relative intensity and contrast between each tissue in the resulting image. Thus, the appearance of an MR image can be directly defined by a particular setting of the pulsed sequences and gradient fields. In particular, the radio frequency signals are parameterized by their repetition time (TR) and echo time (TE), which respectively define the interval between successive pulse sequences applied to the same slice and the time between pulse emission and echo signal reception. As tissues can be characterized by two different relaxation times (T1 and T2), employing either short or longer TR and TE respectively produce T1-weighted or T2-weighted scans highlighting different anatomical structures. While T1 and T2 are the most common MR sequences employed in practice, other image sequences have been defined based on other physical properties and for a variety of clinical applications (e.g., diffusion-weighted MRI to analyze brain and neuronal activities, or proton density weighted MRI to study joint injuries) [113], [114]. As our methodologies are developed on MR modality, the advantages and weaknesses of using MR images to assess pediatric musculoskeletal disorders will be be described in Section 2.3.1.

Similar to CT, it is also possible to employ contrast agents to enhance image quality and facilitate diagnosis by altering the magnetic resonance relaxation time of hydrogen nuclei within a targeted structure. A key application of such contrast agent lies in the visualization of blood vessels and arteries (i.e., angiography). Additionally, the development of dynamic MRI has made it possible to capture a fourth temporal dimension which is essential for evaluating functional disorders. This technique is of particular interest in the study of musculoskeletal disorders because it allows the study of joints in motion. **From a general perspective, MRI provides highly resolute volumetric images without exposition to radiation which is advantageous in pediatric, nevertheless**

the long acquisition time remains a limitation. Finally, while radiation exposure does not represent a safety concern in nuclear magnetic resonance imaging, this technique remains contraindicated for patients with implants (e.g., cochlear implants and cardiac pacemakers) as the scanners are built with powerful magnets [113], [114].

1.3 Recent trends in medical image applications

Through this brief overview, we have seen that medical image acquisition relies on various technologies, each defined by specific characteristics, challenges and limitations. While additional details could have been provided for each modality, such technical considerations are out of scope here, **this summary aimed at exhibiting the richness, diversity, and heterogeneity of medical imaging in terms of involved physical phenomena, image appearance, and image dimensionality (i.e., 2D/3D, dynamic, multi-modal)**. As previously mentioned, we also refrain from considering other type of images such as images captured by digital cameras or through microscopy (i.e., histopathology) that do not provide internal anatomical representations. It should also be emphasized that novel imaging modalities have been developed, such as near-infrared [115] (in the 1990s) and magnetic particle [116] (in the 2000s) imaging techniques. However, their deployments remain limited in clinical practice, and these devices were thus omitted from this survey.

Ultimately, there is a large number of medical image applications emerging from these commonly used acquisition devices (i.e., X-ray, CT, ultrasound, PET, and MRI). Indeed, in parallel with the progress in medical image acquisition, digital image processing, and pattern recognition techniques (i.e., computer vision) have been developed and specifically tailored to the characteristics and challenges of each imaging modality [2], [10].

1.3.1 Diversity of medical image domains

In clinical routine, medical imaging devices provide multi-dimensional anatomical and/or functional information of the human body that is typically analyzed by a radiologist in a summary report and then used by a physician to define a diagnosis and treatment plan. **While each imaging modality presents particular challenges, continuous technological advancements in spatio-temporal resolution, signal-to-noise ratio, and image contrast allow for the acquisition of highly informative, dense,**

and sometimes dynamic or multi-modal imaging data [1], [10]. For instance, the spatial resolution of CT and MRI scans has reached the sub-millimeter level while ultrasound temporal resolution exceeds real-time. At the same time, novel CT and MRI reconstruction algorithms have led to increased acquisition speed and signal-to-noise ratio, while improved PET analytical tools have enabled a diminution in radiation exposure through the use of low-dose radioisotopes [1].

In turn, these improvements in image acquisition have resulted in more accurate and robust clinical decisions and prognosis, as well as safer medical procedures for the patients (i.e., lower radiation exposure). Nevertheless, the decision to employ a specific medical imaging device remains subject to a compromise between multiple parameters, such as the safety of the patient, the pathology to be examined, as well as the availability, cost and duration of image acquisition. For example, ultrasound and MRI (without contrast agents) are the techniques of choice for medical imaging during pregnancy, as these instruments are not associated with any degree of exposure to ionizing radiation that, at high doses, could result in miscarriage or congenital disabilities. Ultrasound is favored in clinical routine for its ease of use, cost effectiveness and rapid acquisition time. Nevertheless, MRI allows clinician to evaluate the fetal brain with greater detail and detect abnormalities not visible on ultrasound images. In some scenarios, it also may be needed to employ more than one imaging modality to exploit complementary information. However, obtaining images from a secondary modality may be difficult due the limited availability and high cost of scanners. Moreover, as each imaging modality is characterized by specific physical interactions, some happen to be more adapted to visualize specific anatomical structures. For the diagnosis of Covid-19, chest X-rays and CT scans have been preferred over ultrasound examination of the chest, which is difficult to interpret as sound waves reflect strongly at the air-tissue boundaries of the lungs [1].

Furthermore, these distinct physical phenomena (i.e., radioactivity, electromagnetic radiation, nuclear magnetic resonance, and sound waves) induce specific image characteristics (e.g., intensity distribution, type of noise, contrast between tissues, artifacts) for each modality. For its part, the type of scanner (e.g., manufacturer, model) and acquisition protocol (e.g., contrast agents, MRI sequence, scanner settings, patient positioning) determine the image resolution, intensity distribution, signal-to-noise ratio, and contrast, as well the position of the anatomical structures within the image. It should be emphasized that the dimensionality of imaging data can also vary, although typically 2D or 3D, certain devices enable the acquisition of temporal information while hybrid scanners pro-

vide multi-modal images. When this large number of acquisition parameters is combined with the variety of structures to be imaged (e.g., brain, heart, lungs, abdomen, bones), this results in a great variability of medical image appearance and content. Furthermore, patient characteristics (e.g., gender, age, ethnicity, pathology) can introduce additional image variations in appearance and content resulting from the difference in anatomies, such as the divergence in bone morphology between infants and adults or the prevalence of lung nodules in urban populations compared to rural ones. **Finally, each of these combinations between imaging modality, type of scanner, acquisition protocol, targeted anatomical structure, and patient population defines a distinct medical image domain corresponding to a specific mathematical intensity space.** The diversity of medical image domains represents one of the main characteristics of medical imaging (Figure 1.1).

1.3.2 Overview of medical image analysis tasks

The applications of medical imaging found in clinical practice are numerous and diverse, reflecting the large number of diseases and pathologies present in the human population. As it is nearly impossible to present an exhaustive list, one can provide a few critical applications: lung cancer diagnosis from chest CT/PET scans [117], prediction of Alzheimer's disease from PET neuroimaging [89], quantitative analysis of cardiac vessels from CT angiography [118], lesions detection in abdominal organs from CT or MR abdominal images [119], fetal heart rate monitoring using Doppler ultrasound [120], and evaluation of (pediatric) musculoskeletal disorders from MR images [40]. In all cases, the analysis of the acquired medical images is performed by radiologists whose interpretation can be limited due to human subjectivity, fatigue, and variability among experts [1]. Moreover, radiologists have limited time to review an ever-increasing number of examinations, which leads to missed findings, long turn-around times, and a paucity of numerical results and quantification (i.e., a low amount of expert annotations). In turn, this drastically reduces the ability of the medical community to advance towards more evidence-based personalized healthcare or to analyze large-scale population studies [1], [11].

The development of computerized methodologies to provide automated medical image analysis thus reveals essential to assist physicians in making faster and more reliable diagnoses and clinical decisions. Specifically, **computerized medical image analysis consists of an array of technologies, including reconstruction, enhancement, detection, classification, regression, segmentation, and registration** that arise

from the needs of each medical imaging application [2]. For instance, a lung cancer diagnosis can be formulated as an image classification task, while the quantitative analysis of cardiac vessels or bone morphology can be based on segmentation tools. A brief overview of these medical image analysis tasks is provided below:

- Reconstruction algorithms, as previously mentioned, aim at forming an image from signals acquired by a medical imaging device (e.g., ultrasound, CT, PET, or MRI) so that it can be visually interpreted, and consequently appear first in the image processing pipeline [11]. While reconstruction of high-quality images is primordial for further analysis, enhancement methods can also be employed to help adjusting the intensities of an image via denoising, super-resolution, or harmonization.
- Detection, classification, and regression constitute the main tools to help clinical diagnosis. Detection aims at localizing an object of interest in an image (e.g., tumor, metastasis) by providing a bounding box that frames the structure [2]. Classification and regression tools are then usually applied to either classify the detected structure among predefined categories (e.g., lesion types, benign or malign tumor) or predict a continuous value characterizing the localized structure (e.g., tumor or lesion size). Classification and regression techniques can also estimate discrete (e.g., healthy or impaired patient) or continuous (e.g., degree of impairment) properties defined at the scale of the whole image.
- Segmentation is considered as the "holy grail" of medical image analysis as it aims at providing fine-grained information. Segmentation is the process of assigning a label to every pixel in an image so that pixels with the same label share certain characteristics, thus resulting in multiple segments corresponding to distinct structures [2], [121]. Unlike detection, classification, and regression, segmentation is a low-level (i.e., pixel-wise) task that provides quantitative information about the volume and morphology of targeted anatomical structures, essential for intervention and surgical planning. Deep learning-based segmentation and its mathematical formalism will be introduced in more details in Chapter 3.
- Medical image registration aims at aligning the spatial coordinates of one or more images into a common coordinate system, which reveals useful for performing population-based analysis, longitudinal analysis, and multi-modal fusion [2].

Following these definitions, it appears that each task is associated with a specific type of predicted output defined in a distinct mathematical space. Typically, segmentation maps are multidimensional objects, while a classification label can be formulated as

a unique integer value. Furthermore, depending on the targeted anatomical structures and patient population characteristics, the labels may differ for a given type of task. For instance, from the same domain of abdominal MR images, one can target separate abdominal organs (e.g., liver, spleen) for segmentation or classify different lesions (e.g., tumor, metastasis, cyst) associated with distinct pathologies. The multiplicity of image analysis tasks is also a key feature of medical imaging (Figure 1.1).

Ultimately, with large-scale imaging datasets and sufficient corresponding annotations, it has been illustrated that deep learning networks can achieve accurate and robust performance, for a specific task and in a dedicated medical imaging domain [2]. **Hence, the major challenge of computerized medical image analysis resides in the multiplicity of imaging acquisition protocols and medical applications, making it difficult to develop tools applicable in diverse clinical scenarios** (e.g., across different hospitals) [1]. In particular, deep learning models tend to be task and domain-specific (i.e., "specialists" model) and necessitate dedicated training data which may be not readily available or only available in limited quantities [21]. While our work focuses on the segmentation of pediatric musculoskeletal images (see Chapters 2 and 3), we believe that it is important to provide the general limitations of deep learning-based approaches for medical image analysis. Because these limitations are inherent to the data-driven nature of deep learning, they are present in all clinical applications regardless of the image domain and task, and hinder the deployment of deep learning technologies in real-world scenarios.

1.3.3 Technical challenges of deep learning-based medical image analysis

Deep learning is a machine learning method based on artificial neural networks. In particular, it is a representation learning technique that aims to automatically learn the most relevant features of the image domain and task at hand [13], [14]. As mentioned in Section 1.1, deep learning methodologies contrast with traditional machine learning approaches based on handcrafted features. The principle of deep learning techniques will be detailed in Chapter 3.

Although deep learning has been successfully applied and proven to outperform previous machine learning methodologies in numerous medical image applications, the deployment of deep learning solutions in real-life scenarios remains scarce [17]. **At its core,**

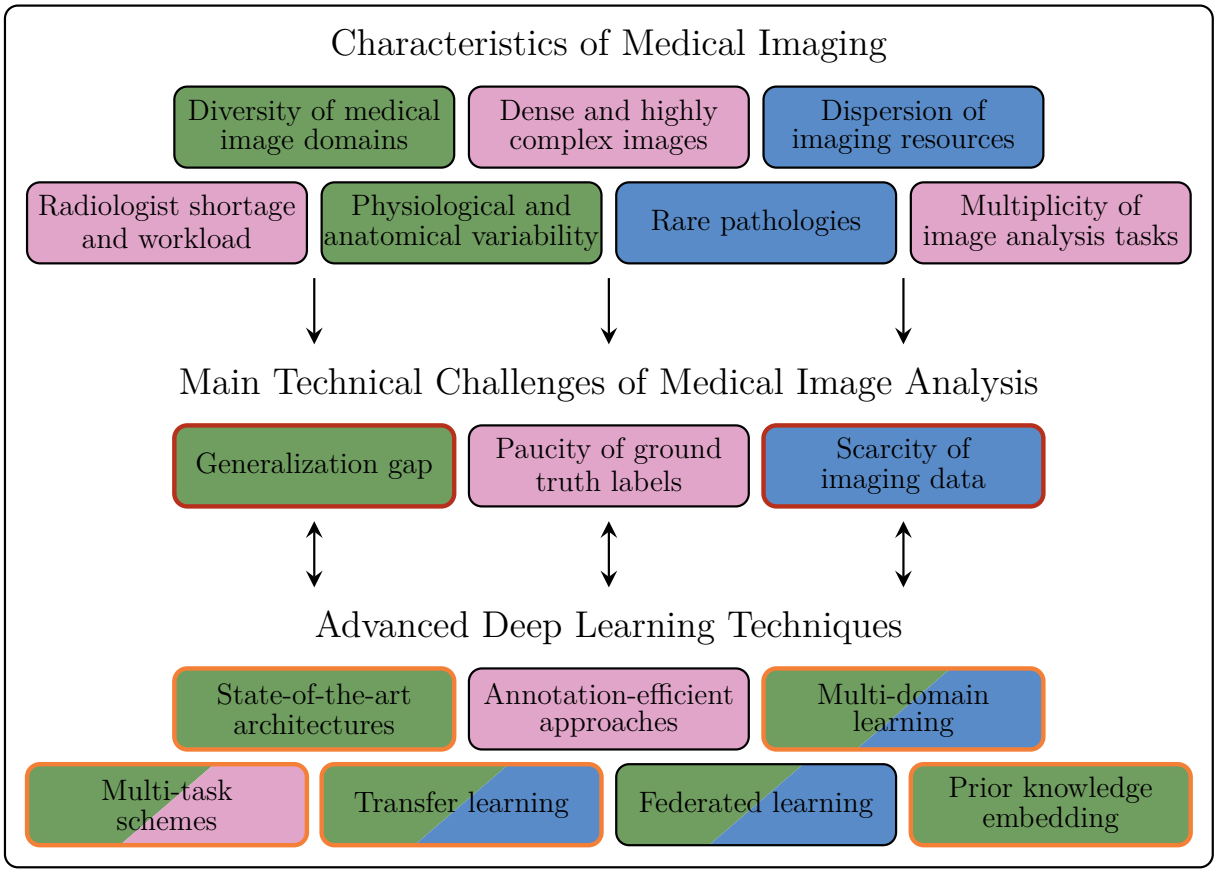


Figure 1.1 – Summary of the main characteristics of medical imaging, the major technical challenges associated with medical image analysis (based on deep learning), and the advanced deep learning techniques developed to address these issues. Each technical challenge is characterized by a specific color: the generalization gap in green (—), the paucity of ground truth label in pink (—), and the scarcity of imaging data in blue (—). This thesis targets the generalization gap and imaging data scarcity issues, circled in red (—) and the deep learning techniques employed in work are highlighted in orange (—).

deep learning is a data-driven technology requiring large amounts of training data, and optimized models tend to be task-specific and display poor domain generalization capabilities on unseen imaging domains. Meanwhile, medical imaging is characterized by a wide variety of acquisition settings (see Section 1.3.1) leading to a problem known as distribution or domain shift, which is defined by the difference in data distribution between the training dataset and data encountered during deployment (i.e., different medical image domains, Figure 1.1). This shift in data distribution is multifactorial and may be due to, among other things, differences in acquisition parameters or differences in the imaged population [17]. One could hope that collecting large and repre-

sentative medical imaging datasets could solve this issue, however such approach is highly challenging given the diversity of imaging protocols, imaging devices, and patient populations [21]. In addition to the lack of standardized acquisition protocols, the dispersion of imaging data among different hospitals and imaging centers, as well as data privacy and clinical data management requirements, make it challenging to construct large-scale open-source medical imaging databases (Figure 1.1) [1]. For its part, the data annotations process is all the more costly and laborious in the context of large-scale dataset, as multiple experts are usually involved. In particular, it is preferable to define a standardized annotation protocol beforehand to limit labeling inconsistency.

It should be emphasized that works addressing the domain shift observed in clinical routine are part of an emerging field known as domain generalization or out-of-distribution generalization. Indeed, a common hypothesis in traditional deep learning is that the training and test data originate from the same data distribution. On the contrary, domain generalization methods propose to address the more challenging setting in which the goal is to learn a model that can generalize to an unseen test domain [72], [74]. In this thesis, we will refrain from considering this distribution shift issue between training and test data and we position our work in a more traditional deep learning setting. **The "standard" generalization gap, which is defined as the difference between the model's performance on training data and its performance on unseen images drawn from the same distribution, still represents a current limitation of deep learning models in medical imaging, especially when associated with the scarcity of imaging data** (Figure 1.1) [1], [2]. Indeed, deep models can present poor performance on test images presenting features (e.g., highly irregular deformities, lower contrast between tissues) different from the ones observed in the training set. From a more mathematical perspective, neural networks can learn to approximate arbitrary functions, but only in the manifold where there is enough density of training data. This notion of density and therefore of distance between data point is extremely challenging to define in medical imaging, where images are highly dimensional objects (i.e., number of pixels typically in the order of thousands or millions). In fact, this problem is closely linked with the "curse of dimensionality", which states that the amount of data needed to approximate a function grows exponentially with the dimensionality. However, a common hypothesis is that medical images lie on a lower-dimensional manifold, which is supposed to 'break' the curse of dimensionality [28]. Indeed, medical images present common features and shared characteristics due to the relative constrained nature of the human anatomy. Hence, it is

crucial to construct imaging datasets with sufficient variability to obtain enough samples in the manifold (and prevent over-fitting), which may be highly challenging with limited imaging resources. In this direction, an ideal model could adapt to the diversity of imaging domains as well as the variability within each domain.

For its part, **optimization of deep networks can be hindered by the lack and poor quality (i.e., paucity) of available labels associated with medical images** (Figure 1.1). As previously mentioned, annotating medical images is a time consuming and expensive process even for moderately sized medical imaging datasets [21]. Because of the variability in experience and environment, both inter-user and intra-user labeling inconsistency is high, and labels must therefore be considered noisy [1], [2]. Moreover, different types of annotations exist in practice, reflecting the variety of feasible tasks, with pixel-wise annotations (i.e., segmentation mask) being extremely tedious to create compared to global image labels, especially when considering the continuous improvements in spatio-temporal resolution [1]. It is also important to note that segmentation is an inherently ambiguous process and a group of annotators typically produces a set of plausible but diverse delineations. It is thus essential that neural networks can integrate and provide such uncertainty in their segmentation predictions.

An additional challenge specific to medical image analysis resides in the long-tailed distribution of diseases. While a small number of common pathologies have sufficient observed cases for large-scale analysis, most disorders are rare in clinical routine. It is also important to note that novel contagious diseases that are not represented in the current ontology may arise (i.e., Covid-19) and thus require the development of new dedicated tools [1]. In particular, **it is highly challenging to adapt data-hungry deep learning models to detect and diagnose rare diseases whose low prevalence may limit data collection to very few subjects** (Figure 1.1) [21]. Furthermore, each pathology presents a large variation among the affected population in terms of severity and development, thus resulting in a great heterogeneity of image content. The population, as a whole, also exhibits a significant diversity due to considerable variation in anatomy and function between individuals. All of these, therefore, reinforce the generalization gap observed in practice (Figure 1.1) and confirm the need to incorporate such variability in deep learning models [1], [2].

The attempts of the medical image community to collect large and representative datasets could represent a solution to solve these issues caused by the data-driven nature of deep learning (Figure 1.1). One primary example of such a large-scale publicly avail-

able dataset is the multimodal brain tumor segmentation challenge (BRATS) dataset [1] [122]. This dataset is composed of multimodal (i.e., T1 and T2) MR scans, acquired from multiple international institutions, and the corresponding expert annotations of brain tumors (i.e., glioblastoma and lower grade glioma). The BRATS dataset represents hundreds of volumetric MR images and associated labels acquired and curated from several imaging centers using multiple scanners. Hence, one could believe such a dataset to be sufficiently representative of the variability encountered in real-world deployment. Deep learning models trained on the BRATS dataset show overall great performance (e.g., $\approx 90\%$ Dice), which might indicate that for this specific dataset, automatic segmentation with per human-level performance is almost reached. We will introduce the metrics, such as Dice, used to assess quantitatively the performance of segmentation models in Section 3.5.2. However, it remains an open question whether these models could generalize well on unseen domains not present in the training set and hence be actually deployed in clinical practice. It should also be emphasized that networks trained on BRATS can solely perform one task: glioblastoma and lower grade glioma segmentation. Furthermore, such attempts at building large-scale datasets are highly expensive as it requires the collaboration of numerous international experts and research centers. It also appears infeasible to build such a large dataset for every pathology encountered in clinical practice due to the low prevalence of some rare diseases.

Hence, **to address the scarcity of imaging data, the paucity of associated imaging labels, and the generalization gap (Figure 1.1), several advanced methodologies and paradigms have emerged in the deep learning field**, including state-of-the-art network architectures (e.g., recurrent layers, attention mechanisms) [1], [2], annotation-efficient approaches (e.g., self-supervision, semi-supervision, weakly-supervision, zero/few shot learning, active learning, interactive learning, synthetic data augmentation) [1], [2], [20], [21], multi-domain learning (e.g., multi-modal, multi-anatomy, multi-site, multi-scanner, domain adaptation) [2], [21], multi-task schemes (e.g., multi-anatomy segmentation, detection and classification, enhancement and segmentation) [21], transfer learning (e.g., from large natural or medical image databases) [20], federated learning (e.g., via distributed computing, model aggregation strategies) [1], and prior knowledge embedding (e.g., via a regularized loss, directly encoded in the architecture) [1]. All these approaches have been demonstrated to improve performance over more traditional and conventional deep learning models. It should be emphasized that this list is

1. `http://braintumorsegmentation.org/`

not exhaustive. One could also mention other approaches such as those exploiting clinical information present in medical reports in conjunction with medical imaging (i.e., learning shared text and image representation) to address the paucity of pixel-wise annotations.

As depicted in Figure 1.1, this thesis will only focus on some advanced deep learning techniques which will be presented and employed in the remainder of the thesis, as follows: state-of-the-art network architectures (Chapters 5 and 7), multi-domain learning (Chapters 6 and 7), multi-task schemes (Chapters 6 and 7), transfer learning (Chapters 5 and 7), and prior knowledge embedding (Chapters 4, 5, 6, and 7). Therefore, **our methods are primarily aimed at addressing the generalization gap and data scarcity issues**.

1.3.4 Theoretical challenges with impact on healthcare

Before introducing the technical challenges specific to pediatric image analysis, it is worth mentioning that there also exist more general theoretical challenges in deep learning that may significantly impact medical image analysis. Indeed, due to a complex succession of layers, **the black-box nature of deep learning models makes it difficult to analyze and understand how the decision process is conducted and how the inputs affect the final output prediction**. In healthcare, it is particularly essential to ensure that deep learning decisions are driven by clinically relevant features and to understand possible failure of network predictions. In this direction, methods and studies enabling the explanation of neural network decisions (i.e., explainable or interpretable deep learning) have thus emerged into an active area of research [123]. It should be noted that the distinction between explainable and interpretable model is not yet clearly defined in the literature, the former generally relying on a post-hoc step to analyze the model while the latter make the decision inherently explicit. In parallel, one can also mention techniques aiming to understand and quantify the uncertainties in the predictions generated by neural networks [124]. Indeed, it is essential that clinician can access the confidence and reliability of the model's predictions for future diagnosis.

From a more general perspective, **one of the major theoretical challenges of deep learning resides in the lack of a rigorous mathematical formalism to design, optimize and analyze deep learning architecture in a principled way**. A notable example that attempts to address this problem is geometric deep learning, which proposes to build a common mathematical framework to study neural networks by analyzing pre-defined regularities through unified geometric principles. If successful,

such novel mathematical formalism could redefine deep learning as a whole and modify the artificial intelligence landscape, with obvious implications for computer vision and medical image analysis [125]. Nevertheless, we refrain from considering these emerging peripheral topics in this thesis, and focus on challenges specific to pediatric image analysis which is, as presented in the next section, the context of our research work.

1.4 Technical challenges specific to pediatric imaging applications

The analysis of pediatric pathologies faces all of the mentioned technical problems (see Section 1.3), which are exacerbated by the pediatric status of the considered population, especially the scarcity of imaging resources, as the acquisition of pediatric imaging is an even more challenging process than for adult cohorts. In addition, unique pediatric disorders may necessitate the development of dedicated tools, while the smallness, thinness, and continued growth of children's anatomical structures can make detection and identification difficult [100], [101].

1.4.1 Pediatric image acquisition

Imaging represents an extremely valuable diagnostic tool in the pediatric population, but it is followed by a number of distinct challenges as compared to image acquisition in adults. **Pediatric imaging requires, among other things, dedicated acquisition protocols, specific training for the healthcare personnel involved, and extensive knowledge and expertise in pediatric anatomy and physiology for accurate image analysis [22]–[27].** In clinical practice, pediatric diagnosis relies on the same image modalities and acquisition technologies as for adults (see Section 1.2). However, specific devices are sometimes needed to adapt to the morphology of children, for example, to maintain infants or toddlers in position to improve image quality (e.g., chest X-ray).

Most importantly, protection and safeguarding against radiation exposure are paramount in this age group, as children have a greater risk for the manifestation of possible harmful effects of radiation. Therefore, the ALARA principle (i.e., As Low As Reasonably Achievable) should be strictly followed, and appropriate imaging modality should be used depending on the clinical indication (e.g., using ultrasound instead of CT in a suspected case of appendicitis). Moreover, numerous recommendations exist to produce qualitative

pediatric images using low radiation dose, typically consisting of specific settings for X-ray/CT equipment such as reduced exposure time or detector coverage proportional to body size. In functional imaging, designing novel PET radiotracers enabling lower radiation dose as well as employing hybrid PET/MR scanners are active research topics for pediatrics [23], [24].

Because ultrasound and magnetic resonance imaging modalities emit no ionizing radiation, these technologies are considered safe for use in pediatrics, especially for longitudinal follow-up [25]. Nevertheless, the smaller and thinner anatomical structures of children can create a challenge in terms of available signal and image resolution. For instance, a higher signal-to-noise ratio is needed in MRI, which can be achieved using pediatric specific coils, high field strengths and by optimizing the field of view (FOV) and slice thickness. It is also worth mentioning that the major challenge with lengthy MRI acquisition procedures, as opposed to rapid ultrasound examinations, is the need for sedation or general anesthesia in younger children, which can involve further health complications and stricter ethical considerations [26].

For clinical care workers, specific training is crucial to manage all of these pediatric image acquisition procedures as well as to obtain children's cooperation before and throughout the duration of an examination in order to guarantee the acquisition of qualitative images and prevent repeated scans [23]. Despite all these acquisition challenges, pediatric imaging is rich in content, and numerous medical image-based applications have been deployed in practice. We will see in Section 2.3.1 the advantages and weaknesses of each imaging modality for the analysis of pediatric musculoskeletal disorders.

1.4.2 Difficulties in pediatric image analysis

Although providing an exhaustive list of pediatric medical image applications is out of scope here, notable examples include quantification of the development of post-treatment myocardial ischemia from cardiac PET/CT scans [22], assessment of knee joint injuries in young athletes from MR images [126], presurgical localization of seizure focus using PET/MRI [127], diagnosis of soft-tissue vascular anomalies and lesions from MR scans [128], and detection of hemodynamic abnormalities associated with congenital heart diseases using Doppler echocardiography [25]. These pediatric disorders may be acquired or congenital, and while most are also present in the adult population, it should be emphasized that **the specificity of the pediatric anatomy and physiology results in unique pathological patterns and a distinct appearance of pediatric images**.

Moreover, as the entire body of children is rapidly developing, this can lead to medical image applications specific to pediatricians, such as the monitoring of bone growth from X-rays [100] or brain development from MRI scans [101]. Nevertheless, as with adult imaging, automated analysis based on computing tools is crucial to help physicians provide a more reliable diagnosis. Radiologists reporting pediatric cases must have in-depth knowledge and expertise of the mechanisms modifying the human anatomy from early childhood to adolescence, and the pathologies afflicting children that can differ from those in adults or be unique to the pediatric population [100], [101]. Specifically, the developing anatomy can have many healthy variations that must be distinguished from pathological development.

Like in the rest of the medical image research community, deep learning has become the standard computerized method for pediatric image analysis. We have already presented in Section 1.3.3 **the major challenges associated with deep learning for medical image analysis (i.e., the scarcity of imaging data, the paucity of associated imaging labels, and the generalization gap)**, and these technical problems **are heightened when considering the pediatric population**. For instance, as pediatric imaging typically involves specific protocols and specialized healthcare personnel, these requirements can limit the number of acquisitions performed in practice. Furthermore, ethical considerations (e.g., sedation) and the relatively small-scale of children cohorts compared to adult studies can also impact the scarcity of pediatric images. It is generally more difficult (and sometimes even impossible) to recruit healthy child volunteers than to recruit adults through an ethics committee. For its part, the undersupply of radiologists with expertise in pediatric anatomy and function can aggravate the paucity and quality of associated labels [26], [27].

Furthermore, **the many healthy variations observed in the pediatric anatomy may induce poor generalization performance on unseen images**. For instance, let's consider the bone ossification process observed during childhood, which typically results in altered bone intensity in MR images. In the context of segmentation, deep models may produce poor bone delineations if the test sample presents a delayed ossification stage (linked with modify intensity patterns), as compared with the pediatric cohort used for training. Finally, it should be emphasized that due to the heterogeneity in anatomy and function between age groups, images acquired in infancy typically present extremely different features than images extracted from adolescent cohorts. Consequently, models trained on childhood imaging data would generally fail and be inapplicable to the analysis of adolescent images, each corresponding to a distinct image domain (see Section 1.3.1).

1.4.3 Advanced deep learning techniques for pediatric image analysis

Because of all these intrinsic technical challenges, the deployment of deep learning tools for pediatric image analysis remains rare in clinical practice. As the generalization gap, the paucity of ground truth labels, and the scarcity of imaging data issues are exacerbated in pediatrics, advanced deep learning techniques (see Section 1.4.3) appear essential to build image analysis tools applicable in real-world scenarios. To the best of our knowledge, these novel approaches have rarely been applied to pediatric image analysis, with the exception of multi-task scheme in combination with semi-supervision for early prediction of neurodevelopment [129], multi-center model with Inception modules for pediatric bone age prediction [130], transfer learning and Xception models for pediatric otitis media classification [131], and quality assessment of pediatric MR images via semi-supervised models [132]. **Hence, this thesis proposes to develop and apply computerized techniques belonging to these novel deep learning paradigms for the analysis of the pediatric musculoskeletal system.**

Therefore, the segmentation methodologies developed in this thesis rely on state-of-the-art architectures, multi-domain learning, multi-task scheme, transfer learning, and prior knowledge embedding (see Parts II and III). **Most significantly, our methodologies leverage multi-anatomy learning and shape priors regularization integration.** Indeed, multi-anatomy learning appears as a natural way to combine information about human anatomy acquired for different purposes and to leverage features across datasets in order to build more powerful and robust models, as well as to drastically reduce the cost of curating task-specific datasets [21]. For their part, shape priors have been studied extensively since the work of Kendall et al. [133], and have found numerous applications in medical imaging due to the constrain nature of anatomical objects. Nevertheless, deep learning-based shape priors slightly differ from the traditional definition given in Kendall et al. [133], as we will see in Chapter 4.

1.5 Conclusion

This first chapter provided a **general and historical background for medical imaging to allow a better understanding of the current issues facing medical image analysis in the age of deep learning**. The main technical challenges that hinder

the development and deployment of deep learning solutions in clinical practice can be grouped under the following concepts: the generalization gap, the paucity of ground truth labels, and the scarcity of imaging data. To address these challenges, novel paradigms and advanced deep learning techniques have emerged, which appear all the more essential for the analysis of pediatric images. Indeed, as illustrated in this chapter, pediatric imaging acquisition and analysis characteristics reinforce these technical challenges, particularly the scarcity of imaging resources.

In this thesis, we propose to develop methodologies following recent paradigms in the context of pediatric musculoskeletal image analysis. In particular, we aim at addressing the generalization gap and data scarcity issues which are aggravated in pediatric. The clinical motivations for the study of the pediatric musculoskeletal system are introduced in subsequent Chapter 2, while Chapter 3 introduces the mathematical formalism and backbone neural networks used for deep learning-based medical image segmentation.

Chapter 2

Analysis of the pediatric musculoskeletal system

2.1 Introduction

The human musculoskeletal system refers to the organ system that allows the body to move, support itself, and maintain stability during locomotion. This complex system is constituted of various tissues, including bones (i.e., skeleton), muscles, cartilage, tendons, ligaments, and other connective tissue that support and bind tissues and organs together. From a global perspective, **the skeleton serves as a framework for tissues and organs, giving shape to the body and protecting vital organs, while the primary function of skeletal muscles is to enable movement**. It is worth mentioning that cardiac and smooth muscles, found in the heart and abdominal organs, are not part of the musculoskeletal system but are instead controlled by the automatic nervous system. For their part, bones are covered and protected from directly rubbing against each other at musculoskeletal joints by articular cartilage, a resilient, smooth, and viscoelastic tissue. Cartilaginous tissues, also present in other structures such as the ears, nose, ribs, or intervertebral discs, are softer and more flexible than bones, while being stiffer and more rigid than muscles or tendons. At last, tendons are tough, flexible fibrous tissues connecting muscle to bones, whereas ligaments are small, dense, elastic fibrous tissues that tether bones together. Both tendinous and ligamentous tissues are primarily composed of collagen fibers, giving them mechanical resistance to stretching. Similarly, one can also mention fasciae which are bands of connective tissues (i.e., primarily collagen) that attach, stabilize, and separate muscles and other internal organs [134].

An anatomical joint defines an articulation between bones allowing various degrees and types of movement, and is characterized by the number and shapes of the articular surfaces as well as the type of connecting tissue [134]. **This chapter focuses on synovial joints that permit free movement between the articulating bones at the**

point of contact. There exist several types of synovial joints (e.g., condyloid, bicondyloid, spheroid), each allowing different types of movements among abduction, adduction, extension, flexion, and rotation. In a simplified manner, movement originates from muscle contraction stimulated by a motor neuron, and the resulting force is then transmitted to the connected bone via the tendon, in turn resulting in movement of the corresponding joint. Ligaments and tendons thus maintain joint stability and control joint range of motion limits. Meanwhile, articular cartilage protects bone and allows smooth contact between bones during movement, thanks to its viscoelastic properties. Specifically, cartilage transmits and distributes the mechanical forces when the joints are solicited. It should be noted that the provided description of the anatomical joint and the role of its constituents is very simplified [134]. For instance, one should distinguish between positional tendons that position limbs and energy-storing tendons that act as springs to make locomotion more efficient.

We illustrate the definition of anatomical joints with the knee as an example. The knee consists of four bones (femur, fibula, patella, and tibia), and is composed of two joints: the tibiofemoral joint and the patellofemoral joint (Figure 2.1) [134], [135]. Both permit flexion and extension of the leg, as well as slight rotations movements, which are all controlled by muscles and enabled by tendons present in the leg. For instance, the quadriceps femoris muscle, which covers the front and sides of the femur, is a powerful extensor of the knee joint and is crucial for walking, running, jumping, and squatting. The quadriceps insert into the tibia via the patella, where the quadriceps tendon becomes the patellar ligament. During movement, cartilage present at the end of long bones (femur, fibula, and tibia) and surrounding the patella ensure supple knee movement, while several ligaments (e.g., anterior cruciate ligament connecting the femur and tibia) help in stabilizing the joint [134], [135]. As previously mentioned, there exist other structures (e.g., meniscus, bursae, articular capsule) supporting movement that we will refrain from considering here.

From a historical perspective, the study and analysis of the musculoskeletal system originated with the ancient Greek and Roman philosophers. In particular, the Greek physician Galen (2^{nd} century) was the first to describe the muscle system as a complex but unified organ of locomotion and to define the brain as the center of the neuromuscular system. Most notably, his studies on muscle contraction, based on the dissection of animals, laid the foundations of muscle mechanics. Galen's views dominated and influenced Western medical science for more than a millennium. Indeed, very few medical advances were achieved during the Middle Ages, as medical research was generally discouraged.

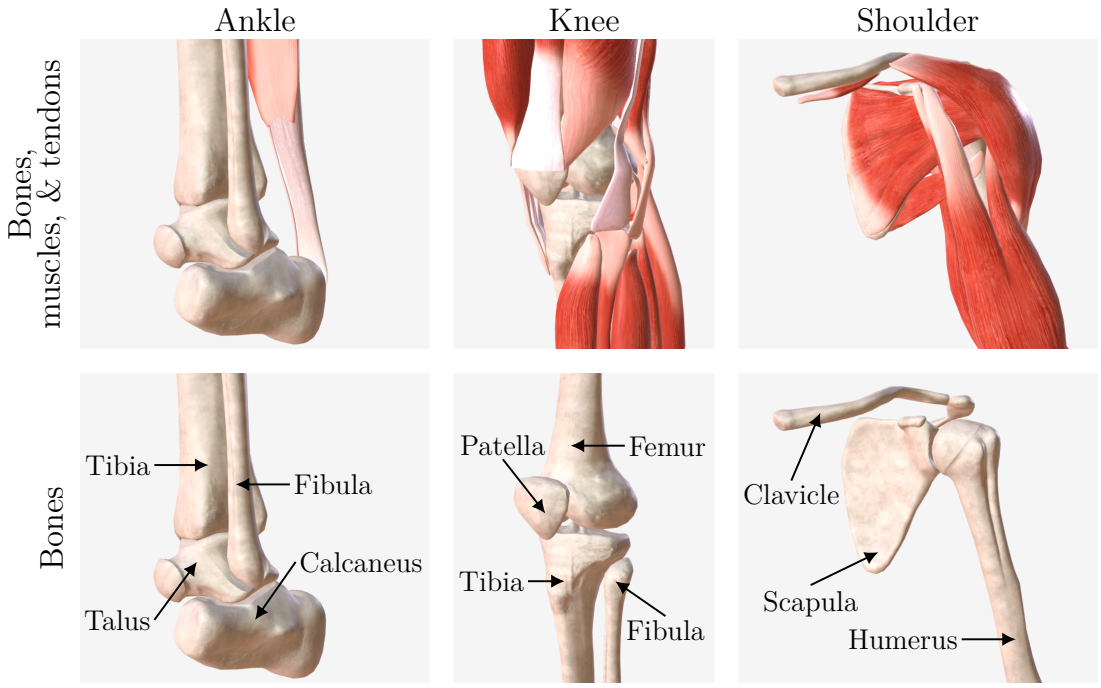


Figure 2.1 – Anatomical representation of the ankle, knee, and shoulder joints. Images extracted from the BioDigital Human Plateform™ visualization software: `https://human.biodigital.com/`. Visualization of bone, muscle, and tendon tissues using the adult male anatomical model.

During the Renaissance, scientific approaches were again accepted in medical research, with whole-body dissections and experimental approaches more commonly performed. During this period, two anatomists made lasting contributions to the current understanding of muscles: Leonardo Da Vinci and Andreas Vesalius. Indeed, Da Vinci's sketchbook, "Anatomical Manuscript B" (1511), contained at the time the most comprehensive physical description of the human body, with practically every muscle being reported and drawn. For its part, Vesalius' book on human anatomy, "De Humani Corporis Fabrica Libri Septem" (1543), was a significant advance in the history of anatomy over the long-dominant work of Galen. For instance, Vesalius studied the link between nerves and muscle action. From the 17$^{\text{th}}$ century until today, many medical advances resulted from progress in chemistry, physics, and mathematics. Numerous scientists have provided an increased understanding of the musculoskeletal system. One can mention the work of Albrecht von Haller which showed nerve impulses to be a physiological reaction separate from but controlling muscle contraction. In orthopedics, Jacques Mathieu Delpech examined the role of muscle and ligaments in joint stability and proposed a surgical process (i.e., tenotomy or tendon lengthening) to correct contracture abnormalities. Finally, Guillaume Duchenne

and Jean-Martin Charcot, who established the field of neurology, discovered that the nervous system controlled purposeful muscle movements. The understanding of musculoskeletal structures and metabolic processes further progressed during the second half of the 20th century, mainly through muscle-tendon modeling [136] and movement simulations [137] that became fundamental tools. **Nevertheless, comprehension of the musculoskeletal system is incomplete especially when considering the pediatric anatomy whose growth can induce many healthy variations among bone and muscle structures. Similarly, the origins of neuromuscular disorders and their treatments remain mostly undetermined.**

Following this brief overview, it appears essential to distinguish morphological analysis from physiological analysis: the former focuses on form and structures (e.g., shape of a muscle or bone, components of a joint) while the latter is concerned with function (e.g., biomechanical interaction during movement). Morphology and physiology are extremely close and interlinked, due to the evident relationship between form and function. Indeed, studies have reported that joint congruence, defined as the morphological adequacy of an articular surface with the opposing surface, is in part determined by its motion. Conversely, the morphology of the bones forming the joint can constrain their relative position, and thus regulating the joint's range of motion and function [138]–[140].

For its part, anatomical analysis allows to describe and characterize organs through, for instance, simple anatomical measurements (i.e., organ length and orientation) [141]–[143], regions of interest (i.e. muscle insertion area) [144], or global morphological characteristics (i.e., quadratic surface fitting) [145]. In turn, these descriptors help to compare the shape of musculoskeletal tissues (typically bones) and have proven useful in distinguishing healthy from pathological shape as well as guiding surgeons during the pre-planning phase. In this direction, statistical shape models (SSMs) are a popular tool to represent the shape distribution of bones and to perform morphological analysis of osseous structures. In particular, SSMs can model the shape correlation between distinct part of a bone (i.e., distal and proximal femur), as well as between multiple bones from the same joint (i.e., femur and tibia) [138], [146]. Therefore, such approaches could help understand morphological modifications occurring in the joint following an injury or a pathology.

As presented by Unal et al. [147], musculoskeletal biomechanics enables the understanding of the behavior of the musculoskeletal system under external and internal forces that are applied during movements. From a general perspective, biomechanics consists in the study of biological systems (e.g., cardiovascular or musculoskeletal systems) based

on the applications of the fundamental principles of mechanics, including statics, dynamics, deformable-body, and fluid mechanics. For instance, in kinematic analysis, Newton's laws, the work-energy relationship, and the principle of energy conservation can be used to document the relationship between force, moment, and motion. One can also perform an analysis at the material level to obtain stress and strain relations which reveal material properties such as elasticity, resilience, strength, and toughness. In turn, this information provides an assessment of the resistance of the material to trauma or fatigue. However, the analysis of biological tissues (in our case, bone, muscle, tendons, ligaments, cartilage) can be hindered by, among other things, their complex viscoelastic, composite, and anisotropic properties. Nevertheless, musculoskeletal biomechanics is highly relevant for predicting physiological ranges of forces acting on musculoskeletal tissues during daily activities, especially for post-surgery rehabilitation. Such analysis allows to design novel implants with specific materials, dimensions, and positioning to improve patient rehabilitation outcomes. Finally, in the context of pediatric musculoskeletal disorders, musculoskeletal biomechanics enables the study of abnormal gaits and movements resulting from muscle weakness or impaired bone growth [147].

Most pathologies affecting the musculoskeletal system result in reduced joint range of motion and impaired movements. Typical injuries caused by traumatic events encompass bone fracture, ligament rupture, muscle tear, or joint dislocation, while disorders such as cartilage wear (i.e., osteoarthritis) and tendon inflammation (i.e., tendinopathy) can arise from sudden exertion, repeated motion, or previous injuries, and result in pain and swelling [148], [149]. These injuries and disorders have a lasting impact on the musculoskeletal system and can be associated with other medical complications. As previously stated, understanding these pathologies is thus essential and relies on morphological and physiological analysis of the impaired system, which in turn allows the design of more efficient and sustainable treatment and rehabilitation strategies. In this context, it is notably crucial to compare impaired patients to the healthy population to assess the anatomical and functional modifications induced by the pathology [6]. In this direction, **information provided by medical imaging could help generating anatomical and physiological patient-specific models of the musculoskeletal system, as well as performing population-wise quantitative comparisons**. Such approaches are especially needed in the pediatric population, where musculoskeletal disorders may have a debilitating effect on a child's growth and development. Indeed, as explained in Section 1.4.2, large anatomical and physiological variations exist between the pediatric

and adult body systems. These differences can produce unique and variable responses to injuries and healing that are not seen in the mature skeletons of adults.

This chapter is organized as follows. First, Section 2.2 provides an overview of pediatric musculoskeletal pathologies, including: bone growth complications (Section 2.2.1), neuromuscular disorders (Section 2.2.2), sports-related injuries (Section 2.2.3), and the two musculoskeletal conditions present in the imaging resources employed in this thesis (Section 2.2.4). Next, Section 2.3 presents the techniques and challenges associated with pediatric musculoskeletal image acquisition (Section 2.3.1) and analysis (Section 2.3.2). Finally, Section 2.4 introduces the clinical motivations (Section 2.4.1), pediatric imaging resources (Section 2.4.2), and technical challenges (Section 2.4.3) of this thesis.

2.2 Pediatric musculoskeletal pathologies

2.2.1 Bone growth complications

The pediatric musculoskeletal system differs significantly from that of adults, and these anatomical and physiological differences result in unique disorders and injury patterns. Like all parts of the child's body, the musculoskeletal system is considered immature and still in development, so pathologies affecting any of its component tissues (i.e., bones, muscles, cartilages, ligaments, tendons) could result in growth disturbance and possible disabling complications [149]–[151]. For instance, bone growth is a complex process that involves a cartilage plate at the end of long bones (i.e., physis or growth plate) of children with an immature skeleton. The physis can be divided into three layers based on histology and function: the germinal zone, which contains stem cells of chondrocytes (i.e., cells constituting the cartilage) and an abundance of extracellular matrix; the proliferative zone, where chondrocytes are rapidly dividing, allowing for longitudinal bone growth; and the hypertrophic zone, where chondrocytes are expanding and will undergo apoptosis (i.e., programmed cell death). The bone is formed in the succeeding zone of provisional calcification, where hyaline cartilage is converted into bone during the mineralization process [150], [152].

A second ossification center localized in the epiphysis (i.e., rounded end of long bone) is surrounded by a spherical growth plate that undergoes enchondral ossification analogous to that of the primary physis. The second ossification center is initially spherical, but with growth it conforms to the contours of the epiphysis and becomes more hemispherical and

eventually borders the physis [152]. Bone growth continues until a physeal line replaces the physis in a process known as physeal closure or growth plate fusion, that typically occurs during adolescence [150]. Therefore, direct injury to the physis from a fracture or a chronic disorder can lead to premature physeal closure with bone bridge formation, one of the most common growth disturbances. Specifically, longitudinal injury across the physis may allow for the formation of transphyseal vessels, along which osteoprogenitor cells can deposit bone, thus forming a bridge across the physis that disturbs bone growth [150], [152]. Growth complications of physeal bridges depend upon the location of the bridge within a particular physis, with bridges located within the central portion of the physis leading to longitudinal growth restriction, while bridges located at the periphery of the physis can cause angular deformities [150]. In turn, lesions of epiphysis may also result in the loss of the discoidal shape of its growth plate and the curvature of the slowed growth zone with consequent angular deformation [150], [152].

Although fracture is the most common bone-related trauma, other pathologies affecting pediatric bones may include: osteochondrosis (e.g., Legg-Calvé-Perthes or Blount diseases), osteomyelitis (i.e., inflammation or swelling due to an infection), juvenile idiopathic arthritis (i.e., type of arthritis in children), or bone tumors (i.e., abnormal growth of bone tissue) [150]. **While the mechanisms of these pathologies vastly differ, each can be associated with disturbed bone growth** and result in bone bridge formation if the physis is damaged. Therefore, **imaging diagnosis of bone pathologies is essential in the juvenile population**, as these disorders can impact the rest of the musculoskeletal system, leading to limb length disparity and eventually permanent extremity dysfunction.

2.2.2 Pediatric neuromuscular disorders

While bone disorders can lead to localized growth restriction, pathologies affecting muscles and, more generally, the motor system can result in more serious disability, with locomotion and movement becoming almost impossible [151], [153]. Pediatric neuromuscular disorders encompass the spectrum of diseases affecting the peripheral nervous system, neuromuscular junction, or skeletal muscle. Neuromuscular disorders can generally lead to abnormal muscle function, muscle atrophy, muscle weakness, impaired movement, skeletal deformities or respiratory failure, and are frequently associated with severe debilitation and premature mortality. In pediatrics, the majority of neuromuscular disorders are genetic, with the most commonly encountered condition being Duchenne muscular

dystrophy, spinal muscular atrophy, and Charcot-Marie-Tooth disease [151], [153], [154].

Specifically, Duchenne muscular dystrophy is characterized by progressive weakness and breakdown of skeletal muscles over time due to fat replacement of muscle fiber, with the thighs and calves usually being first affected, resulting in difficulty walking. Eventually, the disorder progresses to all muscles, and complications typically consist of skeletal deformities and respiratory impairment. For its part, spinal muscular atrophy results in the loss of motor neurons and progressive muscle wasting, while Charcot-Marie-Tooth disease affects the peripheral nervous system and is characterized by progressive loss of muscle tissue and touch sensation. Both disorders are also associated with locomotion trouble, skeleton deformities, and respiratory failure. It should be noted that while there is no known cure for these disorders, supportive care, including physical therapy, occupational therapy, respiratory support, nutritional support, orthopedic interventions, and mobility support, may help relieving some symptoms and improve life expectancy [151], [153], [154].

Another special type of pathology affecting the pediatric musculoskeletal system is **cerebral palsy, a group of movement disorders that appear in early childhood caused by damage to the motor cortex of the developing brain during pregnancy, delivery, or shortly after birth** [155], [156]. Cerebral palsy is the most common movement disorder in the pediatric population and is characterized by abnormal motor development, posture, and balance. The most frequent movement disorders associated with cerebral palsy are spasticity (i.e., muscle tightness), dyskinesia (i.e., involuntary muscle movements), and ataxia (i.e., clumsy voluntary movements). These movement disorders and associated muscle weakness may result in secondary problems, including hip dislocation, hand dysfunction, joint contractures, articular cartilage atrophy, tendon tightness, equinus deformity, and thinner bones. In turn, these musculoskeletal injuries cause gait abnormalities such as tip-toeing gait due to tightness of the Achilles tendon and scissoring gait due to tightness of the hip adductors, which are typical of children with cerebral palsy. Similar to other neuromuscular disorders, there is no known cure for cerebral palsy, but supportive treatments, medication, and surgery may help many individuals. In clinical practice, medical imaging can help analyze the impaired musculoskeletal system, and evaluate abnormal development or damage to the motor cortex in the pediatric brain [155], [156].

2.2.3 Sports-related injuries in young athletes

A notable trend in the pediatric population is the increasing number of sport-related trauma affecting the musculoskeletal system, as sports activities have become an integral part of children's extracurricular activities [157]. These lesions encompass acute and chronic injuries such as joint dislocation, bone fracture, muscle contusion, tendon inflammation, or ligament tear [157], [158]. As previously explained, bone fracture and chronic repetitive trauma can lead to injuries to the growth cartilage and may involve the formation of a bone bridge.

For instance, a little leaguer's shoulder refers to a stress-related injury characterized by the widening and irregularity of the proximal humeral physis due to repetitive and poor throwing practice. Repetitive stress is also associated with osteochondritis, as in Osgood-Schlatter syndrome (i.e., inflamed bone or cartilage), while joint dislocation can result in recurrent and global instability. These injuries and resulting conditions are rarely followed by lasting growth impairment but are typically associated with pain and swelling, and may restrict movement. Hence, medical imaging is needed to provide robust clinical diagnosis and optimally plan physical therapy [157], [158].

2.2.4 Equinus and obstetrical brachial plexus palsy conditions

This thesis focuses on two pediatric musculoskeletal conditions: equinus and obstetrical brachial plexus palsy (OBPP), which are briefly introduced below:

- **Equinus deformity** is a clinical condition that affects the ankle joint's function by restricting its range of motion [159]–[161]. The ankle joint (or talocrural joint) refers to the articulation of the talus between the lateral and medial malleoli of the fibula and tibia (Figure 2.1) and its movements allow for dorsiflexion and plantarflexion of the foot. The subtalar joint between the calcaneus and talus also contributes significantly to foot positioning (i.e., inversion and eversion) but plays a minimal role in dorsiflexion or plantarflexion movements. Calf muscles attached to the calcaneus via the Achilles tendon are responsible for the plantarflexion movement. Additionally, the joint surfaces of all bones in the ankle are covered with articular cartilage, and each joint is bounded by strong and supporting deltoid and lateral ligaments. Ankle equinus is most notably characterized by reduced dorsiflexion, with the magnitude of diminution in movement being variable among patients

[159]. However, the etiology of this condition is poorly understood, and numerous causes have been discussed, such as muscle spasticity due to cerebral palsy, bone block in ankle joint, and tendon or calf muscle stiffness [160]. Moreover, ankle equinus is typically associated with increased forefoot loading, aggravated risk of ankle sprain, and, more importantly, possible bony deformity due to excessive and repetitive compensation for the unstable and inefficient gait (i.e., toe-walking) induced by reduced dorsiflexion [160].

- For its part, **obstetrical brachial plexus palsy** is a common birth injury associated with complex or assisted delivery during which the peripheral nervous system is disrupted [162]. The brachial plexus is formed by cervical and thoracic nerves that innervate the upper limb, and despite having the same mechanism of injury, the severity of nerve lesions can largely differ among individuals. Indeed, the various possible sequelae affecting the shoulder, elbow, or forearm depend on the localization and intensity of the nerve damage. This thesis focuses on complications involving the shoulder joint, as this nerve injury may result in shoulder muscle atrophy, impedes bone growth, and osseous deformity [163]. More precisely, OBPP is associated with delayed ossification and malformed bones, including hypo-plastic humeral head, non-spherical humeral head, hypoplastic scapula, elevated scapula, and abnormal scapula glenoid [164]. It should be noted here that the shoulder joint defines the articulation between the glenoid fossa of the scapula and the head of the humerus (Figure 2.1), which allows for free movement of the arm. Rotator cuff muscles and tendons enable different types of movements (i.e., flexion, extension, abduction, adduction, circumduction, and rotation) and, along with ligaments, maintain the glenohumeral joint stability. The articular cartilage at the interface of the scapula and humerus provides smooth joint motion with minimal friction. Modifications in muscle and bone morphology thus lead to shoulder strength imbalance and joint range of motion reduction, which consequently limit the function of the pediatric shoulder [165].

For both pathologies, medical imaging enables the acquisition of patient-specific information related to the degree of organ deformity, which is key to understanding morphological modifications for better diagnosis, treatment planning and follow-up [166].

2.3 Acquisition and analysis of pediatric musculoskeletal images

2.3.1 Background on pediatric musculoskeletal image acquisition

Imaging of the pediatric musculoskeletal system is required in a variety of clinical scenarios. Nevertheless, as introduced in Section 1.3.1, the image modality (i.e., X-ray, PET, ultrasound, CT, and MRI) and acquisition protocol may differ depending on the medical application [148]. For instance, X-ray imaging techniques are used for studying osseous structures and their disorders. However, the low contrast between soft tissues remains inadequate for the complete diagnosis and evaluation of most musculoskeletal pathologies. Consequently, plain radiographs are typically limited to the assessment of bone fractures and injuries, while CT and fluoroscopy can be used to guide and assist surgeons during therapeutic procedures [148]. Meanwhile, the evaluation of bone and soft tissue tumors and, to a lesser degree, the monitoring of bone infection in pediatric patients can be performed via PET/CT hybrid scanners [167]. However, following the ALARA [1] principle, ionizing radiation should be restricted when acquiring images of pediatric patients. CT and PET scans should therefore only be employed based on informed clinical indications [168].

For their part, ultrasound and magnetic resonance devices enable imaging of soft musculoskeletal tissues (i.e., muscles, cartilage, ligaments, tendons). Ultrasound is a particularly useful technology for studying muscular injuries due to its relative inexpensiveness and wide availability, as well as, its ability to capture real-time and functional information via the Doppler effect [169], [170], as mentioned in Section 1.2.1. Moreover, the presence of non-ossified cartilage portions in pediatric bones can improve the diagnostic capability of ultrasound through better visibility of muscle, tendon, and cartilage structures [169]. However, ultrasound examinations strongly depend on the physician and require necessary background experience and knowledge [157], [169], [170].

In clinical practice, MRI is thus preferred over ultrasound despite its high equipment cost and long scanning time [7], [171]. Indeed, as opposed to ultrasonography which is limited by the range of sound waves, **magnetic resonance scanners can efficiently capture deep and complex osseous structures** [7], [157], [168]. Moreover, **the high**

1. As Low As Reasonably Achievable, as defined in Chapter 1.

spatial resolution and superior soft-tissue contrast make MR imaging the most suitable technique to depict a complete picture of the pediatric musculoskeletal system [157], [168], [171], [172]. Most notably, **thanks to its ability to image the cartilaginous structures of developing bones, MRI has become the modality of choice for evaluating children with growth disorders and directing surgical management** [157]. The absence of radiation exposure during MRI scans makes it particularly recommended in the management of pediatric patients, especially when several scanning sessions are needed to evaluate disorders' progression or to assess complete restoration after a traumatic injury. Nevertheless, sedation may be required to perform the scan, raising practical issues and ethical considerations.

For its part, dynamic MRI devices allow to capture the moving anatomy and provide an informed diagnosis based on joint kinematics. Nevertheless, the use of real-time dynamic MRI to evaluate joints remains limited in clinical routine [173], [174]. Additionally, although both T1 and T2 MRI sequences are employed in practice, the visualization of the pre-ossification centers in T2-weighted images may present in-homogeneous signal intensity that, in turn, may induce diagnostic errors [171]. **Therefore, due to all of its imaging specificity, the analysis of pediatric musculoskeletal images requires specialized radiologists and dedicated computerized tools.**

2.3.2 Recent trends in pediatric musculoskeletal image analysis

The analysis of pediatric musculoskeletal images relies on a precise knowledge of the anatomical characteristics that differentiate it from an adult and the pathological processes that impact the development of children [169], [171], [175]. For example, radiologists evaluating pediatric bone deformity must understand bone growth mechanisms and associated disorders [148], and a comparison to the healthy population may be necessary to assess the magnitude of morphological abnormalities and resulting impairment. Additionally, the growth plate can display several confusing but normal appearances, including normal physeal undulations or focal periphyseal edema, which should not be mistaken for pathological findings such as physeal fracture, infection, or bridge [171]. As for the rest of medical imaging, **human-based analysis of pediatric musculoskeletal images involves a considerable workload, large resource consumption, and is prone to practitioner errors, the latter being especially aggravated by these developing and sometimes confusing anatomical features** [171]. All of these aspects reinforce the need for automated and reliable computerized techniques.

Even though deep learning has become the standard approach for pediatric musculoskeletal image analysis, studies remain scarce compared to the corpus on adult imaging, which may be due to the challenges specific to pediatrics. Nevertheless, some key applications of deep learning technologies (defined in Section 1.3.2) in pediatric musculoskeletal imaging include bone age prediction via regression [176], landmarks localization via detection [177], and muscle and bone shape extraction via segmentation [40], [178], [179]. Bone age prediction is a common technique based on the determination of bone development characteristics to obtain a numerical assessment of human development [176]. This information is typically extracted by medical experts from X-ray images (e.g., hand bones X-ray) and is in turn used to assess the physical development of adolescents and to discover or prevent growth disorders [176].

For its part, localization aims at positioning anatomical landmarks or biologically meaningful loci that must be unambiguously defined and repeatedly located with a high degree of accuracy and precision [177]. Such analysis can be performed on 2D (e.g., plan radiographs) or 3D (e.g., CT or MR scans) images, and the obtained landmarks are then typically used to compute lengths and angles characterizing the degree of impairment and deformity of a given patient. For instance, one can extract femoral and tibial landmarks from full-length anteroposterior X-ray radiographs to measure the respective bone lengths. Based on this anatomical information, clinicians can monitor growth and alignment of the lower extremities, as well as assess limb length discrepancy disorders that may be congenital or acquired [177].

Finally, segmentation of musculoskeletal tissues, particularly from 3D CT or MRI scans, can provide comprehensive surface and volumetric information through the generation of polygon meshes. In turn, these can be used to create patient-specific models for further kinematic, mechanical and morphological analysis. For example, muscle segmentation can help assessing both fat replacement and the degree of muscle weakness caused by neuromuscular disorders [40], [179]. In contrast, bone segmentation may provide an accurate understanding of the pathomechanics of joints [173], [174]. As previously mentioned, all of these tasks are typically performed manually by radiologists, but neural networks have recently been shown to automatically achieve these types of outcomes [40]. Nevertheless, it should be emphasized that the literature on cartilaginous, ligamentous, and tendinous pediatric tissue segmentation remains limited.

2.4 Clinical motivations, pediatric imaging resources, and technical challenges

2.4.1 Clinical motivations

This thesis aims at automatically extracting and segmenting bone shapes from pediatric MR images. For the analysis of pediatric musculoskeletal disorders, segmentation helps clinicians in effectively quantifying morphology and possible deformity by providing 3D solid or surface models of bones [3]–[5]. As with all medical image segmentation tasks, delineating pediatric bone is difficult, tedious, and time-consuming. However, it also presents additional challenges, such as the ongoing bone ossification process, the large anatomical variability between age groups, and the thinness of osseous structures. Hence, segmentation automation could improve the reliability and robustness of the generated delineations while reducing the need for human intervention in image processing tasks. More precisely, **this thesis aims at developing fully-automatic bone segmentation methods from pediatric MR image datasets of three musculoskeletal joints: ankle, knee, and shoulder** (Figure 2.2). It should be emphasized that MR scanners provide complete and highly resolved 3D imaging of each anatomical joint, with the growth plate visible inside bones, especially in T1-weighted images.

In any case, fully-automated and reliable bone segmentation of pediatric examinations could provide a rapid evaluation of the patient's level of impairment, guide surgery, and help optimize rehabilitation programs. Furthermore, patient-specific 3D bone models could also assist clinicians in analyzing strength imbalance and the kinematics and dynamics of pathological and healthy joints [3]–[5]. For instance, it has been reported a clear relationship between muscle atrophy and strength loss in the context of OBPP. Hence, accurately quantifying muscle morphology can directly translate to a better understanding of shoulder muscles strength imbalance and other biomechanical properties [40], [165].

2.4.2 Pediatric imaging resources

Ankle and shoulder imaging datasets were acquired at Centre Hospitalier Régional Universitaire (CHRU) La Cavale Blanche, Brest, France, using a 3T Achieva scanner (Philips Healthcare, Best, Netherlands) [2] while knee imaging datasets were obtained ret-

2. Data were acquired with the support of Fondation motrice (2015/7), Fondation de l'Avenir (AP-RM-16-041), PHRC 2015 (POPB 282), and Innoveo (CHRU Brest).

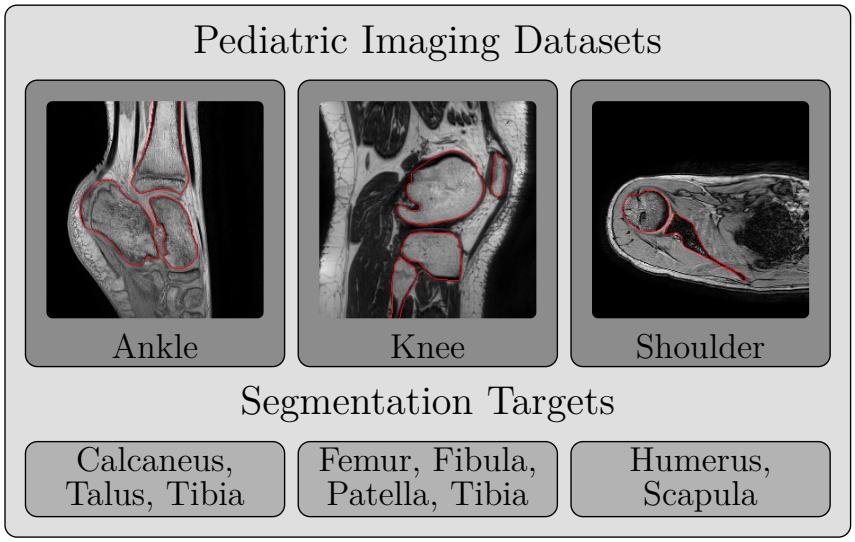


Figure 2.2 – Samples from the pediatric ankle, knee, and shoulder joint imaging datasets and their respective segmentation masks consisting of the following bones: `[calcaneus, talus, tibia (distal)]`, `[femur (distal), fibula (proximal), patella, tibia (proximal)]`, and `[humerus, scapula]`. Ground truth delineations are in red (–).

rospectively from the Children's Mercy Hospital, Kansas City, United States [3]. The knee data was acquired using a 3T MRI scanner (MAGNETOM Skyra, Siemens Healthineers, Siemens AG). MRI data acquisition was performed in line with the principles of the Declaration of Helsinki. Ethical approvals were respectively granted by the Ethics Committee (Comité Protection de Personnes Ouest VI) of CHRU Brest (2015-A01409-40) and by the research ethics committee of the Children's Mercy Hospital, Kansas City, United States. Additional information on the imaging acquisition protocols and patient cohorts is provided for each dataset, as follows:

- **Ankle joint dataset.** The ankle joint dataset contained 20 MR examinations acquired on pediatric individuals aged from 7 to 13 years (average age: 10.1 ± 2.1 years). A T1-weighted gradient echo sequence was employed during image acquisition (TR: 7.9 ms, TE: 2.8 ms, FOV: 140×161 mm^2), with resolutions varying from $0.25 \times 0.25 \times 0.50$ mm^3 to $0.28 \times 0.28 \times 0.80$ mm^3.
- **Knee joint dataset.** The knee imaging dataset consisted of 17 MR examinations extracted from a pediatric cohort composed of patients aged from 13 to 18 years old (average age: 15.4 ± 1.6 years). Images were acquired using a 3D Gradient

3. We would like to acknowledge Dr. Antonis Stylianou from the University of Missouri-Kansas City, Kansas City, United States and Dr. Donna Pacicca from Children's Mercy Hospital, Kansas City, United States for sharing the anonymized knee joint image dataset.

Recall Echo (GRE) sequence (TR: 13.0 ms, TE: 4.4 ms, FOV: 320×320 mm^2), with resolutions ranging from $0.47 \times 0.47 \times 0.5$ mm^3 to $0.625 \times 0.625 \times 0.63$ mm^3.

- **Shoulder joint dataset.** MR images of 15 shoulder joints were obtained from pediatric individuals aged from 5 to 17 years old (average age: 11.6 ± 4.4 years). Images were acquired using an eTHRIVE (enhanced T1-weighted High-Resolution Isotropic Volume Examination) sequence (TR: 8.4 ms, TE: 4.2 ms, FOV: 260×210 mm^2). Image resolution varied across subjects from $0.24 \times 0.24 \times 0.60$ mm^3 to $0.37 \times 0.37 \times 1.00$ mm^3.

As previously indicated, ankle and shoulder images were respectively acquired to study the equinus condition and OBPP disorder. More precisely, both ankle and shoulder imaging datasets presented a mixture of healthy and pathological examinations, including 9 pathological and 11 healthy ankle joints, and respectively 7 pathological and 8 healthy shoulder joints. For its part, the pediatric knee dataset only comprised healthy cases, which are essential in understanding the development of the normal and unimpaired morphology. Furthermore, the imaging datasets were acquired from three unpaired and distinct pediatric cohorts, which included both male and female patients. The children ranged in age from 5 to 18 years old, with on average, older patients in the knee cohort (mean age 15.4 years) and younger patients in the ankle cohort (mean age 10.1 years), while the shoulder cohort presented the largest age variability (age ranging from 5 to 17 years). Imaging data thus included a large anatomical variability due to the presence of examinations from early childhood to late adolescence.

To provide a quantitative assessment of each joint, we targeted the segmentation of three ankle bones (calcaneus, talus, and tibia), four knee bones (femur, fibula, patella, and tibia) and two shoulder bones (humerus and scapula), as seen in Figure 2.2. Indeed, this allows us to study the ankle talocrural and subtalar joints, both tibiofemoral and the patellofemoral knee joints, as well as the glenohumeral shoulder joint. It should be noted that all images were annotated by a medically trained expert (15 years of experience) using the ITK-SNAP software [4] to get ground truth labels of each bone. The growth cartilage was simultaneously included with completely ossified areas in the ground truth labels of each bone. These ground truth segmentation masks are needed for algorithm optimization and performance evaluation. Finally, although the T1 MRI sequence allows for soft tissue (i.e., muscle, cartilage, tendons) visualization, the ground truth labels of these structures were not manually produced due to resource and time constraints.

4. `http://www.itksnap.org/`

Ultimately, following the definitions given in Chapter 1, each pediatric MR dataset defines a distinct image domain and segmentation task. However, it should be noted that these datasets present similar characteristics: pediatric population, targeted bone structures, and MR modality with T1 sequence.

2.4.3 Technical challenges

Following their widespread success in medical imaging [1], [2], we propose to employ deep learning algorithms for the segmentation of pediatric bone in MR images. However, we face numerous technical challenges that can be divided into three categories.

- First, generic challenges present in all medical image applications including the scarcity of imaging data, the paucity of labels, and the generalization gap, which are reinforced due to the pediatric nature of our imaging resources (see Section 1.3.3).
- Second, challenges specific to pediatric musculoskeletal analysis such as the ongoing bone ossification process, the large anatomical variability between age groups, and the thinness and smallness of osseous structures (see Section 2.3.2).
- Third, challenges specific to our datasets: high level of shape heterogeneity due to a mixture of healthy and pathological examinations, varying ages from childhood to adolescence (see Section 2.4.2), as well as motion noise present in some sample images due to patient movement during acquisition.

This thesis, therefore, aims to address these challenges by developing novel deep learning techniques following recent paradigms, including multi-anatomy learning and regularization integration as explained in Chapter 1. In particular, multi-anatomy learning and shape priors regularization appear natural in the context of the pediatric musculoskeletal image segmentation. Indeed, one can assume that multi-anatomy learning could be beneficial to leverage shared features (e.g., shape, pose, intensity) present in MR imaging datasets arising from distinct anatomical joints (i.e., ankle, knee, shoulder). Moreover, even considering healthy and pathological variations, anatomical structures, especially bones, are globally constrained in terms of shape and position. It is thus relevant to learn shape priors in a data-driven manner, as we will see in Chapter 4.

While the methodologies developed in this thesis share certain global limitations which will be discussed in Parts II and III (e.g., amount of imaging data, number and types of annotated structures, evaluated joints and pathologies), we believe that our approaches are generic enough to have a great impact for the analysis of other pediatric musculoskeletal

pathologies (e.g., impeded bone growth, neuromuscular disorders) and structures (e.g., joint, tissues, age-group). This thesis also provides some insights on possible directions for better management of pediatric imaging resources and associated ground truth labels.

2.5 Conclusion

This second chapter introduced the clinical context of this thesis which aims at providing segmentation tools for the analysis of the pediatric musculoskeletal system. The pediatric musculoskeletal is a complex system that largely differs from that of adults and which can be affected by numerous disorders and conditions. Therefore, the unique anatomical features and pathologies patterns present in pediatrics can hinder the analysis of medical images by radiologists and reinforce the need for automatic tools to assist clinical diagnosis. While this thesis targets the segmentation of ankle, knee, and shoulder bones, we believe that our methodologies could have a greater impact on the analysis of other pediatric musculoskeletal structures.

Following its widespread success in medical image analysis, we propose to develop deep learning-based methods for pediatric bone segmentation. Chapter 3 introduces the basic concepts of medical image segmentation using deep learning that will be employed throughout this thesis manuscript.

Chapter 3

Deep learning for medical image segmentation

3.1 Introduction

Although the success of deep learning methodologies over traditional machine learning dates back nearly a decade, the inception and implementation of the first artificial neural networks emerged in the 1960s. Indeed, the perceptron model, which Rosenblatt implemented in 1962, is generally considered as the starting point of the deep learning field [14], [180]. At its core, the perceptron was an algorithm for learning a linear binary classifier in a supervised manner. The perceptron was the simplest neural network, composed of a single "neuron" predicting a binary value. This model was indeed limited and was notably unable to represent the XOR logical function that is not linearly separable. To address this issue, researchers developed a two-layer perceptron where the first layer learned a hidden representation of the data. This "hidden neuron" encompassed an affine transformation controlled by learnable parameters, followed by a fixed non-linear function called an activation function. Specifically, this layer learned to map the input into a new space where the data was linearly separable and from which a second neuron could now solve the problem [14]. Following this idea, multi-layer perceptron (MLP), defined by a cascade of hidden layers, were implemented to solve more complex problems. Currently, the architecture of an MLP is determined by its depth and width, which respectively correspond to the number of hidden layers and the dimension (or number of features) of each neuron. These values are set empirically by practitioners and generally depend on the nature of the problem to be solved. It should be noted that the design of the perceptron was initially motivated by the study of biological neurons, hence the term "neural" network, but the comparison with the human neuronal system does not go any further [14], [180].

In computer vision, the Neocognitron developed by Fukushima in 1980 is typically

considered as the first successful implementation of a neural network for an image processing task [181]. In turn, this model, which targeted handwritten character recognition, later inspired LeCun to design his LeNet in 1989 [182]. **In retrospect, the LeNet integrates and combines all the essential components of modern convolutional neural networks (CNNs) [13], [14]. In particular, CNNs are special neural networks that use convolution operators as linear transformations.** This enforces a desirable characteristic in neural networks for pattern recognition: translation-equivariant representations. Furthermore, while each convolutional layer only learns local interactions, the cascade of convolutions allows these simple blocks to build complex hierarchical representations. The network also incorporates downsampling layers to obtain representations with separated spatial scales incorporating fine-grained to global features [13], [14]. Last but not least, LeCun employed the back-propagation method, which was freshly introduced by Rumelhart in 1986 [183], to learn the weights of LeNet using the gradient descent algorithm [184]. While this approach was successful in solving handwritten digit recognition tasks, other machine learning algorithms such as support vector machines (SVMs) were typically preferred due to their lower computational requirement and better interpretability [1], [2], [14]. Indeed, the LeNet training required multiple days of computation, and it remained difficult to interpret the weights learned during the optimization steps. This lack of interpretability is commonly known as the "black-box" nature of neural networks [13], [14], as discussed in Section 1.3.4.

In parallel, more theoretical works led from 1989 to 1999 [185]–[188], demonstrated that neural networks with one hidden layer could approximate any continuous functions on compact subsets of $\mathbb{R}^d$. However, this universal approximation theorem states that the cost of such approximation is an arbitrarily large network's width. It should be noted that this theorem requires only that the activation function of the neural network be point-wise and non-polynomial, which is a very mild condition [187]. Nevertheless, to approximate a function with a given precision, the width of the network is exponential with respect to the dimension d of the input data. For image analysis, the dimensionality of the data, which corresponds to the number of pixels, is typically large. The required number of neurons, therefore, makes any practical implementation infeasible. Moreover, while this theorem suggests that neural networks can represent a wide variety of functions, it fails to provide a way to obtain the network's weights. Ultimately, this theorem has very limited practical insights for the design of CNNs and their optimization.

In the 2000s, machine learning algorithms (e.g., SVMs, random forests, active contour

models) were still favored over neural networks for computer vision tasks and medical image analysis. At the time, deep learning researchers aimed at improving the performance of CNNs through better hardware implementation or more carefully designed activation functions. In particular, one can mention the introduction of the rectified linear unit (`ReLU`), which lead to more stable optimization by preventing vanishing gradient problems encountered with previous non-linearity [189]. The deep learning field experienced a resurgence in 2012 when AlexNet [190] surpassed traditional machine learning approaches on the ImageNet natural image dataset [45]. The ImageNet challenge evaluates the image classification performance over 1000 classes, and in the past decade, deep learning approaches improved the top-1 accuracy from 63.3% for AlexNet to 84.3% for EfficientNet [191]. This resurgence is multi-factorial and can be partly explained by the experimentation made during the 2000s (e.g., better hardware implementation and integration of `ReLU` non-linearity). One can also mention global technological advancements, such as the wider availability and better performance of graphical processing units (GPUs) or the collection of large-scale open-access image databases (i.e., big data). Both were required to accelerate deep neural networks optimization and enhance image processing tasks' performance [13], [14]. Following this success, deep learning methodologies were largely applied to medical image analysis tasks [1], [2]. As a first step, CNNs were developed and optimized for medical image classification (e.g., detecting tumors or metastasis) [192]. For its part, image segmentation represents a more challenging task. Indeed, instead of returning a single class prediction, segmentation aims at providing pixel-wise class prediction. Hence, novel convolutional encoder-decoder architectures have been designed, the best-known example being the UNet model [29]. **The UNet is one of the most successful deep learning models and has been declined in many variants. Thanks to its efficiency, it has been employed in various clinical applications, including the segmentation of brain tumors, abdominal organs, or lungs [1], [2], [91], [117], [193].**

In the context of pediatric musculoskeletal imaging, the morphological information obtained through segmentation can help clinicians assess disorder progression and design novel treatment strategies (see Chapter 2). Hence, **this chapter presents a general mathematical framework for deep learning-based image segmentation.** As mentioned in Chapter 1, the novel architectures and training schemes proposed in this thesis are based on UNet and aim to mitigate the generalization gap and data scarcity issue present in pediatric medical imaging. We also introduce the baseline architecture used in the rest of this thesis, as well as implementation details for the experiments performed

in Parts II and III. Most importantly, we aim to provide background on deep learning to build more advanced architectures and training schemes in the following chapters.

The remainder of this chapter is organized as follows. First, Section 3.2 provides the mathematical framework for deep segmentation, which is formulated as a function approximation problem (Section 3.2.2) and thus necessitates optimization tools to be solved (Section 3.2.3). Then, we present the convolutional encoder-decoder architecture for image segmentation in Section 3.3. We briefly recall the role of the convolution operator (Section 3.3.1), non-linear activation, pooling layer (Section 3.3.2), segmentation decoder (Section 3.3.3), and skip connections (3.3.4) of the UNet model. Next, Section 3.4 introduces standard modifications of the UNet model, such as batch normalization (Section 3.4.1) and spatial attention gates (Section 3.4.3). Finally, Section 3.5 provides technical details on deep learning model implementation (Section 3.5.1) and performance assessment (Section 3.5.2).

3.2 Mathematical framework for deep segmentation

3.2.1 Image domain and label space

Let $x = \{x_u \in \mathbb{R}, u \in \Omega\}$ be a grayscale image embedded in the image grid $\Omega \subset \mathbb{Z} \times \mathbb{Z}$ with x_u the intensity value of the pixel at position u in the grid. The corresponding image class labels $y = \{y_{c,u} \in \{0,1\}, c \in \{0,...,C\}, u \in \Omega\}$ (or pixel-wise annotation maps) represent the $C+1$ different anatomical objects of interest (plus background) present within the image. Although we consider a 2D setting (i.e., image grid $\Omega \subset \mathbb{Z} \times \mathbb{Z}$), the following can easily be extended to volumetric images or higher dimensional objects (e.g., volumetric dynamic images). We make the traditional assumption that the $C+1$ classes are mutually exclusive, as each class c represents a distinct anatomical structure (or background). Specifically, for each pixel position u in the image, only one class is present and assigned the value 1, while the remaining classes are set to 0 (i.e., absent). More formally, the segmentation label y must respect the following condition of mutual class exclusion: $\forall u \in \Omega : \sum_c y_{c,u} = 1$.

In the rest of this thesis, we will refrain from employing such a pixel-level description of the image x and label y. Nevertheless, it is important to emphasize that an image is characterized by a highly organized data structure arising from the image grid Ω, which naturally induces a notion of distance (and therefore locality) between pixels [14], [28].

Moreover, an image's content is highly organized because an image typically comprises distinct objects and structures defined by specific characteristics. These objects are also interrelated or correlated to each other and can present shared features between them. This is even more relevant in the context of medical imaging, where the human body is highly organized with global and local systems and structures (see the the musculoskeletal system definition in Chapter 2). These properties specific to images (i.e., highly organized data structure and content) have motivated the design of initial CNNs for image processing, as we will explore in Section 3.3.

Furthermore, **one can define the image x and label annotations y as respectively belonging to the intensity space $\mathcal{I}$ and the label space $\mathcal{C}$**. As introduced in Chapter 1, the image intensity domain and segmentation label task are defined by multiple characteristics, including the imaging modality, the setting of the acquisition device, the anatomical region of interest, and the targeted anatomical structures. Hence, **in this thesis, the intensity domain $\mathcal{I}$ typically corresponds to the space spanned by pediatric musculoskeletal MR images of a specific anatomical joint, while the label space $\mathcal{C}$ is the space spanned by the segmentation maps of the associated pediatric bones**, as defined in Chapter 2.

3.2.2 Segmentation as a function approximation problem

In deep learning, segmentation is formulated as a function approximation problem in which the goal is to learn a mapping S^* between the image intensity domain $\mathcal{I}$ and the segmentation label space $\mathcal{C}$. The function S^* is approximated by a neural network S parameterized by the weights Θ, and which generates a segmentation prediction $\hat{y}$ given an intensity image x, as follows:

$$\begin{aligned} S &: \mathcal{I} \longrightarrow \mathcal{C} \\ x &\mapsto \hat{y} = S(x;\Theta) \end{aligned} \tag{3.1}$$

The network S can be typically decomposed in a succession of simple and generic functions (i.e., layers), which can be organized into a directed acyclic graph whose length defines the depth of the model. In particular, the intermediate hidden (or latent) layers extract more and more abstract information and characteristics (i.e., features) from the input image, while the final output layer generates the prediction segmentation. For image processing, we will see in Section 3.3.1 that neural networks are

usually built upon convolutional layers.

These layers are parameterized by the weights Θ which must be learned during an optimization step thanks to a dataset of n training samples $\mathcal{D} = \{x_i, y_i\}_{1 \leq i \leq n}$. From this perspective, deep learning is considered as a data-driven technology, as stated earlier in Chapter 1. The training set is composed of n couples of images x_i and corresponding ground truth segmentation maps $y_i \approx S(x_i)$, which provide noisy approximate of the function S^* evaluated at different data points. **The training procedure encourages the network to produce prediction $\hat{y}_i$ as close as possible to the ground truth y_i for each image x_i, and adjust (or tune) the weights Θ to produce the desired output and implement the best approximation of the desired segmentation mapping S^*.** This setting is known as a supervised learning framework. As the behavior of the latent layers is not directly specified during optimization, the learned weights are typically not interpretable, which makes deep learning models and their training procedure, difficult to analysis as previously discussed in Chapter 1.

In practice, the training procedure of deep learning models leverages a loss function $\mathcal{L}$ to learn the parameters Θ (or neurons) of S and approximate S^*. In the context of supervised learning, the loss or cost function typically measures the error between the model output predictions and the ground truth annotations. It should be noted that more advanced learning schemes can include additional regularization terms, as we will explore in Parts II and III. Regularization is a key concept in deep learning, which encompasses all the techniques aimed at reducing over-fitting and improving generalization [14], [43]. One such technique relies on the addition of a penalty term to the loss function which will constrain the optimization procedure to enforce suitable characteristics in the neural network. The traditional training procedure thus aims at finding the parameters Θ^* that minimize the loss function $\mathcal{L}$, as follows [1]:

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}(\Theta) \tag{3.2}$$

If we assume that $\mathcal{L}$ is differentiable with respect to Θ, the gradient descent algorithm is a standard optimization tool for finding a local minimum of such a function [184]. By applying the chain rule of calculus, this condition signifies that each layer of the neural network and the loss must be differentiable with respect to their inputs. If one follows standard deep learning practices, this condition is always respected, and gradient descent

1. The expression of the regularized optimization problem is given by $\Theta^* = \arg\min_{\Theta} \mathcal{L}(\Theta) + \mathcal{R}(\Theta)$, where $\mathcal{R}$ is a regularization penalty.

is thus the standard approach to optimize neural networks. In particular, computing the gradient of the loss allows us to "tune" the weights to decrease the loss. Specifically, we iteratively update the weights in the direction of steepest descent, which is the opposite direction of the gradient, to find the local minimum, as follows:

$$\Theta \leftarrow \Theta - \alpha \nabla_{\Theta} \mathcal{L}(\Theta) \tag{3.3}$$

The learning rate α hyper-parameter controls the step size at each iteration while converging towards the minimum of the loss function. This hyper-parameter is set by a deep learning practitioner and typically represents a trade-off between convergence speed and unstable optimization.

Computing the gradient of the loss with respect to the weights $\nabla_{\Theta} \mathcal{L}(\Theta)$ can be challenging due to the number of parameters and usually requires the use of the back-propagation algorithm [183]. This popular algorithm leverages the chain rule to evaluate the gradient of one layer at a time and iterates backward to avoid redundant calculations of intermediate terms in the chain rule. **In essence, back-propagation refers only to the method for computing the gradient, while the gradient descent algorithm is used to perform learning based on this gradient.** These algorithms are already implemented in most deep learning software libraries and are easily operational, as we will see in Section 3.5.1.

3.2.3 Challenges of optimization in high-dimension

Although the combination of back-propagation and gradient descent is a successful approach to train deep learning models in many applications [13], **optimizing neural networks remains a very challenging task due to the high-dimensional and non-convex nature of their loss function**. In the following, we will present some limitations of neural network optimization, without being exhaustive:

- First, computing the gradient over the whole training set proves intractable in most scenarios. However, the loss function usually decomposes as a sum over the training examples. Hence one can approximate the gradient by computing an expected value of the cost function estimated using only a subset of samples from the dataset. This approach is referred to as *mini-batch* stochastic gradient descent or simply stochastic gradient descent [13], [14], [184]. Most modern deep learning models are trained using this scheme, and the size of the mini-batch is typically a compromise

between memory limitations and performance. Interestingly, some deep learning guidelines suggest defining the batch size, the number of convolutional filters, and the image resolution as a power of two to more efficiently fit the mini-batch data in the memory during optimization. It is worth mentioning that larger batches provide a more accurate estimate of the gradient, while smaller batches may provide a regularization effect, perhaps due to the noise they add to the learning process, but at the cost of unstable learning due to the high variance of the gradient estimate.

- Second, in practice, each image in the batch usually undergoes random geometric transformations, including translation, rotation, scaling, shear, or flipping, in addition to random intensity modifications such as normalization, blurring, or contrast adjustment. This technique, known as data augmentation, helps reduce over-fitting during optimization by increasing the amount of data with slightly modified copies of already existing images [13], [14]. It has been noted that data augmentation, such as random translations by a few pixels in each direction, can often greatly improve generalization, even if the model has already been designed to be partially translation invariant by using the convolution and pooling techniques (see Section 3.3).
- Third, for high-dimensional non-convex functions such as neural networks, it is possible to have many local minima as well as saddle points, plateaus, and other flat regions. The learning process may get "stuck" in those regions as the gradient can become very small (i.e., vanishing gradient). On the contrary, neural networks may also present steep regions (i.e., cliffs), resulting in exploding gradient that can "move" the parameters extremely far from the optimal solution. In both cases, the learning process may converge to a poor estimate of Θ^* or even diverge. To address these issues, several gradient descent strategies have been devised, most notably, the momentum which accumulates an average of past gradients to continue moving in the same direction, and adaptive learning rates that apply separate learning rates α for each parameter to adapt these hyper-parameters throughout the course of learning. It has been illustrated that both strategies help preventing high variance (i.e., noise) in the stochastic gradient, "moving" through flat regions, and avoiding areas with high curvature. Thus, many gradient descent algorithm have been designed based on variants and combinations of these techniques, including Adadelta [194], AdaGrad [195], Adam [196], and RMSprop, which are widely used to optimize modern deep learning networks [13], [14].

- Fourth, because the gradient descent is an iterative algorithm, it is needed to initialize the parameters Θ to start the optimization. However, most deep learning networks are strongly affected by the choice of initialization. Indeed, the initial point can determine whether the algorithm converges at all, how quickly learning converges and whether it converges to a point with high or low cost. Hence, several initialization strategies have been devised to achieve some properties when the network is initialized. For instance, one can mention the Xavier initialization [197], which respectively imposes the same activation variance and the same gradient variance for all layers. Specifically, the general goal of Xavier initialization is to prevent the gradients of the network from vanishing or exploding. However, assessing whether these properties are preserved after learning began, remains difficult. Furthermore, a difficulty is that some initial points may be beneficial from the optimization point of view but detrimental from the generalization viewpoint. The understanding of the relationship, between initial point selection and generalization, remains very limited, offering little practical guidance on how to initialize deep learning models [14].

Understanding and designing optimization strategies for deep learning models remains a challenge, but these are active fields of research. While not the focus of this thesis, it remains essential to recognize the limitation of modern gradient descent algorithms that are mostly heuristic, especially with respect to initialization strategies. Most importantly, it should be emphasized that there exist a lack of standardized approaches to optimize neural networks and ensure the best generalization performance when learning has ended. In parallel, there are also no clear guidelines on how to determine the most adapted gradient descent algorithm given a network architecture and training dataset. The selection of optimal hyper-parameters, typically learning rate and batch size, remains entirely empirical. We will see in Section 3.3 that, in the same way, there exist general motivations behind the design of CNNs for image segmentation, but the guidelines used to build modern architectures remain empirical.

3.2.4 The principle of maximum likelihood and cross-entropy loss function

The last ingredient needed to optimize our segmentation network lies in the definition of a suitable loss function. In the context of image segmentation, our parametric model S

defines a probability distribution $p_{\text{model}}(y|x;\Theta)$ over the segmentation masks. Specifically, since the segmentation task corresponds to a pixel-wise classification task, our model outputs a probability value between 0 and 1 for each pixel and each class. If we suppose that the training dataset $\mathcal{D}$ consists of samples drawn from a true distribution $p_{\text{data}}(y|x)$, our goal is thus to model the true distribution using our parametric model. Assuming that samples are independent and identically distributed, the maximum likelihood estimator allows to obtain the best Θ that yields the most similar distribution $p_{\text{model}}(y|x;\Theta)$ to $p_{\text{data}}(y|x)$. In particular, for independent and identically distributed random variables, the likelihood function can be directly decomposed as a product of univariate density functions [14].

Hence, by using the principle of maximum likelihood, the cost function is the negative log-likelihood, equivalently described as the cross-entropy between the training data and the model distribution. **Minimizing the cross-entropy loss thus leads to the maximum likelihood estimator of the parameter Θ^* that yields the best model according to the training examples.** The cross-entropy loss $\mathcal{L}_{\text{CE}}$ is defined as below [2]:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} := \frac{1}{n}\sum_{i=1}^{n} -y_i \log(\hat{y}_i) \tag{3.4}$$

We adopt the notation $\mathcal{L} = \mathcal{L}(\Theta)$, as the parameter Θ appears implicitly in $\hat{y}_i = S(x_i, \Theta)$. The cross-entropy loss ranges from $-\infty$ to 0, with a perfect model characterized by a zero cross-entropy loss. Due its logarithmic nature, for $y_i = 1$, the loss will slowly decreases as $\hat{y}_i$ approaches 1, however, the loss will rapidly increases as $\hat{y}_i$ decreases (i.e., $\mathcal{L}_{\text{CE}}$ penalizes confident and wrong predictions).

It should be noted that the cross-entropy loss can be used with any deep learning architecture and for any segmentation task. This loss is typically implemented in all deep learning software packages (see Section 3.5.1). Hence, it has been widely used in numerous medical image segmentation applications [1], [2]. Although the architecture of the employed segmentation networks can differ, the core of their designs follow common and shared principles which will present in the next section.

2. The full expression of the loss is given by $\mathcal{L}_{\text{CE}} = \frac{1}{n(C+1)|\Omega|}\sum_{i=1}^{n}\sum_{c=0}^{C}\sum_{u\in\Omega} -y_{i,c,u}\log(\hat{y}_{i,c,u})$, corresponding to an average over classes and pixels, with $|\Omega|$ as the cardinality of the image grid.

3.3 Convolutional encoder-decoder for image segmentation

3.3.1 The convolution operator

As previously mentioned, most deep segmentation models belong to the family of CNNs, which are based on convolutional filters particularly suited to process data with a grid-like Euclidean topology, such as images embedded in a pixel grid ($\Omega \subset \mathbb{Z} \times \mathbb{Z}$). Each convolution operator is parameterized by a kernel which extracts a specific pattern and transforms the input data into a feature map localizing the pattern in the image. The learnable parameters Θ thus include the kernel weights of each convolutional operator at each layer of the network. Each convolutional layer typically consists of multiple convolutional operators applied in parallel, and the number of filters defines the channel size (i.e., number of features) of the succeeding feature map (Figure 3.1). For its part, the size of each supporting kernel is typically small (e.g., 3×3, 5×5, or 7×7 pixels), and each convolution operation thus involves very few pixels (i.e., sparse interaction). This property is contrary to traditional neural networks, which employ dense matrix multiplication where each output depends on every input. In CNNs, the first layer aims at detecting small meaningful features (e.g., edges), while convolutions in deeper layers may indirectly interact with a larger portion of the image through a larger receptive field. This allows the network to efficiently learn complex interactions between many structures in the image and construct a hierarchical representation from simple building blocks, each describing only sparse interactions. As one can expect a greater number of complex and global features than simple and local ones, practitioners typically increase the number of channels in deeper convolutional layers (Figure 3.1) [13], [14], [28].

Furthermore, each convolution operation is applied with the same kernel weights at every pixel position of the input image (i.e., weight sharing). The pattern to be detected is thus translated across the whole image. This property is again unlike traditional neural networks based on matrix multiplication, where each element of the weights matrix is used only once per input. It is important to emphasize that weight sharing causes the network representation to be translation-equivariant, a characteristic highly desired in image analysis as feature maps should shift with their input. In particular, the model does not need to learn separate detectors for the same object occurring at different positions in an image. Nevertheless, the convolution is not equivariant to other transformations such

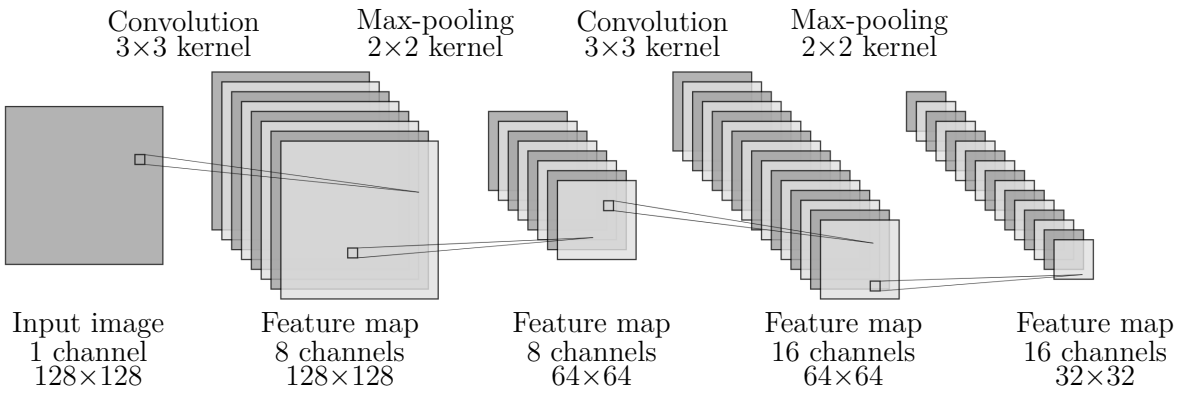


Figure 3.1 – Architecture of a CNN with convolutional and max-pooling layers. The input image is a grayscale 2D image. Each feature map is defined by its number of channels and spatial resolution. The convolution and max-pooling transformations are applied in cascade and each operates on a small 2×2 or 3×3 kernel. Schematic generated using the open-access illustration tool: `https://alexlenail.me/NN-SVG/LeNet.html`.

as rotation or scaling. Finally, **the combination of sparse interaction and parameter sharing greatly reduces the number of parameters to learn by acting as a strong prior on the network. Specifically, this prior states that the function that S should approximate, contains only local interactions and is equivariant to translation.** These properties can be seen as regularization constraints imposed on the network architecture, and they partly explain the success of CNN for image processing tasks. However, as stated in Chapter 1, the literature is currently lacking a better mathematical understanding of these models. For instance, finding the optimal number of filters per layer (i.e. network width) and, the optimal number of layers (i.e. network depth) for a given imaging dataset and task, remains primarily based on an empirical search and heuristic approach [13], [14], [28].

3.3.2 Non-linearity and pooling layer

To recapitulate, CNNs are built on a cascade of convolutional filters producing more and more abstract representations with larger and larger receptive fields. Nevertheless, convolution remains a linear transformation, and cascading convolutions will only be able to approximate linear functions. It thus becomes essential to employ non-linearity between each convolution layer in order to approximate highly non-linear image processing tasks such as segmentation. Hence, convolutional networks integrate point-wise non-linear

activation functions ρ (e.g., `ReLU`, `SiLU`, `Sigmoid`) after each filter to act as features detector. This combination of convolution operator and non-linearity transformation can be expressed as below:

$$u_{i,l+1} = \rho(\Theta_l * u_{i,l} + b_l) \tag{3.5}$$

where $u_{i,l+1}$ is the output activations map generated by the l^{th} block with the i^{th} image of the dataset $\mathcal{D}$ as input, ρ is a non-linearity, Θ_l is the convolution filter of the l^{th} layer, and b_l is the associated bias. As a convention, the input image corresponds to the input of the first layer $u_{i,0} = x_i$.

During optimization, the network learns all the weights $\{\Theta_l, b_l\}_l$ present at each layer. As the non-linearity is usually centered around zero, the role of the bias is only to shift the activation function. Activation functions were originally inspired by the biological process of brain neurons and were an abstract representation of the cell action potential acting as a binary switch. Nowadays, activation functions depart from this concept, and research is now focused on seeking non-linearity that helps stabilize optimization (e.g., `ReLU` and `LeakyReLU` reduce vanishing gradient problems compared to `Sigmoid` [189]) or improve performance (e.g., `SiLU` for deeper models [198]). Once more, such guidelines are based on empirical findings and, in the literature, few studies investigate the role of point-wise non-linearity in deep network.

Furthermore, while convolutions allow obtaining translation-equivariant representations, one may also desire to build hidden representations invariant to small translations, particularly in deeper layers that extract more abstract and global features. Pooling layers (e.g., max-pooling, average-pooling) yield such property by computing summary statistics of nearby pixels (e.g., 2×2 kernel) resulting in latent representations with smaller spatial resolution (Figure 3.1). **Pooling also induces scale separation between representations, with coarser resolutions in deeper layers extracting image-wise features.** Moreover, from a practical perspective, pooling allows for reduced memory consumption for storing and optimizing the network weights, especially considering the standard recommendation to increase the number of filters in deeper layers. Practitioners also employ strided-convolution to down-sample the feature map. Nevertheless, the translation invariance property no longer holds in such a case. Once again, the exact role and benefits of pooling and down-sampling layers are not yet clearly understood, and it remains difficult to assess which strategy one should employ in a given setting [13], [14], [28].

3.3.3 Segmentation decoder

In the end, the cascade of convolution, non-linearity, and occasional pooling or down-sampling layers allows for extracting more and more abstract features from the image (Figure 3.1). However, image segmentation aims to obtain pixel-wise classification, with the output prediction characterized by the same spatial dimension as the input. To address this issue, inverse transformations have been defined to increase the spatial dimension of features maps, including max-unpooling, transpose convolution (or up-convolution), and up-sampling. In essence, max-unpooling corresponds to the inverse operation of max-pooling and leverages the indices of the maxima extracted from the pooling layer, while the transpose convolution is an extension of the convolution, which can handle output with larger spatial dimensions than its input. More specifically, transpose convolution is based on a learnable kernel and maintains a connectivity pattern as opposed to max-unpooling. Finally, up-sampling layers are based on nearest-neighbor, linear, or bi-linear interpolations and thus require no additional trainable parameters [29], [199]–[202].

Hence, **convolutional segmentation networks are typically divided into two components: an encoder that extracts abstract features and a decoder that generates the segmentation given the encoded representations**. The encoder thus progressively increases the number of features and reduces the spatial dimension (i.e., contracting path), while the decoder is symmetric and performs the opposite transformations (i.e., expanding path). In the end, the decoder produces a feature map with the same spatial dimension as the input. In order to perform pixel-wise classification, the last layer is usually based on a point-wise convolution W (i.e., 1×1 kernel which serves as a linear projector) with associated bias b, followed by a `Softmax` non-linearity. The last convolution comprises $C+1$ filters to extract the desired number of targeted objects [29], [199]–[202]. Specifically, if u_i denotes the output of the penultimate layer, then:

$$\hat{y}_i = \texttt{Softmax}(W * u_i + b) \tag{3.6}$$

The `Softmax` function results in a probability distribution $p_{\text{model}}(y|x;\Theta)$ over the segmentation masks defining the $C+1$ classes of interest (see Section 3.2.4). More specifically, each pixel is assigned a vector whose values represent the probability of each of the $C+1$ classes, and the vector is normalized to one. At inference, the class with maximum probability is assigned as the final prediction using an `arg max` function.

For convolutional encoder-decoder based segmentation networks, the trainable param-

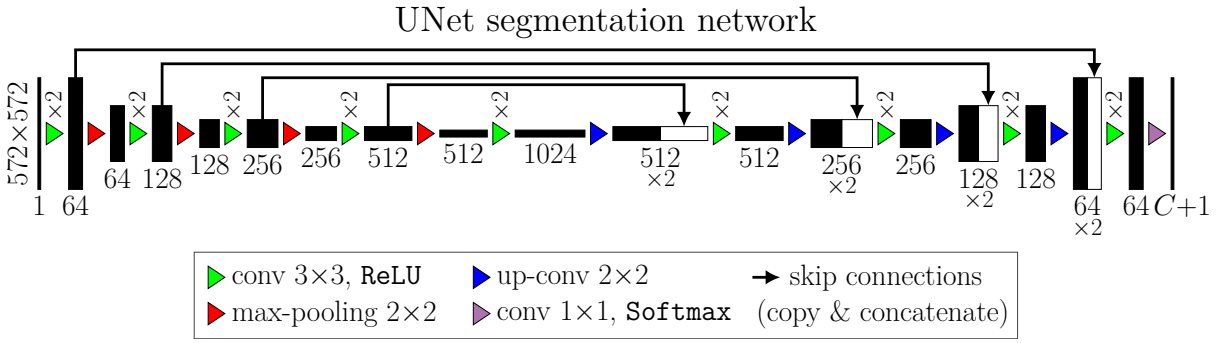


Figure 3.2 – Architecture of the UNet segmentation network [29]. Each black box corresponds to a multi-channel feature map, with the number of channels denoted at the bottom. White boxes represent feature map copied by the skip connections. The colored arrows indicate the different operation including, convolution with non-linearity (i.e., `ReLU` or `Softmax`), max-pooling, and up-convolution.

eters Θ encompass all the convolutional filters (i.e., classical, point-wise, up-convolution) and associated bias. It should be noted that, in the context of binary segmentation (i.e., typically one anatomical structure and background), the `Sigmoid` function is preferred as the last non-linearity, and in this scenario, the loss function is based on binary cross-entropy. Finally, convolutional encoder-decoders are not limited to segmentation and can perform other tasks requiring dense pixel-wise predictions, such as computing optical flow [203] or disparity maps [204] notably used in computer vision.

3.3.4 UNet for medical image segmentation

The motivations and key layers (i.e., convolution, non-linearity, pooling, un-pooling) of the convolutional encoder-decoder architecture have been introduced in the previous section. We should reinforce that the design of deep learning architectures (including convolutional encoder-decoders) remains mostly empirical and based on heuristic approaches. **Nowadays, the convolutional encoder-decoder is one of the most popular backbones for medical image segmentation due to its ability to model the features (e.g., shape, appearance) of multiple objects and the long-range spatial relationships between these structures.** Although the UNet architecture [29] is the most well-known encoder-decoder architecture applied in medical imaging, it is worth mentioning some precursors such as DeconvNet [202], SegNet [200], DeepLab [201], and FCN [199]. While these architectures employed a similar encoder branch based on a traditional succession of convolution, non-linearity, and pooling, the architecture of their decoder

slightly differed, with models based on either max-unpooling (DeconvNet, SegNet), up-sampling (DeepLab), and up-convolutional layers (FCN). Moreover, both DeepLab and FCN employed similar concepts of reusing finer resolution features of the encoder to refine the final output segmentation. In particular, the combination of semantic information from deep, coarse layers with appearance information from shallow, fine layers improved the accuracy of the segmentation [199], [201].

For its part, the UNet architecture [29] employed up-convolutional layers to retrieve the original spatial resolution of the input, as depicted in Figure 3.2. Most importantly, **the addition of skip connections, which copy and concatenate features between the encoder and decoder branches (Figure 3.2), allowed UNet to recover fine-grained details in the prediction and supplement previous architectures**. In particular, the combination of high-resolution features from the contracting path with the up-convolution output resulted in enhanced segmentation localization, with the successive convolution layer learning to assemble a more precise output based on this information (Figure 3.2). Furthermore, it has been empirically noticed that the skip connections help stabilize optimization and convergence by preventing vanishing gradient issues in shallow layers. Nevertheless, there is no theoretical justification for the success and incredible efficiency of symmetrical long skip connections in dense prediction tasks, such as medical image segmentation. Finally, unlike patch-based approaches, UNet is able to simultaneously use the context of the whole image and provide good location prediction [29].

Recently, the medical image research community has employed the UNet architecture (and recent derivatives, Section 3.4) in countless medical applications, each defined by imaging modality (e.g., X-ray, PET, ultrasound, CT, and MRI), anatomical region of interest (e.g., brain, heart, lung, abdomen), and targeted structure to segment (e.g., tumor, metastasis, cyst) [1], [2]. For instance, methods based on UNet were developed for the detection and segmentation of multiple brain metastases on MR images [205] or the distinct bone segmentation from upper-body CT scans [206]. **One can consider each of these application as corresponding to a different function approximation problem with a distinct imaging domain $\mathcal{I}$ and segmentation label space $\mathcal{C}$.** The enhanced segmentation performance obtained in these multiple scenarios illustrates the incredible versatility of the UNet encoder-decoder convolutional architecture. However, it should be emphasized that the architecture and training hyper-parameters (e.g., number of layer, number of features per layer, learning rate, optimizer) need to be fine-tuned by a deep learning practitioner and the weights learned during optimization are

domain and task specific. As mentioned in Chapter 1, these limitations, that are inherent to the data-driven nature of deep learning, can hinder the deployment of such models in real-world scenario. It is worth mentioning that the nn-UNet (no-new-UNet) [207] framework offers the ability to configure some hyper-parameters automatically but is currently limited to the standard UNet architecture as a backbone.

From a more practical perspective, one common limitation, specific to UNet, is its ability to process 2D data, while medical images usually comprise three spatial dimensions (e.g., PET, CT, and MRI). This can limit the accuracy of the segmentation prediction due to the lack of complete spatial context. Furthermore, the network is biased in one spatial direction and would provide poor performance if applied along another axis. Hence, 3D counterparts of UNet based on 3D convolutional layers have been proposed, such as VNet [30] and 3D UNet [208]. Nevertheless, these 3D architectures typically require more computational power and memory consumption than standard 2D UNet. In practice, this results in a more unstable training procedure due to smaller batch size. Moreover, it has been noticed that the performance improvements of 3D models over 2D UNet may not be consistent and rather depend on the task, targeted structures, available imaging data, and computational capacity [209]. This partly explains the ongoing popularity of the 2D UNet over its 3D counterpart among the medical image research community. Consequently, **all the segmentation models employed during this thesis (see Parts II and III) will be based on 2D architectures**.

3.4 Standard modifications of the UNet model

Following its success for medical image segmentation, the UNet model has been adapted, modified, and extended by the research community to improve performance further or tackle new challenges. While providing an exhaustive list is out of scope here, we aim to introduce some key standard extensions of the UNet model, which will be relevant for the rest of this thesis.

3.4.1 Batch normalization

Batch normalization [41], which was independently developed the same year as UNet, aims at improving the convergence speed of CNNs by normalizing their internal activations. Although the first implementation of UNet [29] did not incorporate batch normal-

ization layers, this transformation has now become ubiquitous in the computer vision and medical imaging fields. As previously mentioned, optimizing deep models is a challenging task, partly due to the large number of learnable parameters that are organized into a cascade of layers. During training, the distribution of each layer's inputs changes as the parameters of the previous layers are updated, and this problem is known as the internal covariate shift. This slows the training by requiring lower learning rates and careful parameter initialization [41].

To address these issues, batch normalization provides an elegant way to reparameterize the layers and reduce the problem of coordinating updates across many layers. Batch normalization allows practitioners to use much higher learning rates and to be less careful about initialization, while acting as a regularizer [14], [41]. The batch normalization transformation (BN) is defined as follows [3]:

$$\texttt{BN}_{\beta_l,\gamma_l}(v_{i,l}) = \gamma_l \frac{v_{i,l} - \mu_l}{\sqrt{\sigma_l^2 + \epsilon}} + \beta_l \tag{3.7}$$

where $v_{i,l} = \Theta_l * u_{i,l}$ denoted the feature-map at the l^{th} layer produced by the i^{th} image, μ_l and σ_l are the mini-batch mean and standard deviation. ϵ = 1e-5 is a small positive value added for numerical stability. However, normalizing the mean and standard deviation of a layer can reduce the expressive power of the neural network. To maintain this power, batch normalization employs learnable shift β_l and scale γ_l at each layer. This new parameterization can represent the same family of functions as the previous parameterization, but it is much easier to learn with gradient descent [14], [41].

Following the definition of BN, one can update Equation 3.5 given in Section 3.3.2:

$$u_{i,l+1} = \rho(\texttt{BN}_{\beta_l,\gamma_l}(\Theta_l * u_{i,l})) \tag{3.8}$$

It should be noted that the bias b_l of the convolutional layer can be ignored, as its role is subsumed by the shift β_l of the batch normalization transformation. This succession of convolution, batch normalization, and non-linearity thus defines a new building block for convolutional models. Following its widespread success in computer vision, the batch normalization layer has been adopted in medical imaging, with most recent UNet imple-

3. In practice, batch normalization is performed with respect to the feature m at layer l, as follows: $\texttt{BN}_{\beta_{l,m},\gamma_{l,m}}(v_{i,l,m}) = \gamma_{l,m} \frac{v_{i,l,m} - \mu_{l,m}}{\sqrt{\sigma_{l,m}^2 + \epsilon}} + \beta_{l,m}$

mentations leveraging this normalization [1], [2]. Finally, it should be emphasized that while initial motivation aimed at mitigating the internal covariate shift issue, the exact reasons for the effectiveness of batch normalization are still poorly understood. **Recent works argue that batch normalization does not, in fact, reduce internal covariate shift but instead smooths the objective function, which in turn, induces a more predictive and stable gradients behavior, allowing for faster training and better generalization capabilities [210]. In any case, we integrate batch normalization in the baseline segmentation architecture employed in the remainder of this thesis (see Parts II and III), and we present an extension of this layer for multi-domain learning (see Chapter 6).**

3.4.2 Loss functions specific to image segmentation

In parallel, several works have introduced new loss functions in the context of medical image segmentation. In particular, it has been observed that when the segmentation process targets rare objects, the cross-entropy loss $\mathcal{L}_{\text{CE}}$ can result in sub-optimal performance due to the severe class-imbalance occurring between foreground and background labels. In order to mitigate these imbalanced class scenarios, strategies such as the weighted cross-entropy loss, the focal loss, the Dice loss, or the Tversky loss have been proposed [211], [212]. It should be emphasized that these approaches do not involve any modifications of the UNet encoder-decoder architecture. In particular, the novel loss functions still rely on the `Softmax` activation layer to provide the segmentation output prediction.

One loss function has been of particular interest for the medical imaging community: the Dice loss. The $\mathcal{L}_{\text{Dice}}$ loss is based on the Dice coefficient, which is a widely used metric in computer vision assessing the similarity between two areas labeled by segmentation. The definition of the loss is given below [4]:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{2y_i\hat{y}_i}{y_i + \hat{y}_i + \epsilon} \tag{3.9}$$

The Dice loss ranges from 1 denoting completely dissimilar labelisations to 0 for perfect overlap. Minimizing this loss thus encourages the network to produce segmented images with labels $\hat{y}_i$ similar to the reference labels y_i (considered as ground truth) in the training set. The loss comprises a small positive value ϵ added for numerical stability, needed

4. The full expression of the loss is given by $\mathcal{L}_{\text{Dice}} = 1 - \frac{1}{n(C+1)}\sum_{i=1}^{n}\sum_{c=0}^{C}\frac{\sum_{u\in\Omega}2y_{i,c,u}\hat{y}_{i,c,u}}{\sum_{u\in\Omega}y_{i,c,u} + \hat{y}_{i,c,u} + \epsilon}$.

during optimization. It should be noted that $\mathcal{L}_{\text{Dice}}$ is sometimes referred to as soft or fuzzy Dice loss as it is based on the real value $\hat{y}_i$ prediction, rather than on binary masks (i.e., Boolean set) obtained after applying a threshold (or `arg max` function). Indeed, the Dice coefficient metric, whose definition will be given in Section 3.5.2, usually operates on binary masks for both ground truth and prediction segmentation. However, to be able to use the gradient descent algorithm, one must define a differentiable loss. For this reason, the non-differentiable threshold is omitted, and instead of using an intersection between Boolean sets, one uses the element-wise product to approximate this non-differentiable operation.

While the Dice loss has been reported to improve class imbalance issues, it also allows for directly maximizing the Dice coefficient value during optimization, which is one of the most common metrics used to assess the performance of a segmentation model. In particular, this offers a better comprehension of the optimization process, as opposed to the cross-entropy loss $\mathcal{L}_{\text{CE}}$ which is derived from probability theory and lacks practical insight on the expected performance of a segmentation model. However, the gradients associated with $\mathcal{L}_{\text{Dice}}$ are more complex and can lead to unstable training compared to $\mathcal{L}_{\text{CE}}$. In addition, other metrics than the Dice coefficient are used in practice and it is unclear whether minimizing the Dice loss maximizes the performance of the Dice coefficient at the expense of other metrics. Hence, practitioners may use a linear combination of $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{Dice}}$, known as combo loss.

Finally, with the same goal of optimizing a network to maximize a segmentation metric directly, some works have proposed loss functions leveraging the Hausdorff distance. This metric, which will be presented in Section 3.5.2, is an indicator of the largest segmentation error and, as such, is used extensively in evaluating segmentation algorithms. However, this metric computed between predicted and ground truth segmentation contours is non-differentiable. Hence, various approaches based on boundary distance maps or morphological operations have been developed to obtain differentiable approximations of the Hausdorff distance [213]. Nevertheless, such losses are unstable during training and must be used in combination with $\mathcal{L}_{\text{CE}}$ or $\mathcal{L}_{\text{Dice}}$ losses, especially at the start of the optimization procedure.

In the context of pediatric bone segmentation, we did not observe class imbalance issues and we preferred the $\mathcal{L}_{\text{CE}}$ over $\mathcal{L}_{\text{Dice}}$ which leads to more unstable training. Furthermore, instead of employing a loss based on Hausdorff distance surrogates, our approaches, presented in Parts II and III, leverage shape priors

regularization to enforce globally consistent segmentation predictions and maximize the performance of the network.

3.4.3 Spatial attention gate

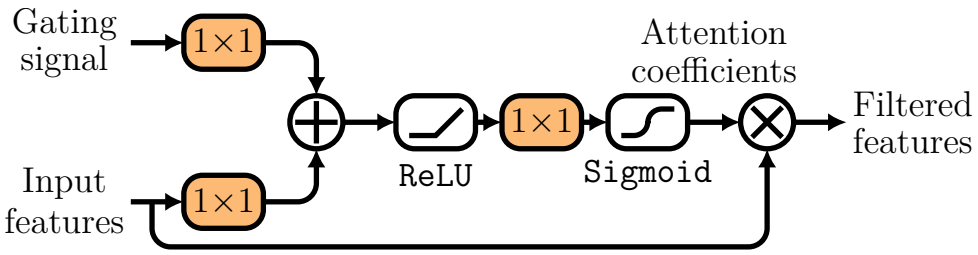


Figure 3.3 – Schematic of the spatial attention gate [42]. The input features are multiplied with attention coefficients to focus on salient regions. Spatial regions are selected by analysing both the activations and contextual information provided by the gating signal. Feature transformations include 1×1 convolution and `ReLU` and `Sigmoid` non-linearity.

One simple extension of the UNet model consists of the addition of spatial attention gates to the skip connections between the encoder and decoder branches. The Attention-UNet (Att-UNet) [42] model exploits attention gates to automatically learns to focus on the anatomical structures of interest. In particular, the attention gates filter the features propagated through the skip connections to suppress irrelevant regions while highlighting salient features useful for the segmentation task. As shown in Figure 3.3, the output of an attention gate is the element-wise multiplication of the input feature-maps originating from the skip connection and attention coefficients. The attention coefficients are computed using both the input features and the gating signal, which are linearly mapped using 1×1 convolutions, then combined through an addition and a `ReLU` non-linearity (i.e., additive attention). The resulting signal contains contextual information to prune lower-level feature responses and determine focus regions. The following 1×1 convolution transforms the signal into a 1-channel dimensional map, and a `Sigmoid` activation then results in the attention coefficient map. Thus, attention coefficients which range from 0 to 1, identify salient image regions (i.e., value close to 1) and prune irrelevant areas (i.e., value close to 0). It has been illustrated that the attention gates consistently improve the sensitivity and prediction accuracy of UNet with minimal computational overhead. Most importantly, the visualization of the obtained attention maps enables us to examine the inference process of segmentation networks. Specifically, one can verify that the network

learns to focus on the correct structures of interest at different scales, as attention gates typically provide a rough outline of the targeted organs [42].

Due to its ability to provide better segmentation localization and interpretable attention maps, the Att-UNet will constitute the baseline model for the experiments performed in Parts II and III. It should be noted that, we also integrate batch normalization (see Section 3.4.1) in our baseline Att-UNet model.

3.5 Technical aspects of deep segmentation in medical imaging

3.5.1 Implementation details

Nowadays, most neural network implementations are based on GPUs due to their high memory bandwidth and massively parallel computing capabilities. Indeed, as neural networks usually involve millions of parameters that must be updated during every step of optimization, the high memory bandwidth of GPUs allows training deep learning models efficiently compared to traditional central processing units (CPUs). Additionally, since each individual parameter can be processed independently from the other parameters in the same layer, neural networks easily benefit from the parallelism of GPU computing. One can also distribute the workload of training and inference across many GPUs in a large-scale distributed system (i.e., data/model parallelism) with more computation power than on a single machine. While GPU hardware was originally specialized for graphics tasks, it has become more flexible over time. In particular, the advent of general purpose GPU, which could execute arbitrary code, is an other important factor in the popularity of graphics cards for neural network training. One should mention the CUDA framework, which provides a way to write and implement code for Nvidia's GPUs. This platform was rapidly adopted by deep learning researchers and used to develop deep learning software such as PyTorch [5] [214], TensorFlow [6] [215], or Keras [7] [216]. These libraries have been of particular interest to deep learning practitioners, as they offer a free, open-source, and simple Python interface to develop their own deep learning algorithms. In particular, Keras

5. `https://pytorch.org/`
6. `https://tensorflow.org/`
7. `https://keras.io/`

libraries uses TensorFlow as backend and provides a modular and extensible framework to implement neural networks, while PyTorch offers more flexibility through its dynamic computational graph used to optimize deep learning models. It should be noted that other deep learning software and libraries have been developed. Nevertheless, we refrain from providing such an exhaustive list which is outside the scope of this thesis. These software have been used in endless applications, including computer vision, speech recognition, natural language processing, and most notably in the context of this thesis, medical image analysis [14], [214]–[216].

The neural networks developed during this thesis were implemented in Python using either PyTorch, Tensoflow, or Keras libraries. Although the combination of Tensoflow and Keras libraries provided a simple framework for implementing the multi-structure segmentation methods developed in Part II. **For multi-task, multi-domain learning (see Part III), it was essential to employ the PyTorch library to implement the domain-specific layers (see Section 6.2.2) which relied on a dynamic computational graph.** Indeed, in the Tensoflow and Keras coding paradigms, the computational graph that defines a deep learning model is static and cannot be modified during training or inference [215], [216]. On the contrary, PyTorch allowed us to implement dynamic graphs that enabled us to use control flow statements and to change the operations performed at every iteration [214]. In practice, the domain-specific layers needed to be selected according to the domain of the input domain and the computational graph of the model is thus dynamically modified during learning.

As previously mentioned, these deep learning packages provide access to traditional layers (e.g., convolution, activation, pooling), loss functions (e.g., cross-entropy, mean-squared error), and optimization algorithms (e.g., Adadelta, AdaGrad, Adam, or RMSprop). Each of these can be easily parameterized and tailored for very diverse scenarios. For instance, one can define a convolutional layer with a specific number of filters, kernel size, and stride. Ultimately, one can assemble these simple building blocks into complex deep learning models. It should be noted that no matter the depth and complexity of the model, the gradients used to optimize the models are computed automatically without any user intervention. In particular, a reverse automatic differentiation system keep records of the operations (i.e., layers) applied on the input data, and provides the directed acyclic computational graph used to back-propagate the gradients (see Section 3.2.2). Hence, once the model, loss function, and optimization algorithm have been selected, the training procedure is straightforward for the deep learning practitioner. Furthermore, these

deep learning software packages give access to state-of-the-art architectures (e.g., `VGG19`, `DenseNet121`, `ResNet50`, `InceptionV3`, `EfficientB3`) with weights pre-trained on large natural image datasets [214]–[216]. We will employ these models in Chapters 5 and 7, in the context of transfer learning.

All the deep learning architectures implemented in this thesis were optimized using a Nvidia RTX 2080 Ti GPU with 12 GB of RAM. As previously mentioned, the amount of available memory is an important factor in the design of neural networks and the selection of optimization hyper-parameters. Most notably, the depth and width of the employed networks, as well as their respective batch size, were selected based on a compromise between performance and hardware capacity. Hence, the results obtained in this thesis could be potentially improved if one had access to a higher computational power and memory capacity. In the same line of thought, it should also be emphasized that we did not employ any multi-GPU computing techniques due to resource constraints. Nevertheless, all networks were trained with extensive on-the-fly data augmentation to teach the models the desired invariance, covariance, and robustness properties, consequently improving generalization performance. Data augmentation comprised random geometric deformations, which were applied to both grayscale intensity images and corresponding annotation maps (reference labels).

Following standard machine learning guidelines, our frameworks included pre-processing and post-processing steps. For each dataset introduced in Chapter 2, we extracted 2D images from the 3D MR imaging data along a fixed axis (axial for shoulder images and sagittal for ankle and knee images). The corresponding 2D annotations were extracted in a similar manner from the 3D ground truth labels. As previously mentioned, all the networks developed in this thesis were based on 2D architectures. In the following pre-processing step, the obtained 2D images and annotation maps were downsampled to 256×256 pixels. This specific value was chosen due to memory limitations. The grayscale intensity images were then normalized to have zero-mean and unit variance for each dataset, while the ground truth segmentations were encoded as one-hot vectors with $C + 1$ classes. At inference, the predicted 2D segmentation were stacked to form a 3D volume. The predicted probabilities resulting from the `Softmax` last layer were transformed to final class prediction using an `arg max` function. Finally, post-processing also included connected component analysis to ensure that each predicted bone was represented by a connected set and binary morphological operations [217] to smooth the resulting boundaries.

3.5.2 Performance metrics for medical image segmentation

As for any machine learning algorithm, the performance of deep learning medical image segmentation models is evaluated on unseen test data to assess the model's generalization capabilities. In machine learning, it is also recommended to build an additional separated validation set that can be used to fine-tune the hyper-parameters of the segmentation algorithm (e.g., learning rate, batch size). Therefore, during experimentation, medical imaging datasets must be split between training, validation, and testing sets. In practice, when developing 2D models for 3D image segmentation, the partition between training, validation, and test data is performed with respect to the 3D examinations to avoid any data leakage issues during experiments. Furthermore, in the context of this thesis, due to the low amount of 3D examinations in each pediatric dataset (see Section 2.4.2), we employed a leave-one-out strategy in which one 3D examination was retained for testing, one for validation, and the rest consisted in the training set. We iterated through the dataset so that each sample was used once for testing and once for validation. This allowed us to obtain a reliable and unbiased estimate of the model performance.

To assess the performance of segmentation algorithms, medical image researchers have access to a wide array of metrics based on region overlap (e.g., Dice coefficient, intersection over union, sensitivity, specificity, Jaccard index, F1 score), distance between boundaries (e.g., maximum symmetric surface distance, average symmetric surface distance, Hausdorff distance 95% percentile, normalized surface distance), volume metrics (e.g., relative absolute volume difference, symmetric relative volume difference), or connectivity metrics (e.g., centerline Dice similarity coefficient). These metrics are computed between the segmentation prediction generated by the algorithm and the ground truth produced by an expert. These scores can be extracted from either 2D or 3D masks. Each of these metrics aims at quantifying distinct information about the prediction errors provided by the segmentation model, and each presents its own advantages and drawbacks. **In medical imaging, it is thus essential to employ multiple metrics to have a complete assessment of the performance and errors of a segmentation model [218].**

In this thesis, we used six metrics to assess the performance of the models developed in Parts II and III. The metrics were computed using the 3D segmentation masks to provide clinically relevant evaluation of the performance of our models. These metrics included the Dice coefficient, sensitivity, specificity, maximum symmetric surface distance (MSSD), average symmetric surface distance (ASSD), and relative absolute volume difference (RAVD). The metrics were defined as follows, let GT and P be the ground

truth and predicted 3D segmentation masks of one specific structure of interest and let S_{GT} and S_P be the surface voxels of the corresponding sets.

$$\text{Dice} = \frac{2|GT \cap P|}{|GT| + |P|} \tag{3.10}$$

$$\text{Sensitivity} = \frac{|GT \cap P|}{|GT|} \tag{3.11}$$

$$\text{Specificity} = \frac{|\overline{GT} \cap \overline{P}|}{|\overline{GT}|} \tag{3.12}$$

$$\text{MSSD} = \max(h(S_{GT}, S_P), h(S_P, S_{GT})) \tag{3.13}$$

$$\text{with } h(S, S') = \max_{s \in S} \min_{s' \in S'} \|s - s'\|_2$$

$$\text{ASSD} = \frac{1}{|S_{GT}| + |S_P|} \left(\sum_{s \in S_{GT}} d(s, S_P) + \sum_{s \in S_P} d(s, S_{GT}) \right) \tag{3.14}$$

$$\text{with } d(s, S') = \min_{s' \in S'} \|s - s'\|_2$$

$$\text{RAVD} = \frac{|\,|GT| - |P|\,|}{|GT|} \tag{3.15}$$

Here, $|.|$ denoted the number of elements equal to 1 in the binary set and $\overline{.}$ indicated the complement of the set. It should be emphasized that GT anf P corresponded to a binary set representing only one targeted structure and P was obtained after applying the post-processing steps described in Section 3.5.1.

As previously mentioned, the Dice coefficient is the most widely used metric in medical image analysis. It measures the similarity between two sets ranging from 0 (completely dissimilar) to 1 (perfect overlap). The Dice coefficient is identical to the F1 score and closely related to the intersection over union (also referred to as the Jaccard index). In practice, it is sufficient to calculate only one of these measures [218]. For their part, sensitivity and specificity evaluate the true positive and true negative rates, which provide complementary information with respect to Dice. The sensitivity and specificity metrics indicate the proportion of targeted object and background correctly detected, respectively. As for Dice, sensitivity and specificity range from 0 for completely incorrect detection, to 1 for perfect detection. The MSSD and ASSD assess the models' ability to generate the same contours as those produced manually. In particular, the MSSD calculates the maximum of all shortest distances for all points from one object boundary to the other, while the ASSD measures the average of all distances for every point from one object to

the other. For both metrics, a value of 0 refers to a perfect prediction (i.e., distance of 0 to the reference boundary), while no fixed upper bounds exist [218]. It should be noted that the MSSD is also referred as the symmetric Hausdorff distance. A major problem related to these boundary-based metrics is the error-prone reference annotations, as domain experts often disagree on the definition and annotation of objects and their boundaries. Finally, the RAVD metric determines the volumetric difference between volumes, with 0 indicating that the prediction presents the same volume as the ground truth. This metric is of particular interest in pediatric musculoskeletal segmentation as bone volumetric information allows clinicians to assess impaired development quickly or easily compare pathological and healthy patients. However, it should be noted that this metric does not consider the location of the objects as opposed to region overlap and boundary-based metrics [218]. **In this thesis, Dice, specificity, sensitivity, and RAVD metrics are denoted as percentages. For their part, MSSD and ASSD distance measures are transformed to millimeters using voxel size information extracted from DICOM metadata.**

After computing segmentation metrics, one usually needs to assess whether the performance of a proposed segmentation algorithm corresponds to a statistically significant improvement over the baseline method. To do so, one must perform a statistical analysis of the results obtained by each model. **In this thesis, due to the scarcity of 3D pediatric examinations, we used metrics computed on 2D slices to perform tests with enough statistical power. Additional details on the statistical analyses are provided in Chapters 5 and 7.** Furthermore, although it is essential to employ complementary metrics to assess the performance of segmentation models, it can reveal challenging to simultaneously compare the performance of various segmentation strategies across multiple metrics [218], [219]. Hence, a common approach to mitigate this issue involves designing a ranking system that aggregates all the segmentation metrics into a unique score. The different segmentation algorithms can then be ranked using this score. Creating a ranking system can be arduous and strongly depends on the targeted application and employed metrics. **In the context of this thesis, we employed 3D metrics to conceive our ranking system, which will be further developed in Chapter 5. Finally, for segmentation it is also essential to perform visual comparison between the algorithms. In this direction, we performed extensive qualitative validation of our methods in Parts II and III.**

3.6 Conclusion

This chapter introduced the mathematical framework for deep learning-based image segmentation, which is formulated as a function approximation problem. However, solving this optimization problem is a challenging task due to the highly dimensional and non-convex nature of deep learning models. Moreover, we have seen that almost all deep learning segmentation models rely on the convolutional encoder-decoder architecture, UNet being a widely used example. We also presented standard modifications of the UNet model, which will be used as a baseline in the rest of this thesis (see Parts II and III). Finally, this chapter described the more technical aspects of this thesis, including the hardware and software implementation details and the metrics and schemes used to assess the performance of the proposed models.

As mentioned in Chapter 1, despite the great potential of deep learning, real-world deployment of neural networks remain limited in clinical practice. We previously identified three main technical challenges of medical image analysis at the age of deep learning, as well as novel deep learning paradigms to mitigate these issues. **In this thesis, we propose to develop novel methodologies addressing the generalization gap and data scarcity issues in the context of pediatric musculoskeletal image analysis (see Chapter 2)**. These methods are built upon novel deep learning paradigms, including state-of-the-art network architectures (Chapters 5 and 7), multi-domain learning (Chapters 6 and 7), multi-task schemes (Chapters 6 and 7), transfer learning (Chapters 5 and 7), and prior knowledge embedding (Chapters 4, 5, 6, and 7).

Part II

Improved multi-structure segmentation via combined regularization from shape priors and adversarial networks

Chapter 4

Shape priors-based regularization for multi-structure segmentation

4.1 Introduction

This chapter focuses on the data scarcity issue associated with pediatric imaging databases, which makes it challenging to develop deep learning models that produce robust delineations on unseen images. As discussed in Part I, since deep learning is a data-driven methodology requiring a large amount of training data, the lack of pediatric imaging resources can induce over-fitting issues and insufficient generalization capabilities on new examinations. In particular, the trained neural networks may be unable to produce accurate segmentation predictions on test images, limiting their applications in real-life clinical scenarios.

To mitigate these issues, **recent works aim to incorporate regularization into deep learning-based segmentation models to further avoid over-fitting and improve generalizability**. In deep learning, the regularization concept covers techniques that can affect the network architecture, the learned weights, the training data, or the loss function [43]. As introduced in Chapter 3, the UNet architecture [29] already contains regularization in the form of convolutional layers which enforce local and translation-equivariant hidden units, pooling-layers that impose translation invariant feature extraction, and skip-connections which assume a correlation between low-level and high-level features [13], [14], [28]. Moreover, data augmentation and batch normalization are two data-based regularization techniques that are commonly incorporated into deep learning models. Data augmentation incites the network to learn invariance, covariance, and robustness properties [14] while the randomization inherent in batch normalization enforces robust data representations [41].

Concerning regularization schemes affecting network weights, one can mention transfer learning [44] and dropout [220]. In deep learning, transfer learning refers to employing

weights pre-trained on a similar image domain or task. Transfer learning from natural image datasets, particularly ImageNet, has proven to be a successful approach for medical image analysis. Indeed, transfer learning assumes that low-level features are shared between image domains and tasks, and the pre-trained weights provide a robust initialization for optimization [44]. For its part, the key idea of dropout is to randomly drop layers from the neural network during training. The motivation is to prevent model over-fitting, and dropout can thus be considered as a bagging (i.e., bootstrap aggregating) approach involving multiple sub-networks [220].

Finally, in machine learning, one of the simplest and most common loss function-based regularization is the L_2 norm penalty, known as weight decay or Tikhonov regularization. This regularization strategy consists in enforcing the weights of the model to be close to zero and preventing over-fitting [14]. Weight decay was initially introduced to regularize ill-posed linear regression problems, and the theoretical motivations for using this type of regularization have been extensively studied (i.e., existence, uniqueness, and stability of the solution) [221]. In the context of deep learning, weight decay has been applied to neural networks and often leads to improved performance in practical settings [222]. However, the mathematical motivation for its use is less clear [223]. Similarly, one can mention the L_1 norm penalty, which enforces sparse weights (i.e., most weights equals to zero). The properties of this regularization are clearly understood in the linear case, but its effect on deep learning models remains uncertain [14]. Even though studies dedicated to loss function-based regularization are infrequent, defining a suitable loss function for training deep learning models can lead to improved performance. In particular, **recent works aim at designing regularization terms specific to deep learning-based medical image segmentation**.

In medical imaging, such regularization schemes can arise from different prior information related to the anatomical structures of interest, such as boundaries [47], shape models [48], atlas models [49], or topology. Exploiting prior knowledge is found to be effective in achieving more precise and consistent results for traditional (i.e., machine learning) medical imaging segmentation applications [50]. Specifically, regularization techniques can alleviate the presence of image artifacts that are inherently embedded in an image during its acquisition [50]. Following this, recent works aim at incorporating similar regularization constraints into deep learning-based segmentation models. In this context, one particular loss function-based regularization methodology has shown promising results: shape priors-based regularization [51]–[55]. Most importantly, **shape regularization appears**

as a key strategy to enhance segmentation outcomes and model generalization abilities when targeting scarce pediatric imaging datasets.

Incorporating shape information into medical imaging segmentation algorithms has already proven to be useful in reducing the effect of noise, low contrast, and artifacts [53]. Recent contributions have proposed to learn a representation of the anatomy directly from ground truth annotations using a deep auto-encoder [51]–[55]. Due to the constrained nature of anatomical structures (global position and shape of bones, see Chapter 2), data-driven models are suitable for learning shape prior information. **It should be emphasized that as opposed to the definition of shape given by Kendall, which does not consider (i.e., quotients out) translations, rotations, and dilations [133], shape priors-based on deep auto-encoders and 2D segmentation masks integrate position, orientation, and size characteristics in addition to "pure Kendall" shape features.** The learned non-linear shape representation can then be integrated into the segmentation network during optimization, thanks to specifically designed regularization schemes. For instance, Dalca et al. employ the decoder component of the auto-encoder as a shape prior during training [51], while other works propose to directly regularize the segmentation network by projecting the predicted segmentation into the shape space using the encoder component of the auto-encoder [53], [55]. These approaches rely on a regularization term based on a distance loss (e.g., Euclidean) which enforces the predicted segmentation to be close to the ground truth in shape space [53], [55]. Consequently, such regularization encourages globally consistent shape predictions.

Standard deep learning architectures (e.g., UNet [29] or VNet [30], see Section 3.3.4) have already been applied for the segmentation of musculoskeletal structures in MR images, including knee bones, muscles, cartilages, and ligaments [31]–[36], shoulder bones [37], wrist cartilage [38], and thigh muscle [39]. One can also mention works targeting the segmentation of musculoskeletal structures in other imaging modalities such as metacarpal bones in CT scans [224], knee cartilage in ultrasound images [225], temporal bone skull in CT images [226], [227], whole-skeleton bones in upper-body CT scans [206], [228], metastasis in thorax bone SPECT images [229], bone and bone lesion in PET/CT scans [230], and bone tumors in X-ray radiographs [231]. However, all of these models were developed for the adult population. As discussed in Chapter 2, studies dedicated to pediatric musculoskeletal image segmentation remain scarce in the literature, except for following works targeting pediatric elbow in X-ray [178], pediatric abdominal skeletal muscles in CT [179], and pediatric shoulder muscles in MR [40]. To the best of our knowledge, the

literature on deep learning-based pediatric bone segmentation remains rare.

For bone segmentation, post-processing schemes based on conditional random field [232], deformable models [32], or statistical shape models [31] have been developed to constrain the predicted shapes. However, these methods fail to regularize and incorporate shape information directly into the segmentation network as opposed to loss function-based regularization techniques [53], [55]. Furthermore, as segmentation of the musculoskeletal system typically involves multiple anatomical structures and tissues, two segmentation strategies emerge in the literature: in the first one, a single network predicts all segmentation classes [31], [32], [37], whereas, in the second one, specific networks are trained for each object of interest [40]. **In the context of pediatric bone segmentation, it remains an open question whether bone structure specialization or exploitation of features shared between bones provide better performance.**

4.1.1 Contributions

In this chapter, we propose an automatic and multi-object pediatric bone segmentation method for scarce MR images. To address this limitation, our framework leverages auto-encoder based shape priors to guide the segmentation network to make anatomically consistent predictions with restricted imaging resources. Furthermore, we illustrate that the proposed approach can be easily integrated into various bone segmentation strategies, and we demonstrate the effectiveness of employing a multi-structure learning scheme. Finally, we assess the learned shape representations and model's interpretability through the t-SNE dimensionality reduction algorithm [87] and attention map visualization.

The research conducted in this part has been published in the Artificial Intelligence in Medicine journal [233] and substantially extends a preliminary work presented at the IEEE International Symposium on Biomedical Imaging (ISBI) [234].

The remainder of this chapter is structured as follows. Section 4.2 consists of an overview of the baseline deep learning segmentation framework (Section 4.2.1) and the integration of shape priors-based regularization (Section 4.2.2). The experiments are explained in Section 4.3 and encompass the assessment of multiple multi-structure learning schemes (Section 4.3.2) and the description of the implementation details (Section 4.3.3). Finally, the results are reported and discussed in Section 4.4. Most importantly, we validate the proposed multi-bone pediatric segmentation method based on shape priors (Section 4.4.1) and assess the learned shape representations (Section 4.4.2).

4.2 Integrating shape priors-based regularization into deep segmentation networks

4.2.1 Baseline deep segmentation framework

As introduced in Section 3.2, let $\{x_i, y_i\}_{1 \leq i \leq n}$ be a training set of n couples of images and corresponding labels (segmentation maps). The grayscale image x_i is in the intensity space $\mathcal{I}$ while the corresponding image class labels y_i represent different anatomical objects of interest (plus background) in label space $\mathcal{C}$. In CNN-based segmentation approaches, the aim is to learn a mapping $S : x_i \mapsto S(x_i; \Theta)$ between intensity x_i and class labels y_i images. The function S is a segmentation network composed of a succession of layers whose parameters Θ must be optimized during training. In the following, we note $\hat{y}_i = S(x_i; \Theta)$ as the estimate of y_i having observed x_i. During training, we optimized the loss function $\mathcal{L}$ using stochastic gradient descent to estimate the optimal Θ^* weights.

The segmentation network S is based on the Att-UNet architecture [42] which extends the standard UNet convolutional encoder-decoder [29] with additional attention gates embedded into the skip connections (see Section 3.4.3). Spatial attention gates leverage contextual information from the decoding branch to focus on salient features while suppressing irrelevant areas. More importantly, the attention coefficients for each skip connection aggregate information from multiple imaging scales to achieve better performance [42]. For their parts, the encoder and decoder branches follow the design of UNet [29], with a 512 dimensional feature map corresponding to the central part between the contracting and expanding paths. Encoding layers are composed of a set of convolutional filters with 3×3 kernel followed by batch normalization (`BN`), `ReLU` activation function, and max-pooling with stride 2×2. Similarly, decoding layers are composed of a set of up-convolutional filters with a kernel 2×2 followed by symmetric convolution filters, `BN`, and `ReLU`. A final 1×1 convolutional layer with `Softmax` activation function achieved pixel-wise segmentation.

We employed a loss function based on cross-entropy defined as follows:

$$\mathcal{L}_{\text{CE}} = \frac{1}{n} \sum_{i=1}^{n} -y_i \log(\hat{y}_i) \tag{4.1}$$

The segmentation model is trained through a loss function that operates on individual pixel-level class predictions. As mentioned in Section 3.2, in practice, the full expression

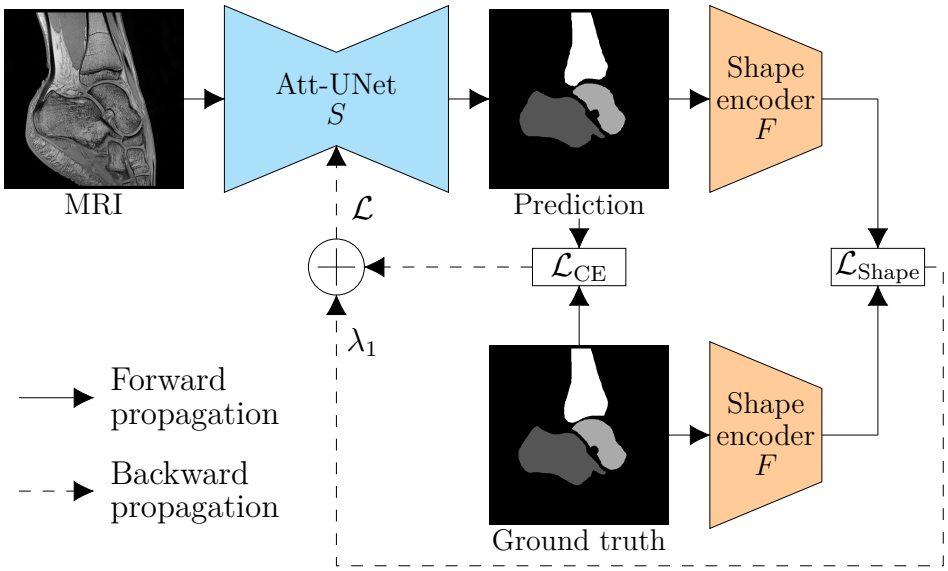


Figure 4.1 – $\text{ShapeReg}^{\text{Multi}}_{\text{Att-UNet}}$ optimization framework composed of multi-structure segmentation network S based on Att-UNet [42] exploiting cross-entropy loss $\mathcal{L}_{\text{CE}}$ and shape priors-based $\mathcal{L}_{\text{Shape}}$ regularization computed by a shape encoder F with fixed weights. The shape encoder corresponds to the encoder component of an auto-encoder previously optimized on ground truth segmentation masks and λ_1 is an empirical weighting hyper-parameter.

of the loss is an average over classes and pixels. However, this pixel-wise loss function fails to exploit contextual inter-structure relationships arising from segmentation masks. Indeed, this loss integrates regional context through the receptive field of the network but fails to include global context [53]–[55]. Hence, we propose to incorporate additional regularization terms which assess the global similarity between predicted and ground truth masks.

4.2.2 Incorporating shape priors-based regularization

In the context of medical image segmentation, one can assume that the ground truth segmentation masks lie in a manifold of true shape, due to the constrained nature of anatomical structures. However, the output prediction of a segmentation network may not lie on the true shape manifold, and it is hence needed to perform a projection onto the correct manifold [53], [55]. While many choices exist for linear and non-linear representations of segmentation shape priors, a convolutional auto-encoder allows us to efficiently learn such low-dimensional shape representation from ground truth segmentation masks,

and to easily compute the projection of segmentation masks using its encoder component [55].

Specifically, an auto-encoder is a neural network composed of an encoder $F : y_i \mapsto F(y_i; \Theta_F)$ and a decoder $G : F(y_i; \Theta_F) \mapsto G(F(y_i; \Theta_F); \Theta_G)$. Θ_F and Θ_G are respectively the learnable parameters of F and G. The encoder F maps the input to a low-dimensional feature space and the decoder G reconstructs the original input from the compact representation. After optimizing the auto-encoder, its encoder component is able to produce a feature map $F(y; \Theta_F)$ which compactly encodes the most salient characteristics of the input mask and each value represents a global feature of a crop of the input binary mask.

The auto-encoder consists of several encoding and decoding layers. Encoding layers are composed of a set of convolutional filters with 3×3 kernel followed by `BN`, `ReLU` activation function, and max-pooling with stride 2×2. Similarly, decoding layers are composed of a set of deconvolutional layers with kernel 2×2 followed by symmetric convolution filters, `BN`, and `ReLU`. A final 1×1 convolutional layer followed by a `Sigmoid` activation function produces the final reconstruction. Following this architecture design, we hypothesize that the learned shape representation is invariant to small translation due to the presence of max pooling layers which induce translation-invariance properties.

The auto-encoder training procedure minimizes a loss function $\mathcal{L}_{\text{AE}}$ which penalizes the reconstruction $(G \circ F)(y_i) = G(F(y_i; \Theta_F); \Theta_G)$ for being dissimilar from the original input y_i. Usual training schemes are based on mean-squared error, Dice, or cross-entropy loss to enforce the auto-encoder to learn the global shape features arising from ground truth annotations [53], [55]. The cross-entropy loss function is employed to optimize both encoder and decoder weights Θ_F and Θ_G, as follows:

$$\mathcal{L}_{\text{AE}} = \mathcal{L}_{\text{CE}} := \frac{1}{n} \sum_{i=1}^{n} -y_i \log(G(F(y_i; \Theta_F); \Theta_G)) \tag{4.2}$$

As a first step, the auto-encoder was trained on ground truth annotations using cross-entropy to learn a shape space in the form of a non-linear low-dimensional manifold which is simply represented by the latent space at the output of its encoder component. While some approaches leverage contour information (e.g., Hausdorff distance) to enforce shape constraints [213], our shape representation is based on complete segmentation masks. **In practice, an auto-encoder would have difficulty to learn a shape representation based on contours due to the high imbalance between contours and background pixels. Hence, we employed a mask-based shape regularization and**

the architecture of the convolutional auto-encoder incorporated traditional convolutional and up-convolutional layers.

After training the auto-encoder, we integrated its encoder component into the baseline segmentation network by computing a shape regularization term $\mathcal{L}_{\text{Shape}}$. To this end, both predictions and ground truth labels were projected onto the latent shape space by the shape encoder with learned weights Θ_F (Figure 4.1). The shape regularization term computed the Euclidean distance between both latent shape representations [53], as follows:

$$\mathcal{L}_{\text{Shape}} = \frac{1}{n}\sum_{i=1}^{n} \|F(\hat{y}_i;\Theta_F) - F(y_i;\Theta_F)\|_2^2 \tag{4.3}$$

The shape regularization loss enforced the predicted segmentation to be in the same low-dimensional manifold as the ground truth mask (i.e., true shape manifold) and thus encouraged anatomically consistent class label prediction [53]. More precisely, minimizing the Euclidean distance led to similar feature maps at the output of the shape encoder (i.e., shape codes) for both segmentation masks. It should be emphasized that **because the weights of the shape encoder were fixed, the two feature maps were in correspondence, with each value encoding the same global shape feature for both ground truth and predicted segmentation masks. However, due to the black-box nature of deep learning models, the interpretability of each shape feature remained limited in practice**. We combined both cross-entropy and shape regularization losses during training, and the updated optimization problem was defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{Shape}} \tag{4.4}$$

where λ_1 was an empirically set weighting factor. The optimization framework of the model $\text{ShapeReg}^{\text{Multi}}_{\text{Att-UNet}}$ is summarized in Figure 4.1.

4.3 Multi-structure segmentation experiments

4.3.1 Imaging datasets

Experiments were conducted on the ankle and shoulder datasets presented in Section 2.4.2, both composed of a mixture of pathological and healthy examinations. We extracted 17 ankles MR images from the ankle joint database including 7 pathological ($A_{P,1}, ..., A_{P,7}$) and 10 healthy ($A_{H,1}, ..., A_{H,10}$) cases. For their part, the 15 MR images of shoulder joints

consisted of 7 pathological ($S_{P,1}, ..., S_{P,7}$) and 8 healthy ($S_{H,1}, ..., S_{H,8}$) examinations. The knee dataset introduced in Chapter 2 was omitted during these experiments as access was granted only to develop our multi-task, multi-domain framework, which will be presented in Part III. For each dataset, all 2D slices were downsampled to 256×256 pixels and intensities were normalized to have zero-mean and unit variance.

4.3.2 Experimental setups

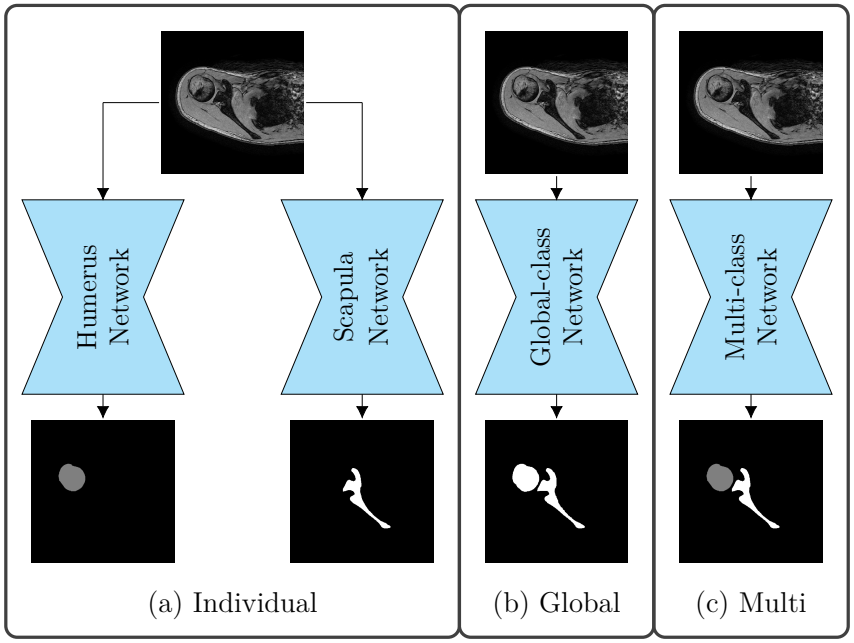


Figure 4.2 – Proposed bone segmentation strategies: (a) individual strategy comprising a specific network for each bone of interest, (b) global strategy constituted of a unique bone class, and (c) multi strategy based on segmentation maps containing multiple classes. The shape encoder was omitted for brevity.

We employed the Att-UNet as backbone architecture to investigate three bone segmentation strategies and to assess which approach would enforce better segmentation outcomes. The segmentation strategies were based on individual-class, global-class, and multi-class labels (Figure 4.2). In the individual-class scheme, we trained an Att-UNet and an auto-encoder for each anatomical class of interest on the individual-class binary masks. The individual-class networks were optimized on each class of interest, and the learned weights were thus specific to a single bone. For the global-class approach, we concatenated the different bone classes into a unique bone-class (i.e., no distinction between

bones), and the learned weights were specific to the global bone class. Finally, in the multi-class strategy, the networks were trained on ground truth segmentation maps containing multiple classes, and the learned weights were shared across all anatomical structures. **We evaluated whether the multi-class strategy promoted more accurate bone segmentation than individual and specific networks, while the global-class approach corresponded to an intermediate strategy between multi-class and individual-class schemes.**

Due to the presence of individual-class (respectively global-class) binary masks in the individual (global) scheme, we modified the last activation function of the individual (global) Att-UNet and auto-encoder from a `Softmax` function to a `Sigmoid` activation resulting in a binary one-channel prediction map. Consequently, we used a binary cross-entropy loss instead of a multi-class cross-entropy loss function during optimization (see Section 3.3.3). The input of the auto-encoder (i.e., segmentation mask) was also characterized by one channel. Moreover, in the individual-class scheme, as the predictions produced by the different networks were independent, a pixel could be predicted as belonging to several classes simultaneously. In this eventuality, we selected the class with the highest probability (i.e., prediction with the highest confidence). Finally, to perform a fair comparison between bone segmentation strategies, predicted individual and multi segmentation masks were transformed into global segmentation masks (i.e., global bone class and background).

Furthermore, to evaluate the contributions of the regularization term, we performed an ablation study for each bone segmentation strategies and compared the baseline Att-UNet and Att-UNet with shape priors-based regularization. The hyper-parameters λ_1 was fixed to 0 to train baseline Att-UNet to ensure a fair comparison. All training hyper-parameters (except λ_1) remained fixed across all methods and all networks were trained from scratch, without relying on any transfer learning and fine-tuning scheme.

The proposed method based on a multi-class Att-UNet with shape priors regularization is referred to as $\text{ShapeReg}^{\text{Multi}}_{\text{Att-UNet}}$. Ultimately, we simultaneously compare baseline Att-UNet and Att-UNet with regularization in individual, global, and multi bone segmentation strategies.

4.3.3 Implementation details

Our training method consisted of two steps. The auto-encoder was first trained using the cross-entropy loss. We explored different hyper-parameters: Adam optimizer with

Network	Batch Size	#Epochs	Learning Rate	#Param.
Auto-encoder	32	10	1e−2	3.1M
Att-UNet	32	20	1e−3	7.9M

Table 4.1 – Summary of the networks employed during experiments: auto-encoder and Att-UNet [42]; along with their corresponding number of trainable parameters and training hyper-parameter values (batch size, number of epochs and learning rate).

initial learning rate 1e-2, batch size set 32 and 10 epochs were found to be optimal. As a second step, we trained the segmentation network using the Adam optimizer with initial learning rate set to 1e-3, batch size set to 32 and number of epochs set to 20 (Table 4.1).

We explored different regularization weighting parameters values and observed $\lambda_1 =$ 1e-1 to be the optimal value. All networks were trained on 2D slices with extensive on-the-fly data augmentation due to limited available training data. Data augmentation comprised random scaling ($\pm 20\%$), rotation ($\pm 20°$), shifting ($\pm 20\%$), and flipping in both directions to teach the networks the desired invariance, covariance and robustness properties. Deep learning architectures were implemented using Keras (on top of TensorFlow) and optimized using a Nvidia RTX 2080 Ti GPU with 12 GB of RAM (see Section 3.5.1).

As a post-processing step, the obtained 2D segmentation masks were stacked together to form a 3D volume. In the individual-class and multi-class schemes, for each anatomical structure, we selected the largest connected set as final 3D predicted mask. In the global-class scheme, we retained the C largest connected sets with C corresponding to the number of bones of interest (3 for ankle and 2 for shoulder). Finally, we applied morphological closing ($5\times5\times5$ spherical kernel) to smooth the resulting boundaries.

4.3.4 Assessment of predicted segmentation

To assess the performance of the different methods, the accuracy of the generated 3D segmentation masks were evaluated against manually annotated ground truths. We computed the Dice coefficient, sensitivity, specificity, maximum symmetric surface distance (MSSD), average symmetric surface distance (ASSD) and relative absolute volume difference (RAVD), as described in Section 3.5.2. We emphasize that although the methods were based on 2D architectures, the segmentation metrics were calculated on 3D volumes. As already mentioned in Section 4.3.2, to perform a fair comparison between bone segmentation strategies, all metrics were computed on global segmentation masks

	Method			Dice ↑	Sens. ↑	Spec. ↑	MSSD ↓	ASSD ↓	RAVD ↓
Ankle	Att-UNet	Indiv	Base	84.0 ± 6.4	82.7 ± 9.5	99.3 ± 0.8	20.4 ± 12.2	2.0 ± 1.4	17.9 ± 16.0
			ShapeReg	87.1 ± 3.9	85.8 ± 7.3	$\underline{99.5 \pm 0.4}$	18.4 ± 11.2	1.5 ± 0.6	$\underline{11.5 \pm 7.5}$
		Glob	Base	87.6 ± 7.8	$\mathbf{92.6 \pm 5.4}$	99.0 ± 1.2	19.9 ± 14.7	1.8 ± 1.6	17.9 ± 27.5
			ShapeReg	87.8 ± 6.2	$\underline{90.6 \pm 7.7}$	99.1 ± 1.0	18.0 ± 12.3	1.7 ± 1.4	13.7 ± 14.3
		Multi	Base	$\underline{88.4 \pm 6.2}$	86.6 ± 10.1	$\mathbf{99.6 \pm 0.4}$	$\underline{17.0 \pm 12.4}$	$\underline{1.3 \pm 0.9}$	12.5 ± 10.8
			ShapeReg	$\mathbf{89.9 \pm 6.2}$	89.1 ± 9.1	$\mathbf{99.6 \pm 0.3}$	$\mathbf{11.1 \pm 4.3}$	$\mathbf{1.0 \pm 0.6}$	$\mathbf{10.1 \pm 7.0}$
Shoulder	Att-UNet	Indiv	Base	82.6 ± 8.9	82.7 ± 10.9	$\underline{99.8 \pm 0.2}$	59.9 ± 31.1	4.6 ± 3.9	12.3 ± 12.3
			ShapeReg	$\underline{84.5 \pm 7.3}$	81.4 ± 11.2	$\mathbf{99.9 \pm 0.1}$	38.1 ± 27.9	2.3 ± 1.7	11.1 ± 11.6
		Glob	Base	82.6 ± 9.1	80.3 ± 8.7	$\underline{99.8 \pm 0.3}$	30.2 ± 17.1	2.0 ± 1.5	$\underline{11.0 \pm 7.5}$
			ShapeReg	$\underline{84.5 \pm 9.1}$	$\underline{83.5 \pm 13.7}$	$\underline{99.8 \pm 0.2}$	$\mathbf{21.6 \pm 9.9}$	$\underline{1.5 \pm 1.3}$	14.9 ± 12.1
		Multi	Base	84.0 ± 12.3	82.8 ± 16.5	$\underline{99.8 \pm 0.2}$	24.7 ± 16.6	2.0 ± 3.3	14.8 ± 17.4
			ShapeReg	$\mathbf{86.9 \pm 5.9}$	$\mathbf{84.8 \pm 9.1}$	$\mathbf{99.9 \pm 0.1}$	$\underline{21.7 \pm 10.5}$	$\mathbf{1.2 \pm 0.9}$	$\mathbf{8.8 \pm 9.2}$

Table 4.2 – Leave-one-out quantitative assessment of Att-UNet [42] on ankle and shoulder datasets. Regularization methods include: baseline and shape priors [53]; while bone segmentation strategies comprise: individual, global and multi. Metrics encompass Dice (%), sensitivity (%), specificity (%), MSSD (mm), ASSD (mm) and RAVD (%). First and second best results for each dataset and for each metric are in bold and underlined italic respectively.

(i.e., global bone class and background). In addition, we performed visual comparison of the regularization methods (baseline and shape priors only) by employing a multi-class Att-UNet.

To evaluate the generalization abilities of each method, experiments were performed in a leave-one-out fashion such that one examination was retained for validation, one for test and the remaining data were used to train the model. The procedure was repeated over all the samples in the dataset to compute the mean and standard deviation for each metric. The hyper-parameters values (e.g., λ_1, batch size, learning rate) were selected based on the performance of the model on the validation set. Moreover, an expert (15 years of experience) visually validated the global anatomical consistency and plausibility of each predicted segmentation.

4.4 Results and discussion

4.4.1 Quantitative and qualitative assessment

The $\text{ShapeReg}^{\text{Multi}}_{\text{Att-UNet}}$ segmentation method based on a multi-class model with shape priors-based regularization achieved the best results on all metrics, except for sensitivity on ankle dataset and MSSD on shoulder dataset. For the ankle dataset, the method improved Dice (+1.5%), MSSD (−5.9 mm), ASSD (−0.3 mm) and RAVD (−1.4%) metrics

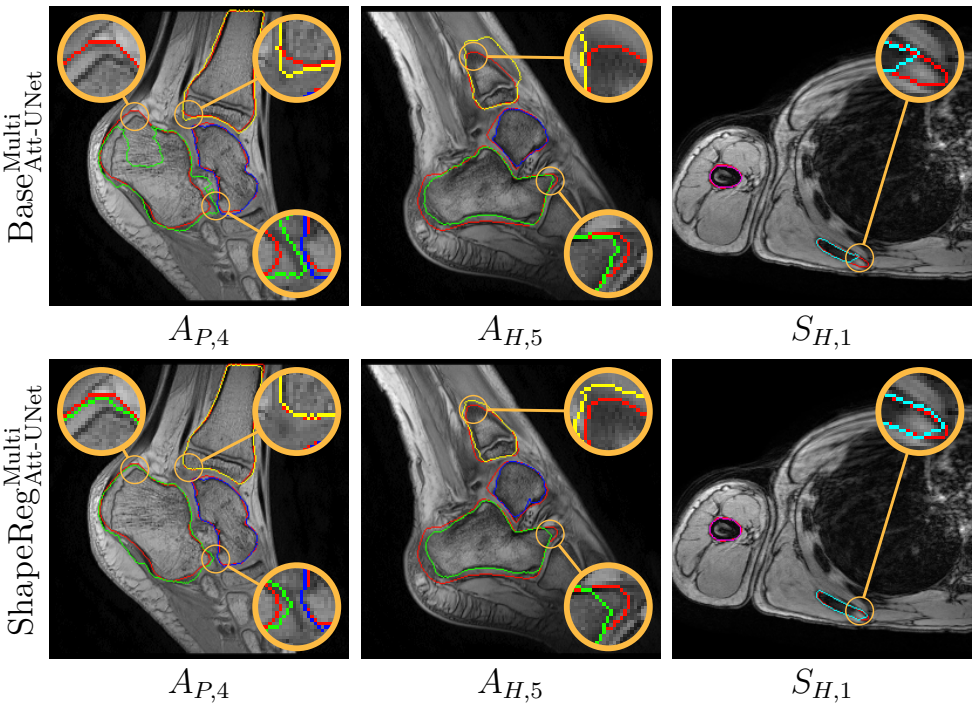


Figure 4.3 – **Visual comparison of regularizations methods using Att-UNet with multi-structure strategy.** Automatic segmentation of ankle and shoulder bones based on Att-UNet [42] with multi-structure strategy using baseline and shape priors [53] regularization. Ground truth delineations are in red (–) while predicted bones comprising calcaneus, talus, tibia, humerus and scapula appear in green (–), blue (–), yellow (–), magenta (–) and cyan (–) respectively.

while remaining 3.5% lower than the best in sensitivity metric (Table 4.2). For shoulder examinations, the method outperformed other approaches in Dice (+2.4%), sensitivity (+1.3%), MSSD (−0.5 mm), ASSD (−0.3 mm) and RAVD (−2.2%) while being 0.1 mm higher than the best in MSSD metric. The specificity metric was excellent in all methods ($>$ 99.3%). Most importantly, results obtained using the Att-UNet architecture demonstrated that shape priors-based regularization improved the performance for each bone segmentation strategy including individual, global, and multi. Finally, we observed that for a fixed regularization scheme, the multi-class strategy outperformed both the global and individual-class strategies except for ankle MSSD and RAVD, as well as shoulder sensitivity. Hence, **the proposed multi-class approach leveraged the benefits of simultaneously learning specific and shared bone features to enhance segmentation performance**.

The visual comparison of the baseline and shape priors regularization approaches

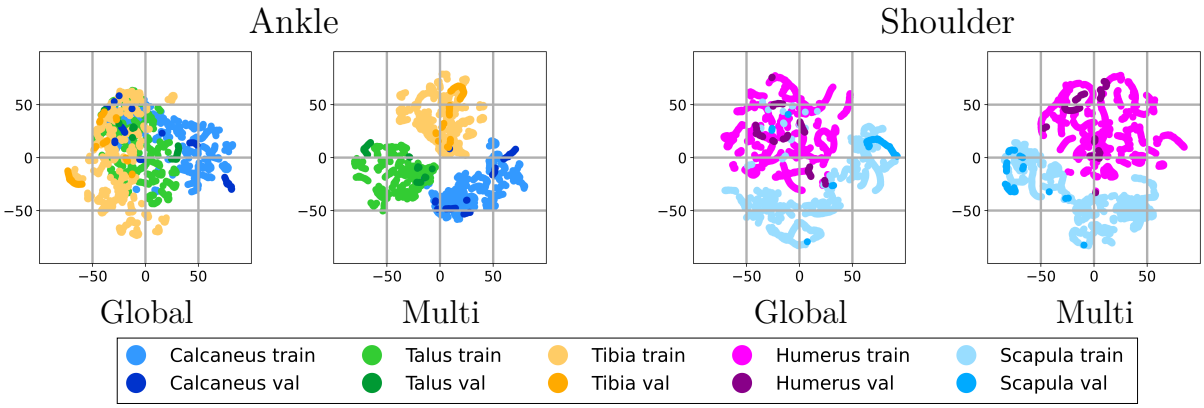


Figure 4.4 – **Visualization of the latent shape spaces learned by the global-class and multi-class auto-encoders on ankle and shoulder datasets.** The visualization was obtained using the t-SNE algorithm [87] in which each colored dot corresponds to a 2D binary mask of one of the anatomical objects of interest. The projection included 2D masks originating from the training (train) and validation (val) sets for each joint.

using multi-class Att-UNet provided the qualitative evidence of gradual improvements in segmentation performance (Figure 4.3). The $\text{Base}^{\text{Multi}}_{\text{Att-UNet}}$ was notably prone to both over-segmentation ($A_{H,5}$ tibia) and under-segmentation ($A_{P,4}$ calcaneus and $S_{H,1}$ scapula) errors. For instance, the intensity difference between ossified and non-ossified areas in the scapula bone resulted in erroneous delineations from baseline Att-UNet ($S_{H,1}$), while the shape regularization ($\text{ShapeReg}^{\text{Multi}}_{\text{Att-UNet}}$) enforced the model to follow the learned shape representation resulting in complete scapula segmentation. As discussed in Chapter 2, it is crucial in the clinical workflow to extract both ossified and non-ossified areas during the automatic segmentation process. In general the shape regularization promoted smoother bone delineations ($A_{P,4}$), nevertheless some bone contours remained difficult to extract ($A_{H,5}$ calcaneus).

4.4.2 Latent shape space analysis

While the work of Biffi et al. [235] demonstrated that a deep auto-encoder could learn to differentiate pathological from healthy cardiac shapes, our study focused on comparing shape representations arising from two different bone segmentation strategies. The pattern recognition behavior of the deep learning networks can be analyzed by visualizing the compact space learned during training. We analyzed the latent representation learned by the global-class and multi-class auto-encoders using the t-SNE dimensionality reduction algorithm [87]. The t-SNE algorithm is a non-linear method for visualizing high-

dimensional data, and it involves an optimization step to construct a 2D visualization. In the resulting visualization, each high-dimensional vector is modeled as a 2D point, with similar vectors represented by nearby points and distant points corresponding to dissimilar vectors. In practice, the L_2 Euclidean norm is employed to assess the similarity between feature vectors in high dimensions [87].

We used the auto-encoders trained on ground truth annotations and employed their encoder components to create latent codes of the bones of both training and validation subjects. We then applied global max pooling and obtained 512 dimensional codes from 2D bone masks. Finally, in order to visualize the 512 dimensional feature vectors, we applied a two-step dimensionality reduction as recommended in [87]. We first employed principal component analysis, which reduced the representations to 50 dimensional feature vectors, then the t-SNE algorithm embedded the data into a 2D space (Figure 4.4). It should be emphasized that the obtained visualizations depend on the selected t-SNE algorithm hyper-parameters [87]. During experiments, the perplexity (i.e., number of nearest neighbor points used for computation during optimization) and learning rate (i.e., gradient descent learning rate) of the t-SNE algorithm were set to 30 and 200, respectively.

For both ankle and shoulder examinations, **the latent representation learned by the global-class auto-encoder did not differentiate shape structures, contrary to the shape representation obtained by the multi-class auto-encoder**, which presented different clusters for each bone (Figure 4.4). Thus, ankle bones were aggregated into a unique cluster in the global-class representation, as opposed to the multi-class one which presented distinctive calcaneus, talus, and tibia clusters. The obtained visualizations reinforced our assumption that the global-class auto-encoder imposed the extraction of shared bone features. In contrast, the multi-class auto-encoder learned to extract discriminative bone features while complying with inter-bone relationships.

4.4.3 Limited interpretability

Although incorporating regularization through the loss function successfully constrains the network's parameters and promotes the desired characteristics for robust bone extraction, it fails to provide a better understanding of the inference process. Additionally, as the computation of the proposed regularization losses is based on a deep learning model (an auto-encoder), the interpretability of the regularization remains also limited. Hence, it would be beneficial to develop more interpretable models (segmentation network and shape encoder) in order to better analyze the internal behavior of the pipeline. More

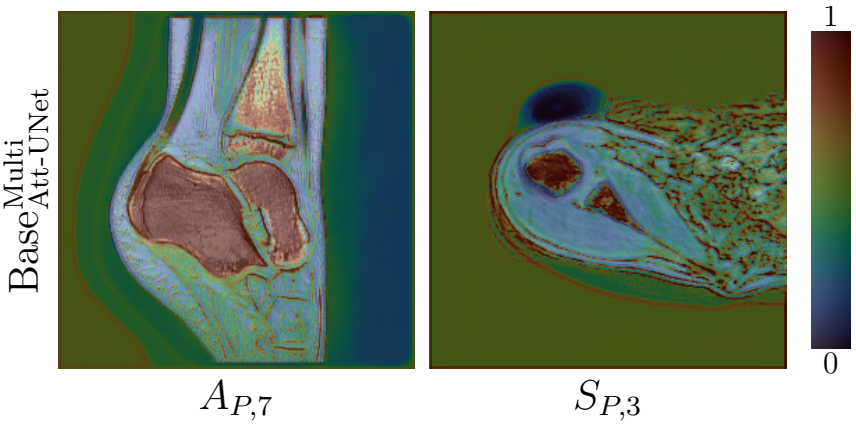


Figure 4.5 – **Visualization of the attention maps computed by the multi-class Att-UNet employed on ankle and shoulder joint images.** Pixel-wise coefficients ranging from 0 in blue to 1 in red indicated low to high attention.

precisely, such interpretable an segmentation model is crucial in medical image analysis applications, as it would allow a better analysis of the network failures (see Section 1.3.4).

Attention maps computed by the attention gates successfully provide a coarse localization of the anatomical structures of interest (Figure 4.5). This confirmed that the segmentation models leveraged the contextual information from the encoder branch to focus on the targeted ankle and shoulder bones. Indeed, **attention maps clearly suppressed irrelevant regions while highlighting calcaneus, talus, and tibia ankle bones as well as humerus and scapula shoulder bones**. Furthermore, one can note that the tibial bone was not uniformly highlighted which may result in less efficient shape extraction. However, these visualizations fail to explain the representation learned by the segmentation models. In this direction, the visualization of the learned feature maps represents the first step toward understanding the internal behavior of the "black box" type CNN models. An example is the work of Kamnitsas et al. [236], which shows that CNN learns concepts similar to the ones used by clinical experts. However, the learned convolutional layer can be activated by a mixture of patterns. Hence Zhang et al. [237] have devised an interpretable CNN in which each filter explicitly memorizes a specific object part without ambiguity and provides a clear semantic representation which could be of great interest for computed-aided musculoskeletal system analysis. Finally, one could mention the work of Biffi et al. [235] on explainable anatomical shape analysis which employed a variational auto-encoder with a two-dimensional latent space, enabling the direct visualization of the classification space and the discrimination of distinct clinical conditions. Thus, it may be thus beneficial to replace our convolutional auto-encoder with

such a model to reinforce the interpretability of the shape priors regularization.

4.5 Conclusion

In this chapter, we proposed and validated an automatic multi-bone segmentation framework that achieved promising performance on scarce and heterogeneous pediatric imaging datasets. The generalization capabilities of the segmentation model were enhanced by exploiting shape priors-based regularization, which enforced globally consistent shape predictions. Furthermore, the proposed method exploited specific as well as shared bone features arising from multi-class annotations in order to improve segmentation performance.

Nevertheless, even though the shape regularization enforces global anatomical consistency in model predictions, it fails to assess the global accuracy of generated masks given intensity images. Indeed, this regularization only exploits mask-based information and does not allow to evaluate the accuracy of the segmentation with respect to the input intensity image. However, pediatric pathological imaging examinations also exhibit irregular and complex pathological structures which are difficult to delineate due to alterations in shape and appearance [8], [9] (see Chapter 2). We will see how to mitigate this issue in Chapter 5. We propose to incorporate into the pipeline a conditional discriminator to reinforce the global realistic aspect of predicted delineations and employ transfer learning to reinforce the model generalizability.

Chapter 5

Leveraging adversarial networks and transfer learning for improved generalizability

5.1 Introduction

In this chapter, we continue to explore and design regularization approaches to limit over-fitting in the context of the management of sparse pediatric imaging datasets. In particular, we focus on two regularization techniques: through a penalty on the loss function (as seen in Chapter 4) and by means of novel neural network architecture designs. We first begin with modifications to the architecture of segmentation networks allowing for better generalization performance. Indeed, recent UNet extensions have been proposed based on more complex architectures incorporating dense, Inception, residual, or, more recently, Transformers-based [238]–[240] modules to provide more efficient optimization and enhanced performance. Additionally, networks integrating encoders (e.g., `VGG19` [46], `ResNet34` [241]) pre-trained on ImageNet [242] leverage low-level features typically shared between different image types to obtain more robust feature extraction. More specifically, transfer learning and fine-tuning from large non-medical datasets has become a widespread method in medical image analysis and has revealed improved performance compared to models with randomly initialized weights [40], [44], [46], [241]. However, the results of transfer learning depend on the task and dataset characteristics, with larger impact in very small data regimes [44]. **Hence, employing pre-trained models appears essential to address the data scarcity issue encountered in pediatric imaging.**

As discussed and illustrated in Chapter 4, regularization penalties incorporated in the optimization scheme allow us to guide the deep models and to promote more consistent delineation predictions in the context of medical image segmentation. In particular, the proposed shape regularization approach enforced the model to follow the learned non-linear

shape representation and thus limited under- and over-segmentation issues (see Figure 4.3). Nevertheless, images of the pediatric and pathological population may also contain irregular and complex pathological structures which are difficult to delineate due to alterations in both shape and appearance [8], [9]. **In this context, shape priors-based regularization may be insufficient and to tackle this limitation, we propose to employ an adversarial regularization based on a conditional discriminator to reinforce the accuracy of predicted delineations.** Specifically, inspired by image-to-image translation approaches [243], medical imaging researchers have employed adversarial networks to refine segmentation outputs. In these frameworks, a segmentation network and a discriminator are concurrently trained in a two-player game fashion in which the former learns to produce valid segmentation while the latter learns to discriminate between synthetic and real data [46], [56]–[58]. The adversarial term computed by the discriminator is added during the segmentation network optimization, which in turn, encourages UNet to fool the discriminator, and produces more plausible segmentation masks given input intensity images.

5.1.1 Contributions

In this chapter, we propose a multi-structure bone segmentation framework based on a partially pre-trained deep learning architecture combining shape priors with adversarial regularization (Figure 5.1). Unlike previous methods [46], [52]–[58], our framework simultaneously leverages both regularizations to guide the segmentation network to make anatomically consistent predictions and produce precise delineations. Specifically, our framework exploits a combination of shape priors and an adversarial regularizer to reduce the data scarcity issue while improving model generalizability. Furthermore, we demonstrate the usefulness of employing pre-trained models along with combining different regularization schemes for deep learning-based medical image segmentation. Finally, we provide an in-depth evaluation of the proposed method's performance by extending the experiments performed in Chapter 4.

As previously mentioned, the research conducted in this part has been published in the Artificial Intelligence in Medicine journal [233] and substantially extends a preliminary work presented at the IEEE International Symposium on Biomedical Imaging (ISBI) [234].

5.2 Incorporating adversarial priors into multi-structure segmentation framework

In this section, we explain the proposed segmentation network built upon Res-UNet and additional regularization terms incorporated into the loss function. We first briefly recall the partially pre-trained Res-UNet architecture (Section 5.2.1). Then, we combine regularization from shape priors and a conditional adversarial network (Section 5.2.2).

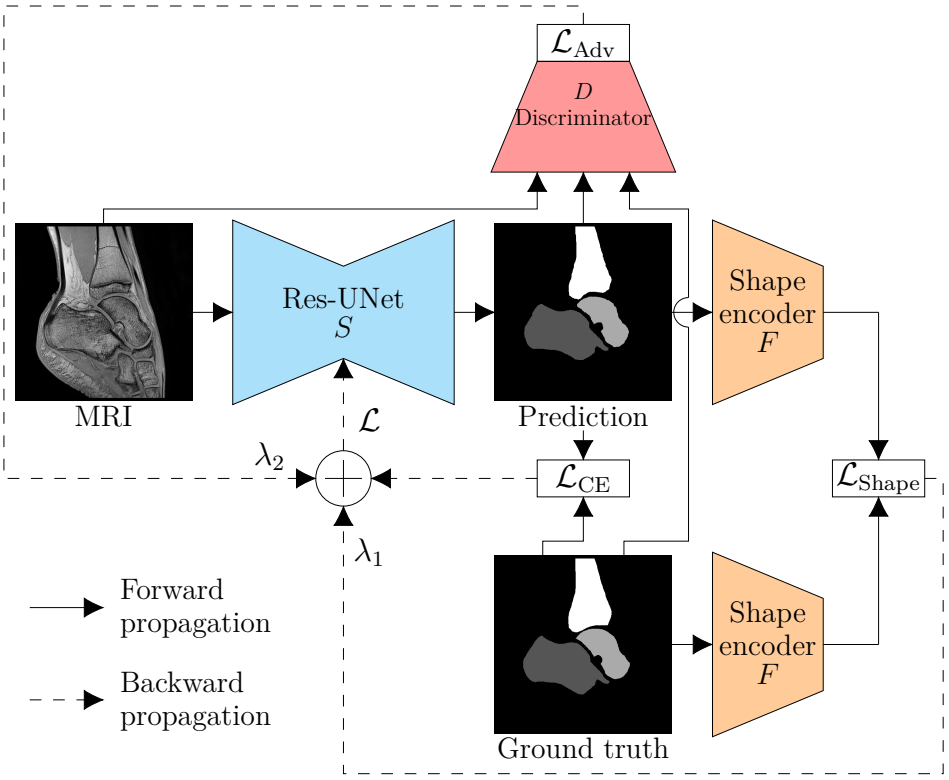


Figure 5.1 – Proposed regularized segmentation network S based on Res-UNet [59] exploiting cross-entropy loss $\mathcal{L}_{\text{CE}}$, shape priors-based $\mathcal{L}_{\text{Shape}}$ and adversarial $\mathcal{L}_{\text{Adv}}$ regularizations respectively computed by a shape encoder F with fixed weights and a discriminator D trained in competition with Res-UNet. The shape encoder corresponds to the encoder component of an auto-encoder previously optimized on ground truth segmentation masks, while the discriminator learns the plausibility of segmentation masks conditioned by their corresponding intensity image. λ_1 and λ_2 are two empirical weighting hyper-parameters.

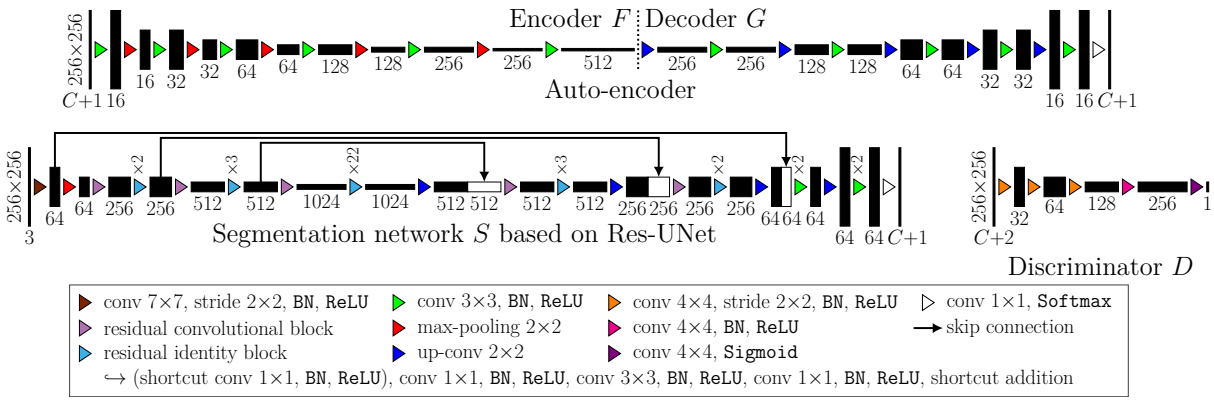


Figure 5.2 – Proposed multi-structure deep architectures with C structures of interest: auto-encoder comprising encoder F and decoder G (top), segmentation network S based on Res-UNet [59] (bottom left), and discriminator D (bottom right). The auto-encoder allows learning a non-linear shape representation from ground truth segmentations, while the discriminator outputs a one-channel likelihood map consisting of values ranging from 0 (fake) to 1 (real). During S training, the shape encoder F and discriminator D respectively compute the shape priors-based and adversarial regularizations to constrain the segmentation network (Figure 5.1). Finally, S integrates `ResNet50` as a pre-trained encoder and to fit the image dimensions, we extended input MR images from single grayscale channel to 3 channels.

5.2.1 Residual segmentation network with pre-trained encoder

We briefly recall the segmentation framework developed in Chapter 4, which incorporates a segmentation network S parameterized by Θ and a shape encoder F with Θ_F weights. The optimization procedure of S minimizes a loss based on cross-entropy and integrates a shape priors regularization computed by F with fixed weights, as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{Shape}} \tag{5.1}$$

Although we previously employed Att-UNet [42] as a backbone architecture (see Section 4.2.1) for the segmentation network S, it should be emphasized that our training strategy is architecture-independent. Hence, following works on transfer learning and fine-tuning from large datasets such as ImageNet, we modified the network design and replaced its encoder component with a classification network with weights previously trained on an image classification task. We assumed that leveraging a pre-trained encoder would lead to better generalization capabilities compared to models with randomly initialized weights [40]. Performance improvements have been particularly reported in low data regimes [44] similar to our scarce pediatric dataset setting.

The neural network S is now based on the UNet architecture [29], and we replaced its encoder branch with the `ResNet50` network, which incorporates residual blocks to allow faster convergence, increase network depth and improve predictive performance. First, to fit the `ResNet50` image dimensions, we concatenated 3 copies of each MR slice to extend the input from a single grayscale channel to 3 channels. The encoder branch built on residual convolutional and identity blocks [59] generated a 1024 dimensional feature map, which corresponds to the central part (i.e., bottleneck) between both contracting and expanding paths (Figure 5.2). We then constructed a symmetrical decoder branch with additional convolutional layers, features channels and residual blocks (Figure 5.2). Finally, it should be noted that contrary to encoder weights that are pre-trained on ImageNet, the decoder weights were randomly initialized.

With respect to the training procedure, while the shape regularization allows to improve the global shape consistency of the predicted segmentation, it is unable to assess the global accuracy of generated masks given intensity images. Hence, to address this problem, we propose to leverage an adversarial regularization reinforcing the realistic global aspect of predicted delineations.

5.2.2 Combining shape priors with adversarial regularization

For semantic segmentation, a conditional discriminator $D : y_i, x_i \mapsto D(y_i, x_i; \Theta_D)$ can assess whether a binary mask is fake or not, given the corresponding grayscale image, which is provided as a condition. The discriminator D is a neural network that returns a one-channel likelihood map $D(y_i, x_i; \Theta_D)$. Each value of the likelihood map (Figure 5.2) represents the degree of likelihood of correct segmentation of a crop of the input image, ranging from 0 (fake) to 1 (plausible or real). The likelihood is learned from the ground truth and generated data, and the discriminator architecture consisted of 4×4 convolutional layers to obtain a large receptive field.

Specifically, the architecture of D consists of five encoding layers with convolutional filters with a kernel of 4×4, stride 2×2 at the first three layers and stride 1×1 at 4^{th} and 5^{th} layers (see Figure 5.2). Batch normalization (`BN`) is applied after 2^{nd}, 3^{rd} convolutional filter and 4^{th} and `ReLU` is applied after each layer except the last. The `Sigmoid` activation function is used after the last convolutional filter. The network input is the concatenation of the 2D MR slice and the associated binary mask to be evaluated (ground truth or predicted). The output segmentation is an array of 32×32 values, each one from 0 (completely fake) to 1 (perfectly plausible or true).

Although traditional GAN approaches aim at generating new images with the same characteristics (i.e., statistics) as the training set, in a segmentation context, the discriminator instead enables to constrain the segmentation network through an adversarial regularization. Specifically, instead of generating new images, the segmentation model predicts synthetic (i.e., fake) masks from intensity images, which should be indistinguishable from ground truth (i.e., real) segmentation. Following typical adversarial learning schemes, the discriminator and the segmentation networks are trained alternatively and competitively, with the role of S being similar to that of the generator. More precisely, the optimization of weights Θ (respectively Θ_D) is done using the loss function $\mathcal{L}$ (respectively $\mathcal{L}_D$) while parameters Θ_D (respectively Θ) are fixed. These losses are defined in such a way that **the discriminator learns to differentiate real from synthetic segmentation masks while the segmentation network learns to generate increasingly plausible masks**.

The binary cross entropy loss $\mathcal{L}_{\text{BCE}}$ is typically used to train the discriminator, with real and fake labels for the likelihood maps of ground truth and generated masks respectively. $\mathcal{L}_{\text{BCE}}$ maximizes the loss value associated with the likelihood map of ground truth masks and minimizes the loss corresponding to the likelihood map of predicted masks, given the intensity image. Therefore, the discriminator learns to discriminate ground truth (i.e., real) from generated (i.e., fake) segmentations during the optimization of Θ_D [46], [56]–[58]. More precisely, the discriminator is optimized to yield a likelihood map with values equal to 1 (respectively 0) for ground truth (respectively predicted) masks.

$$\mathcal{L}_D := \mathcal{L}_{\text{BCE}} = \frac{1}{n}\sum_{i=1}^{n} -\log(1 - D(\hat{y}_i, x_i; \Theta_D)) - \log(D(y_i, x_i; \Theta_D)) \tag{5.2}$$

The discriminator was integrated into our segmentation framework by computing an adversarial term $\mathcal{L}_{\text{Adv}}$ derived from the probability that the network considered the generated mask to be the ground truth segmentation for a given grayscale image (Figure 5.1) [46], [56]–[58]. The loss computed from the discriminator likelihood map given $\hat{y}_i$, x_i and with fixed weights Θ_D was defined as follows:

$$\mathcal{L}_{\text{Adv}} = \frac{1}{n}\sum_{i=1}^{n} -\log(D(\hat{y}_i, x_i; \Theta_D)) \tag{5.3}$$

We modified the segmentation training strategy to combine shape priors-based and conditional adversarial regularizations. The optimization of the adversarial term encour-

aged the segmentation network to fool the discriminator (i.e., discriminator predicting a likelihood map equals to one for a synthetic mask), resulting in a more plausible segmentation mask with respect to the conditional intensity image. At first, the segmentation network will provide a rough prediction of the mask shape, and as the training process progresses, the discriminator will foster an increasingly accurate mask outline, resulting in more precise delineations of the targeted structures [46], [56]–[58]. The proposed loss function was a linear combination of cross-entropy, shape priors, and adversarial regularizations. The novel optimization procedure was defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{Shape}} + \lambda_2 \mathcal{L}_{\text{Adv}} \tag{5.4}$$

where λ_1 and λ_2 were empirically determined.

5.3 Experiments

Experiments were performed on ankle and shoulder pediatric MR datasets (see Section 2.4.2) following the same setting as described in Chapter 4. Furthermore, we extend the ablation study conducted using Att-UNet backbone architecture to include the adversarial regularization. Specifically, for each bone segmentation strategies (i.e., individual, global, and multi), we compared the baseline Att-UNet [42], Att-UNet with shape priors-based regularization [53], Att-UNet with adversarial regularization [57] and Att-UNet with proposed combined regularization scheme. Both hyper-parameters λ_1 and λ_2 were fixed to 0 to train baseline Att-UNet. We set λ_1 (respectively λ_2) to 0 to train Att-UNet with adversarial (respectively shape priors-based) regularization.

5.3.1 Pre-trained architectures performance

In addition to the previous experiments performed on the Att-UNet architecture, we also evaluated several other backbone architectures with and without the proposed combined regularization. In particular, we assessed the performance of our method based on a pre-trained Res-UNet with a multi-class strategy and proposed combined regularization ($\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$) against other backbone architectures pre-trained on a large natural image database. Specifically, we employed two backbone architectures (VGG-UNet [60] and Dense-UNet [61]) and compared pre-trained models with and without combined regularization. Moreover, as the work of Raghu et al. provided an in-depth study of the

benefits of employing transferred weights compared to randomly initialized ones for medical image analysis, especially in low data regimes [44], we omitted such evaluation in our experiments.

The VGG-UNet (respectively Dense-UNet) architecture referred to a UNet model whose encoder was replaced by a `VGG19` [60] (respectively `DenseNet121` [61]) classifier network pre-trained on ImageNet. Similarly to Res-UNet, the decoder components of VGG-UNet and Dense-UNet were extended by adding convolutional filters and more features (as well as dense blocks [61] for Dense-UNet) to get symmetrical networks. Finally, we employed only the multi-class strategy with combined regularization in these transfer learning experiments, as this scheme reached the best performance in the previous comparisons (see Section 5.4.2).

5.3.2 Implementation details

Network	Pre-trained Encoder	Batch Size	#Epochs	Learning Rate	#Param.
Auto-encoder	–	32	10	1e−2	3.1M
Discriminator	–	32	10/20	1e−3	0.7M
Att-UNet	–	32	20	1e−3	7.9M
VGG-UNet	`VGG19`	32	10	1e−4	34.7M
Dense-UNet	`DenseNet121`	16	10	1e−4	18.5M
Res-UNet	`ResNet50`	16	10	1e−4	13.6M

Table 5.1 – Summary of the networks employed during experiments: auto-encoder, discriminator, Att-UNet [42], VGG-UNet [60], Dense-UNet [61], and Res-UNet [59]; along with their corresponding number of trainable parameters and training hyper-parameter values (batch size, number of epochs and learning rate).

As discussed in Chapter 4, our training method consisted of two steps. The auto-encoder was first trained using the cross-entropy loss with batch size set 32 and 10 epochs. As a second step, the segmentation network and the discriminator were trained alternatively, one optimization step for both networks at each batch. We used Adam optimizer with initial learning rate set to 1e-3 for Att-UNet and to 1e-4 for VGG-UNet, Dense-UNet and Res-UNet. The batch size and number of epochs were set to 32 and 20 for Att-UNet, 32 and 10 for VGG-UNet, 16 and 10 for Dense-UNet and Res-UNet (Table 5.1). Additionally, each architecture was characterized by distinct model complexity (i.e., number of trainable parameters): auto-encoder (3.1 million), discriminator (0.7 million), Att-UNet (7.9

Metric	Best	Worst	Threshold
Dice (%)	100	0	> 80
Sensitivity (%)	100	0	> 80
MSSD (mm)	0	δ	< 30
ASSD (mm)	0	δ	< 4
RAVD (%)	0	100	< 10

Table 5.2 – Metrics wise threshold values employed in the ranking system. Metrics included Dice, sensitivity, MSSD, ASSD and RAVD. δ is the longest possible distance in 3D examinations.

million), VGG-UNet (34.7 million), Dense-UNet (18.5 million), and Res-UNet (13.6 million). Finally, since the individual scheme involved individual-class networks, this scheme thus involved C times more networks and parameters than the global- and multi-class strategies (with $C = 3$ in ankle and $C = 2$ in shoulder datasets).

We explored different regularization weighting parameters values and observed $\lambda_1 =$ 1e-1 and $\lambda_2 =$ 1e-2 to be the best combination across all backbone models. As mentioned in Chapter 4, all architectures were trained on 2D slices with extensive on-the-fly data augmentation due to limited available training data, and we employed the same post-processing based on largest connected set selection and morphological closing.

5.3.3 Ranking system

As mentioned in Section 3.5.2, **although it is essential to employ complementary metrics to assess the performance of each segmentation model, simultaneously comparing the performance of each segmentation strategy across multiple metrics can be challenging. Hence we propose to employ a metric-based ranking system.** More specifically, we converted the metrics outputs to normalized scores and used the average scores from all the datasets as a ranking system [91]. The proposed ranking system was created based on Dice, sensitivity, MSSD, ASSD, and RAVD (3D metrics defined in Chapter 3). Specificity was disregarded since excellent results were obtained for all methods ($> 99.3\%$ in Tables 5.3 and 5.4). Furthermore, a threshold was defined for each metric based on expert knowledge to remove non-satisfactory results. Then, we mapped the metric value between the corresponding best value and the threshold (Table 5.2) to the normalized interval $[0, 100]$. Metric values outside this acceptable range were assigned zero scores. The score of the predicted 3D segmentation corresponded to the average over all metric scores, and methods were ranked according to their obtained scalar

score. Separate rankings were performed for each dataset (i.e., shoulder and ankle).

Ranking results via multiple metrics is an arduous task as the selection of thresholds may have an impact on the final ranking [219]. Hence, to assess the robustness of the ranking system, we analyzed the effect of the modification of the threshold values (each resulting in a different ranking system). We tested different threshold values for each metric: Dice ($75-85\%$), sensitivity ($75-85\%$), MSSD ($20-40$ mm), ASSD ($3-5$ mm) and RAVD ($5-15\%$). Thresholds were modified independently. Metric values between the corresponding best value and the modified threshold were mapped to the interval $[0, 100]$.

5.3.4 Quantitative and qualitative assessment of predicted segmentation

Assessment of the predicted segmentation relied on the 3D metrics introduced in Chapter 3 (i.e., Dice, sensitivity, specificity, MSSD, ASSD, and RAVD) and experiments followed the same leave-one-out evaluation design described in Chapter 4.

Furthermore, due to the scarce amount of 3D examinations, we performed the statistical analysis between methods on 2D MR images. We employed the Wilcoxon signed-rank non-parametric test [244] using Dice, sensitivity and specificity scores obtained from the 1446 ankle (respectively 3357 shoulder) 2D slices containing at least one bone of interest and which corresponded to the 17 ankle (respectively 15 shoulder) 3D MR images. The statistical tests were conducted using only the 2D slices containing at least one bone of interest. We preliminary verified the non-normality of the 2D results distributions using the D'Agostino and Pearson normality test [245], [246]. We then performed the statistical analysis between methods and compared the obtained p-values to the typical 0.05 threshold. Due to the skew of the non-normal distributions of 2D scores, we reported as in [228] their mean and the distances from the mean to the upper and lower bound of the 68% confidence interval, which corresponds to the 16 and 84 percentiles. However, we did not perform an preliminary non-parametric analysis of variance such the Kruskal-Wallis test by rank [247]. This represents a limitation of our work.

Finally, we performed visual comparison of predicted segmentation masks at three levels. First, we compared the results of the bone segmentation strategies (individual, global, and multi) using Att-UNet with combined regularization. Second, we extend the visualization of the regularization methods conducted with multi-class Att-UNet in Chapter 4 by including the adversarial and proposed combined methods. Third, we compared the

pre-trained backbone architectures, including VGG-UNet, Dense-UNet, and Res-UNet, employed with multi-class segmentation strategy and combined regularization.

5.4 Results

The proposed $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ method based on pre-trained multi-class Res-UNet with combined regularization was evaluated on two pediatric datasets. In this section we report quantitative results (Section 5.4.1), ranking scores (Section 5.4.2), and qualitative comparisons (Section 5.4.4) for each dataset.

5.4.1 Quantitative assessment

	Method			Dice ↑	Sens. ↑	Spec. ↑	MSSD ↓	ASSD ↓	RAVD ↓
Ankle Dataset	Att-UNet	Indiv	Base	84.0 ± 6.4	82.7 ± 9.5	99.3 ± 0.8	20.4 ± 12.2	2.0 ± 1.4	17.9 ± 16.0
			ShapeReg	87.1 ± 3.9	85.8 ± 7.3	99.5 ± 0.4	18.4 ± 11.2	1.5 ± 0.6	11.5 ± 7.5
			AdvReg	86.4 ± 3.8	83.8 ± 8.4	99.5 ± 0.4	16.5 ± 9.4	1.5 ± 0.7	15.1 ± 8.8
			CombReg	88.0 ± 5.4	86.9 ± 8.0	99.5 ± 0.3	18.0 ± 13.3	1.4 ± 0.7	9.1 ± 5.8
		Glob	Base	87.6 ± 7.8	$\underline{92.6 \pm 5.4}$	99.0 ± 1.2	19.9 ± 14.7	1.8 ± 1.6	17.9 ± 27.5
			ShapeReg	87.8 ± 6.2	90.6 ± 7.7	99.1 ± 1.0	18.0 ± 12.3	1.7 ± 1.4	13.7 ± 14.3
			AdvReg	87.8 ± 5.0	$\mathbf{93.0 \pm 3.8}$	99.1 ± 0.7	18.2 ± 11.3	1.7 ± 1.2	14.4 ± 11.4
			CombReg	89.8 ± 2.4	90.9 ± 4.9	99.4 ± 0.3	18.3 ± 12.4	1.3 ± 0.7	$\underline{7.3 \pm 4.8}$
		Multi	Base	88.4 ± 6.2	86.6 ± 10.1	$\underline{99.6 \pm 0.4}$	17.0 ± 12.4	1.3 ± 0.9	12.5 ± 10.8
			ShapeReg	$\underline{89.9 \pm 6.2}$	89.1 ± 9.1	$\underline{99.6 \pm 0.3}$	$\mathbf{11.1 \pm 4.3}$	$\underline{1.0 \pm 0.6}$	10.1 ± 7.0
			AdvReg	$\underline{89.9 \pm 3.5}$	88.9 ± 6.8	$\underline{99.6 \pm 0.3}$	$\underline{13.6 \pm 7.5}$	1.1 ± 0.5	9.5 ± 5.3
			CombReg	$\mathbf{90.7 \pm 3.2}$	88.8 ± 6.3	$\mathbf{99.7 \pm 0.2}$	$\mathbf{11.1 \pm 3.4}$	$\mathbf{0.9 \pm 0.3}$	$\mathbf{7.1 \pm 5.7}$
Shoulder Dataset	Att-UNet	Indiv	Base	82.6 ± 8.9	82.7 ± 10.9	$\underline{99.8 \pm 0.2}$	59.9 ± 31.1	4.6 ± 3.9	12.3 ± 12.3
			ShapeReg	84.5 ± 7.3	81.4 ± 11.2	$\mathbf{99.9 \pm 0.1}$	38.1 ± 27.9	2.3 ± 1.7	11.1 ± 11.6
			AdvReg	84.3 ± 6.4	82.0 ± 10.0	$\underline{99.8 \pm 0.1}$	28.6 ± 16.0	1.7 ± 1.0	10.8 ± 8.0
			CombReg	85.7 ± 5.6	83.4 ± 9.7	$\underline{99.8 \pm 0.1}$	30.5 ± 19.7	1.8 ± 1.3	9.4 ± 10.0
		Glob	Base	82.6 ± 9.1	80.3 ± 8.7	$\underline{99.8 \pm 0.3}$	30.2 ± 17.1	2.0 ± 1.5	11.0 ± 7.5
			ShapeReg	84.5 ± 9.1	83.5 ± 13.7	$\underline{99.8 \pm 0.2}$	21.6 ± 9.9	1.5 ± 1.3	14.9 ± 12.1
			AdvReg	84.3 ± 9.3	84.5 ± 13.0	$\underline{99.8 \pm 0.3}$	26.7 ± 11.3	1.7 ± 1.3	13.6 ± 15.0
			CombReg	86.1 ± 5.2	85.5 ± 7.1	$\underline{99.8 \pm 0.2}$	25.8 ± 8.9	$\underline{1.4 \pm 0.7}$	$\underline{8.2 \pm 11.1}$
		Multi	Base	84.0 ± 12.3	82.8 ± 16.5	$\underline{99.8 \pm 0.2}$	24.7 ± 16.6	2.0 ± 3.3	14.8 ± 17.4
			ShapeReg	$\underline{86.9 \pm 5.9}$	84.8 ± 9.1	$\mathbf{99.9 \pm 0.1}$	$\underline{21.7 \pm 10.5}$	$\mathbf{1.2 \pm 0.9}$	8.8 ± 9.2
			AdvReg	85.7 ± 7.1	$\underline{86.4 \pm 8.2}$	$\underline{99.8 \pm 0.3}$	23.7 ± 18.5	1.6 ± 1.5	10.4 ± 11.5
			CombReg	$\mathbf{87.8 \pm 5.2}$	$\mathbf{87.1 \pm 5.9}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{21.2 \pm 13.3}$	$\mathbf{1.2 \pm 1.1}$	$\mathbf{4.8 \pm 4.7}$

Table 5.3 – Leave-one-out quantitative assessment of Att-UNet [42] on ankle and shoulder datasets. Regularization methods include: baseline, shape priors [53], adversarial [57] and proposed combined; while bone segmentation strategies comprise: individual, global and multi. Metrics encompass Dice (%), sensitivity (%), specificity (%), MSSD (mm), ASSD (mm) and RAVD (%). First and second best results for each dataset and for each metric are in bold and underlined respectively.

	Method			Dice ↑	Sens. ↑	Spec. ↑	MSSD ↓	ASSD ↓	RAVD ↓
Ankle	VGG	Multi	Base	92.9 ± 1.4	$\mathbf{93.6 \pm 3.7}$	$\underline{99.7 \pm 0.2}$	9.1 ± 4.2	$\underline{0.7 \pm 0.1}$	7.3 ± 3.1
			CombReg	93.1 ± 1.6	$\mathbf{93.6 \pm 3.6}$	$\underline{99.7 \pm 0.2}$	8.1 ± 2.6	$\underline{0.7 \pm 0.1}$	5.9 ± 4.0
	Dense	Multi	Base	93.6 ± 2.0	91.7 ± 4.7	$\mathbf{99.8 \pm 0.1}$	7.9 ± 3.9	$\underline{0.7 \pm 0.2}$	6.8 ± 4.9
			CombReg	$\underline{93.9 \pm 1.8}$	92.2 ± 4.5	$\mathbf{99.8 \pm 0.1}$	$\underline{6.8 \pm 3.4}$	$\mathbf{0.6 \pm 0.2}$	6.1 ± 4.4
	Res	Multi	Base	93.8 ± 1.8	$\underline{92.4 \pm 4.3}$	$\mathbf{99.8 \pm 0.1}$	7.3 ± 3.2	$\mathbf{0.6 \pm 0.2}$	$\underline{5.3 \pm 4.4}$
			CombReg	$\mathbf{94.3 \pm 1.1}$	$\mathbf{93.6 \pm 3.1}$	$\mathbf{99.8 \pm 0.1}$	$\mathbf{6.1 \pm 2.8}$	$\mathbf{0.6 \pm 0.1}$	$\mathbf{4.7 \pm 2.9}$
Shoulder	VGG	Multi	Base	88.7 ± 4.5	$\underline{91.5 \pm 5.0}$	$\underline{99.8 \pm 0.2}$	24.6 ± 26.9	1.3 ± 1.6	8.9 ± 11.6
			CombReg	89.2 ± 3.7	$\mathbf{92.1 \pm 3.5}$	$\underline{99.8 \pm 0.1}$	$\underline{21.7 \pm 22.1}$	$\underline{1.0 \pm 0.8}$	6.8 ± 6.1
	Dense	Multi	Base	$\underline{90.5 \pm 3.2}$	91.1 ± 3.0	$\mathbf{99.9 \pm 0.1}$	29.8 ± 26.4	1.1 ± 0.9	4.5 ± 4.0
			CombReg	$\mathbf{90.7 \pm 3.0}$	90.2 ± 3.6	$\mathbf{99.9 \pm 0.1}$	21.8 ± 21.0	$\mathbf{0.8 \pm 0.6}$	$\underline{4.4 \pm 2.2}$
	Res	Multi	Base	90.1 ± 3.5	90.4 ± 3.1	$\mathbf{99.9 \pm 0.1}$	23.7 ± 22.6	$\underline{1.0 \pm 1.2}$	$\mathbf{3.5 \pm 3.7}$
			CombReg	$\mathbf{90.7 \pm 3.0}$	90.7 ± 3.6	$\mathbf{99.9 \pm 0.1}$	$\mathbf{19.3 \pm 14.2}$	$\mathbf{0.8 \pm 0.5}$	$\mathbf{3.5 \pm 3.4}$

Table 5.4 – Leave-one-out quantitative assessment of the three pre-trained architectures: VGG-UNet [60], Dense-UNet [61] and Res-UNet [59] on ankle and shoulder datasets. Baseline and combined regularization methods are employed along with the multi-structure strategy. Metrics encompass Dice (%), sensitivity (%), specificity (%), MSSD (mm), ASSD (mm) and RAVD (%). First and second best results for each dataset and for each metric are in bold and underlined respectively.

The results obtained using the Att-UNet architecture, which complete the experiments performed in Chapter 4, demonstrated that the segmentation method based on a multi-class model with proposed combined regularization achieved the best results on all metrics, except for sensitivity on ankle dataset (Table 5.3). For the ankle dataset, the method improved Dice (+0.8%), specificity (+0.1%), MSSD (−2.5 mm), ASSD (−0.1 mm) and RAVD (−0.2%) metrics while remaining 4.2% lower than the best in sensitivity metric. For shoulder examinations, the method outperformed other approaches in Dice (+0.9%), sensitivity (+0.7%), specificity (+0.1%), MSSD (−0.5 mm), ASSD (−0.2 mm) and RAVD (−3.4%).

Furthermore, the proposed $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ method outperformed state-of-the-art pre-trained methods on all metrics, except for sensitivity on shoulder dataset (Table 5.4). All methods reached excellent specificity scores ($> 99.7\%$). For ankle examinations, the proposed $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ method ranked best in Dice (94.3%), sensitivity (93.6%), MSSD (6.1 mm), ASSD (0.6 mm) and RAVD (5.1%) metrics. For the shoulder dataset, the proposed $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ method achieved the best results in Dice (90.7%), MSSD (19.3 mm), ASSD (0.8 mm) and RAVD (3.5%) metrics while remaining marginally lower in sensitivity (0.4% lower than the best). It is also worth mentioning that, while the performance improvements are lower than in Att-UNet experiments (Table 5.3), the proposed combined regularization consistently improved performance across all architectures and

metrics except for VGG-UNet and Dense-UNet shoulder sensitivity, and the proposed $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ method was associated with the lowest variance in all metrics except for ankle MSSD and shoulder sensitivity and RAVD.

5.4.2 Rankings

Method			Ankle Dataset		Shoulder Dataset	
			Mean ± STD	Rank	Mean ± STD	Rank
Att-UNet	Indiv	Base	33.3 ± 16.2	18	25.2 ± 18.8	18
		ShapeReg	40.4 ± 10.9	16	33.0 ± 23.9	15
		AdvReg	37.7 ± 13.2	17	31.9 ± 16.3	16
		CombReg	44.2 ± 16.3	15	36.8 ± 21.3	14
	Global	Base	46.2 ± 18.3	14	28.7 ± 20.4	17
		ShapeReg	47.4 ± 20.7	13	37.2 ± 22.6	12
		AdvReg	47.6 ± 17.2	12	37.2 ± 18.3	13
		CombReg	50.9 ± 12.7	9	40.9 ± 18.6	11
	Multi	Base	47.9 ± 21.9	11	41.6 ± 23.3	10
		ShapeReg	54.4 ± 18.5	8	42.4 ± 20.3	9
		AdvReg	49.3 ± 15.7	10	42.5 ± 23.4	8
		CombReg	56.7 ± 16.2	7	48.7 ± 21.3	7
VGG	Multi	Base	63.0 ± 7.8	6	53.7 ± 18.4	6
		CombReg	66.5 ± 9.2	4	54.6 ± 16.0	5
Dense	Multi	Base	65.2 ± 13.8	5	55.2 ± 12.0	4
		CombReg	67.7 ± 13.3	3	56.4 ± 10.6	3
Res	Multi	Base	68.5 ± 14.2	2	56.7 ± 12.9	2
		CombReg	$\mathbf{71.4 \pm 10.0}$	**1**	$\mathbf{59.2 \pm 13.9}$	**1**

Table 5.5 – Scores of the four backbone architectures: Att-UNet [42], VGG-UNet [60], Dense-UNet [61], and Res-UNet [59] on ankle and shoulder datasets. Regularization methods include: baseline, shape priors [53], adversarial [57] and proposed combined; while bone segmentation strategies comprise: individual, global and multi. Results encompass mean, standard deviation (STD) and associated rank. Methods were ranked according to their mean score. Best results are in bold.

$\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ ranked first in performance (Table 5.5) for both datasets with mean scores of 71.4 on ankle dataset and 59.2 on shoulder dataset. Baseline Res-UNet ranked second on both datasets, while individual-class baseline Att-UNet ranked last on ankle (mean score of 33.3) and shoulder (mean score of 25.2) datasets. It was observed in the experiments based on Att-UNet architecture that for a fixed regularization scheme, the multi-class strategy outperformed the global-class strategy, which in turn outranked

Method				Rankings									
				$Dice_{75}$	$Dice_{85}$	$Sens_{75}$	$Sens_{85}$	$MSSD_{20}$	$MSSD_{40}$	$ASSD_3$	$ASSD_5$	$RAVD_5$	$RAVD_{15}$
Ankle Dataset	Att-UNet	Indiv	Base	18	18	18	18	18	18	18	18	18	18
			ShapeReg	16	16	16	16	16	16	16	16	16	16
			AdvReg	17	17	17	17	17	17	17	17	17	17
			CombReg	15	15	15	15	15	15	15	15	15	15
		Global	Base	14	14	14	14	14	14	14	14	14	14
			ShapeReg	13	**12**	13	13	**12**	13	**12**	13	**12**	**12**
			AdvReg	12	**13**	12	12	**13**	**11**	**13**	12	**11**	**13**
			CombReg	9	9	9	9	9	9	9	9	9	9
		Multi	Base	11	11	11	11	11	**12**	11	11	**13**	11
			ShapeReg	8	8	8	8	8	8	8	8	8	8
			AdvReg	10	10	10	10	10	10	10	10	10	10
			CombReg	7	7	7	7	7	7	7	7	7	7
	VGG	Multi	Base	6	6	6	6	6	6	6	6	6	6
			CombReg	4	4	4	4	4	4	4	4	4	4
	Dense	Multi	Base	5	5	5	5	5	5	5	5	5	5
			CombReg	3	3	3	3	3	3	3	3	3	3
	Res	Multi	Base	2	2	2	2	2	2	2	2	2	2
			CombReg	1	1	1	1	1	1	1	1	1	1
Shoulder Dataset	Att-UNet	Indiv	Base	18	18	18	18	18	18	18	18	18	18
			ShapeReg	15	15	15	15	15	**16**	15	15	**16**	15
			AdvReg	16	16	16	16	16	**15**	16	16	**15**	16
			CombReg	**13**	14	**13**	**13**	**12**	14	14	**12**	14	14
		Global	Base	17	17	17	17	17	17	17	17	17	17
			ShapeReg	12	12	**14**	12	**14**	**13**	**13**	**14**	12	**13**
			AdvReg	**14**	13	**12**	**14**	13	**12**	**12**	13	13	**12**
			CombReg	**10**	11	11	11	11	11	11	11	**10**	11
		Multi	Base	**11**	10	10	10	10	10	10	10	**11**	10
			ShapeReg	**8**	9	**8**	9	9	9	9	9	**8**	**8**
			AdvReg	**9**	8	**9**	8	8	8	8	8	**9**	**9**
			CombReg	7	7	7	7	7	7	7	7	7	7
	VGG	Multi	Base	6	6	6	6	6	6	6	6	6	6
			CombReg	5	5	5	5	5	5	5	5	5	5
	Dense	Multi	Base	4	4	4	4	4	4	4	4	4	4
			CombReg	3	3	3	3	3	3	3	3	3	**2**
	Res	Multi	Base	2	2	2	2	2	2	2	2	2	**3**
			CombReg	1	1	1	1	1	1	1	1	1	1

Table 5.6 – Transformed rankings of the four backbone architectures: Att-UNet [42], VGG-UNet [60], Dense-UNet [61], and Res-UNet [59] on ankle and shoulder datasets. Regularization methods include: baseline, shape priors-based regularization [53], adversarial regularization [57] and the proposed combined regularization; and bone segmentation strategies comprise: individual, global and multi. Rankings were computed using different threshold values: Dice = 75 or 85%, Sensitivity = 75 or 85%, MSSD = 20 or 40 mm, ASSD = 3 or 5 mm and RAVD = 5 or 15%. Modified ranks are in bold.

the individual-class scheme. Furthermore, for a fixed bone segmentation strategy, shape priors-based and adversarial regularizations improved the baseline performance, while a combined regularization resulted in the best overall performance. Additionally, the ranks

achieved by the pre-trained architectures (VGG-UNet, Dense-UNet, and Res-UNet) further demonstrated that the proposed combined regularization promoted better performance as compared to baseline training. Hence, **the combined regularization consistently outperformed the compared methods covering various segmentation strategies (individual, global, and multi) and distinct architectures (Att-UNet, VGG-UNet, Dense-UNet, and Res-UNet), demonstrating the effectiveness of the proposed approach**. Finally, to assess the robustness of our ranking system and these observations, several threshold values were tested as reported in Table 5.6 with modified ranks in bold and our conclusions remained unchanged on every transformed ranking. For instance, Dice threshold modification to 85% (Dice_{85}) led to a permutation of $\text{ShapeReg}^{\text{Global}}_{\text{Att-UNet}}$ and $\text{AdvReg}^{\text{Global}}_{\text{Att-UNet}}$ ranks (12^{th} and 13^{th}) on ankle dataset. More importantly, $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ ranked first on both datasets, whatever the selected threshold values, which further confirms the efficiency of the proposed contributions.

5.4.3 Statistical analysis

The statistical analysis performed on 2D slices using Dice, sensitivity and specificity metrics (Table 5.7) indicated that the proposed $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ model produced significant improvements (p-values < 0.05), except compared with: $\text{Base}^{\text{Global}}_{\text{Att-UNet}}$ and $\text{AdvReg}^{\text{Global}}_{\text{Att-UNet}}$ on ankle datasets using sensitivity 2D metrics; as well as $\text{CombReg}^{\text{Multi}}_{\text{Dense-UNet}}$, $\text{Base}^{\text{Multi}}_{\text{Res-UNet}}$ and $\text{ShapeReg}^{\text{Multi}}_{\text{Att-UNet}}$ on shoulder datasets using 2D Dice, 2D sensitivity and 2D specificity metrics respectively. In these particular cases, the difference between results obtained by our model and compared methods was not statistically significant. However, in each case $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ produced statistically significant improvements on the remaining 2D metrics. Hence, we considered the overall improvements achieved by our model to be statistically significant.

It should be noted that in this thesis, we did not take into account the problem of multiple comparisons when reporting the p-values. This a limitation of our work, we could have employed a multiple testing correction such the Holm-Bonferroni method [248] to adjust the rejection criteria for each of the individual hypotheses. However, as most p-values were lower than 1×10^{-6}, we assumed that such corrections would not have impacted our final conclusions.

Furthermore, the results reported from 2D slices were consistent with the performance achieved on 3D examinations. For ankle datasets, our proposed model $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ ranked best in 2D Dice (86.2%) and 2D sensitivity (86.2%) metrics while remaining 0.1%

Method				Dice 2D	p-value	Sens. 2D	p-value	Spec. 2D	p-value
Ankle Dataset	Att-UNet	Indiv	Base	$70.9^{+22.1}_{-28.0}$	$<\mathbf{1\times10^{-6}}$	$72.3^{+22.1}_{-26.3}$	$<\mathbf{1\times10^{-6}}$	$98.4^{+1.5}_{-1.6}$	$<\mathbf{1\times10^{-6}}$
			ShapeReg	$74.3^{+19.6}_{-18.2}$	$<\mathbf{1\times10^{-6}}$	$75.1^{+21.2}_{-20.3}$	$<\mathbf{1\times10^{-6}}$	$98.7^{+1.2}_{-1.5}$	$<\mathbf{1\times10^{-6}}$
			AdvReg	$74.6^{+18.6}_{-20.6}$	$<\mathbf{1\times10^{-6}}$	$74.4^{+20.6}_{-21.5}$	$<\mathbf{1\times10^{-6}}$	$99.0^{+1.0}_{-1.1}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$76.6^{+17.9}_{-19.4}$	$<\mathbf{1\times10^{-6}}$	$77.1^{+18.8}_{-22.3}$	$<\mathbf{1\times10^{-6}}$	$99.0^{+0.9}_{-0.7}$	$<\mathbf{1\times10^{-6}}$
		Global	Base	$77.9^{+16.7}_{-16.0}$	$<\mathbf{1\times10^{-6}}$	$85.3^{+13.6}_{-8.2}$	8.7×10^{-2}	$97.7^{+2.0}_{-1.8}$	$<\mathbf{1\times10^{-6}}$
			ShapeReg	$77.3^{+17.1}_{-16.0}$	$<\mathbf{1\times10^{-6}}$	$82.1^{+16.2}_{-16.1}$	$<\mathbf{1\times10^{-6}}$	$98.1^{+1.7}_{-1.7}$	$<\mathbf{1\times10^{-6}}$
			AdvReg	$77.9^{+15.9}_{-16.1}$	$<\mathbf{1\times10^{-6}}$	$85.8^{+12.4}_{-8.0}$	6.4×10^{-1}	$97.9^{+1.7}_{-1.5}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$79.9^{+14.0}_{-10.3}$	$<\mathbf{1\times10^{-6}}$	$82.3^{+15.4}_{-7.9}$	$<\mathbf{1\times10^{-6}}$	$98.7^{+1.1}_{-1.2}$	$<\mathbf{1\times10^{-6}}$
		Multi	Base	$76.2^{+19.3}_{-23.0}$	$<\mathbf{1\times10^{-6}}$	$76.7^{+19.9}_{-22.8}$	$<\mathbf{1\times10^{-6}}$	$99.1^{+0.8}_{-0.7}$	$<\mathbf{1\times10^{-6}}$
			ShapeReg	$80.5^{+15.4}_{-11.2}$	$<\mathbf{1\times10^{-6}}$	$81.4^{+16.3}_{-17.2}$	$<\mathbf{1\times10^{-6}}$	$99.2^{+0.8}_{-0.7}$	$<\mathbf{1\times10^{-6}}$
			AdvReg	$79.0^{+16.3}_{-13.6}$	$<\mathbf{1\times10^{-6}}$	$79.5^{+18.3}_{-16.6}$	$<\mathbf{1\times10^{-6}}$	$99.0^{+0.9}_{-1.1}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$78.1^{+17.4}_{-15.9}$	$<\mathbf{1\times10^{-6}}$	$77.0^{+19.7}_{-19.7}$	$<\mathbf{1\times10^{-6}}$	$99.3^{+0.7}_{-0.6}$	$<\mathbf{1\times10^{-6}}$
	VGG	Multi	Base	$81.7^{+14.9}_{-10.7}$	$<\mathbf{1\times10^{-6}}$	$83.7^{+14.9}_{-11.8}$	$\mathbf{1.9\times10^{-3}}$	$99.2^{+0.7}_{-0.7}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$82.9^{+13.6}_{-9.7}$	$<\mathbf{1\times10^{-6}}$	$84.2^{+14.4}_{-12.3}$	$\mathbf{2.0\times10^{-4}}$	$99.2^{+0.7}_{-0.6}$	$<\mathbf{1\times10^{-6}}$
	Dense	Multi	Base	$84.8^{+12.1}_{-7.8}$	$<\mathbf{1\times10^{-6}}$	$83.7^{+13.8}_{-10.1}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.6^{+0.4}_{-0.3}}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$84.9^{+12.0}_{-5.4}$	$\mathbf{1.2\times10^{-4}}$	$84.2^{+13.4}_{-7.2}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.6^{+0.4}_{-0.4}}$	$<\mathbf{1\times10^{-6}}$
	Res	Multi	Base	$83.9^{+13.1}_{-6.8}$	$<\mathbf{1\times10^{-6}}$	$83.5^{+14.1}_{-8.2}$	$<\mathbf{1\times10^{-6}}$	$99.5^{+0.4}_{-0.3}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$\mathbf{86.2^{+10.8}_{-6.0}}$	–	$\mathbf{86.2^{+11.5}_{-6.5}}$	–	$99.5^{+0.4}_{-0.3}$	–
Shoulder Dataset	Att-UNet	Indiv	Base	$82.6^{+13.0}_{-12.2}$	$<\mathbf{1\times10^{-6}}$	$83.6^{+12.8}_{-9.3}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.2}$	$\mathbf{3.3\times10^{-2}}$
			ShapeReg	$82.6^{+13.4}_{-8.9}$	$<\mathbf{1\times10^{-6}}$	$81.4^{+15.2}_{-15.2}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
			AdvReg	$83.8^{+11.8}_{-7.7}$	$<\mathbf{1\times10^{-6}}$	$83.1^{+13.9}_{-11.23}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.1}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$83.6^{+11.9}_{-7.9}$	$<\mathbf{1\times10^{-6}}$	$83.2^{+13.4}_{-10.7}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.1}$	$\mathbf{1.4\times10^{-5}}$
		Global	Base	$82.8^{+13.1}_{-13.0}$	$<\mathbf{1\times10^{-6}}$	$82.7^{+14.3}_{-14.7}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.1}$	$\mathbf{3.5\times10^{-2}}$
			ShapeReg	$84.7^{+11.3}_{-6.2}$	$<\mathbf{1\times10^{-6}}$	$85.1^{+12.6}_{-8.1}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.2}$	$<\mathbf{1\times10^{-6}}$
			AdvReg	$84.6^{+11.3}_{-9.3}$	$<\mathbf{1\times10^{-6}}$	$85.7^{+11.8}_{-8.1}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.1}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$86.4^{+9.4}_{-8.0}$	$<\mathbf{1\times10^{-6}}$	$87.2^{+10.1}_{-7.3}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.1}$	$<\mathbf{1\times10^{-6}}$
		Multi	Base	$84.8^{+11.5}_{-6.4}$	$<\mathbf{1\times10^{-6}}$	$84.7^{+13.0}_{-9.2}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.1}$	$<\mathbf{1\times10^{-6}}$
			ShapeReg	$86.5^{+9.7}_{-7.1}$	$<\mathbf{1\times10^{-6}}$	$85.9^{+11.7}_{-8.7}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	5.5×10^{-1}
			AdvReg	$85.9^{+10.2}_{-8.6}$	$<\mathbf{1\times10^{-6}}$	$87.6^{+10.2}_{-8.0}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.2}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$87.1^{+9.2}_{-6.3}$	$<\mathbf{1\times10^{-6}}$	$87.1^{+10.4}_{-7.7}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.1}$	$<\mathbf{1\times10^{-6}}$
	VGG	Multi	Base	$89.1^{+7.0}_{-3.6}$	$<\mathbf{1\times10^{-6}}$	$91.4^{+6.5}_{-4.0}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.2}$	$<\mathbf{1\times10^{-6}}$
			CombReg	$89.6^{+6.2}_{-4.1}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{92.3^{+6.1}_{-4.2}}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.2}$	$<\mathbf{1\times10^{-6}}$
	Dense	Multi	Base	$90.4^{+5.5}_{-4.0}$	$<\mathbf{1\times10^{-6}}$	$91.1^{+6.0}_{-3.8}$	$\mathbf{9.0\times10^{-4}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$\mathbf{1.4\times10^{-4}}$
			CombReg	$90.4^{+5.3}_{-3.7}$	1.6×10^{-1}	$90.2^{+6.9}_{-4.5}$	$\mathbf{1.2\times10^{-5}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	Res	Multi	Base	$89.9^{+5.8}_{-4.0}$	$\mathbf{7.3\times10^{-5}}$	$90.5^{+6.8}_{-4.2}$	8.5×10^{-1}	$\mathbf{99.9^{+0.1}_{-0.1}}$	$\mathbf{2.3\times10^{-2}}$
			CombReg	$\mathbf{90.5^{+5.4}_{-3.3}}$	–	$90.8^{+6.6}_{-5.1}$	–	$\mathbf{99.9^{+0.1}_{-0.1}}$	–

Table 5.7 – Statistical analysis between the proposed model and the four backbone architectures: Att-UNet [42], VGG-UNet [60], Dense-UNet [61] and Res-UNet [59] on ankle and shoulder datasets. Regularization methods include: baseline, shape priors-based regularization [53], adversarial regularization [57] and the proposed combined regularization; and bone segmentation strategies comprise: individual, global and multi. Statistical analysis performed through Wilcoxon signed-rank non-parametric test using Dice (%), sensitivity (%) and specificity (%) computed on 2D slices. Bold p-values (< 0.05) highlight statistically significant results for each dataset and for each metric.

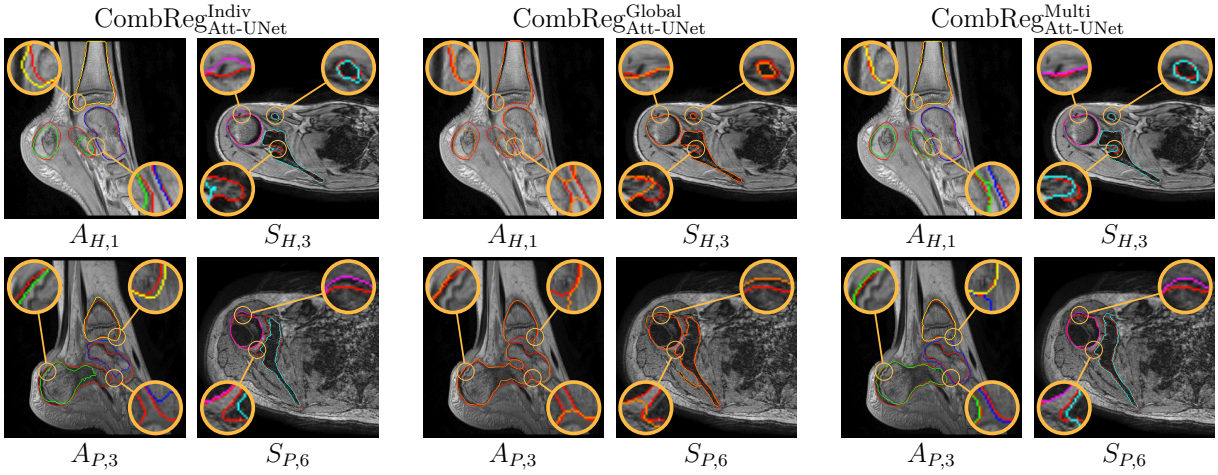


Figure 5.3 – **Visual comparison of bone segmentation strategies using Att-UNet [42] with combined regularization.** Automatic segmentation of ankle and shoulder bones based on Att-UNet with combined regularization using individual-class, global-class and multi-class strategies. Ground truth delineations are in red (–) while predicted bones comprising calcaneus, talus, tibia, humerus and scapula appear in green (–), blue (–), yellow (–), magenta (–) and cyan (–) respectively. Predicted global bone class is in orange (–).

lower than the best method in specificity 2D metric. For shoulder datasets, CombReg$^{\text{Multi}}_{\text{Res-UNet}}$ outperformed other approaches in 2D Dice (90.5%) and 2D specificity (99.9%) metrics, and ranked 1.5% lower than the best model in 2D sensitivity metric.

5.4.4 Qualitative assessment

We first visually compared the combined regularization method using individual-class, global-class, and multi-class Att-UNet models to assess the anatomical validity of the segmentation predictions (Figure 5.3). The individual-class Att-UNet models produced masks based on weights specific to each bone, the global-class Att-UNet models exploited shared features between bones, and multi-class Att-UNet models utilized both specific and shared bone features. It was observed that the global-class mask predictions included fused bone errors in both ankle and shoulder datasets ($A_{H,1}$, $A_{P,3}$, $S_{H,3}$ and $S_{P,6}$ examinations). Thus, exploiting specific bone annotations was necessary to prevent fused-bone errors in predicted delineations. Moreover, shared feature learning in the global-class strategy enforced more accurate delineations ($A_{H,1}$ and $S_{H,3}$). Hence, multi-class Att-UNet leveraged the benefits of learning specific and shared bone features simultaneously, and avoided fused-bones in estimated segmentation masks while producing precise delineations.

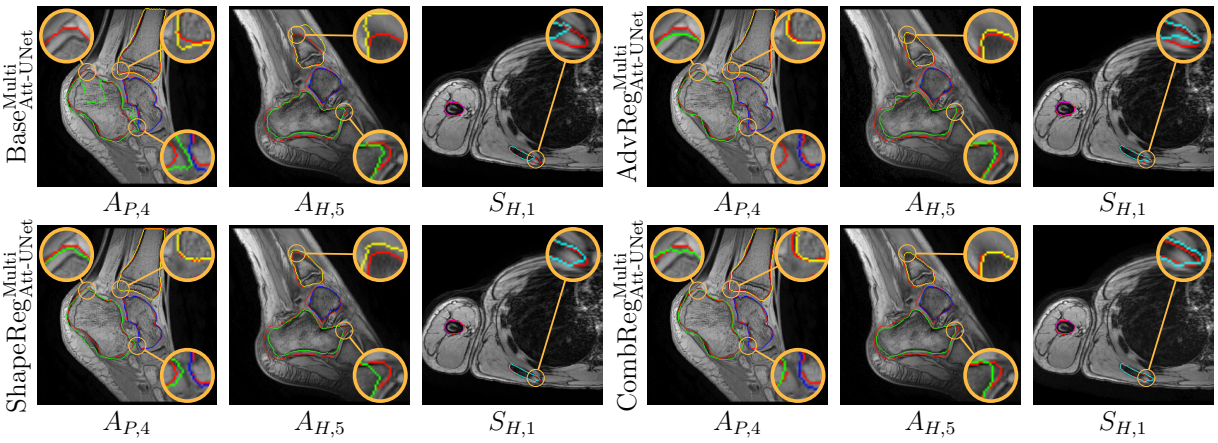


Figure 5.4 – **Visual comparison of regularizations methods using Att-UNet with multi-structure strategy.** Automatic segmentation of ankle and shoulder bones based on Att-UNet [42] with multi-structure strategy using baseline, shape priors [53], adversarial [57], and combined regularizations. Ground truth delineations are in red (–) while predicted bones comprising calcaneus, talus, tibia, humerus and scapula appear in green (–), blue (–), yellow (–), magenta (–) and cyan (–) respectively.

Visual comparison of the four regularization approaches (baseline, shape priors, adversarial, and the proposed combined regularizations) completed the evaluation performed in Figure 4.3. It provided visual evidence of step-wise improvements in segmentation quality from baseline to combined regularization (Figure 5.4). It was clearly observed that each additional regularization improved the segmentation predictions over baseline Att-UNet. Furthermore, baseline Att-UNet did not segment the complete non-ossified area of the scapula, contrary to the compared regularized methods, which incorporated prior knowledge ($S_{H,1}$). More specifically, the shape regularization enforced the model to follow the learned shape representation and promoted smoother bone delineations ($A_{P,4}$), while the adversarial regularization encouraged the model to generate more realistic masks and incited more precise bone delineations ($A_{H,5}$). Meanwhile, the proposed combined regularization fostered the advantages of both former regularizations and provided smoother and more realistic bone extraction ($A_{P,4}$, $A_{H,5}$ and $S_{H,1}$).

Visual comparisons of the pre-trained models (VGG-UNet, Dense-UNet, and Res-UNet) demonstrated that networks benefiting from transfer learning produced highly accurate delineations and captured complex bone shapes (Figure 5.5). The qualitative results further confirmed the advantages of employing networks pre-trained on large non-medical databases along with a combination of shape priors and adversarial regularization to train more generalizable models on scarce pediatric datasets ($A_{P,2}$ and $S_{H,4}$). Most im-

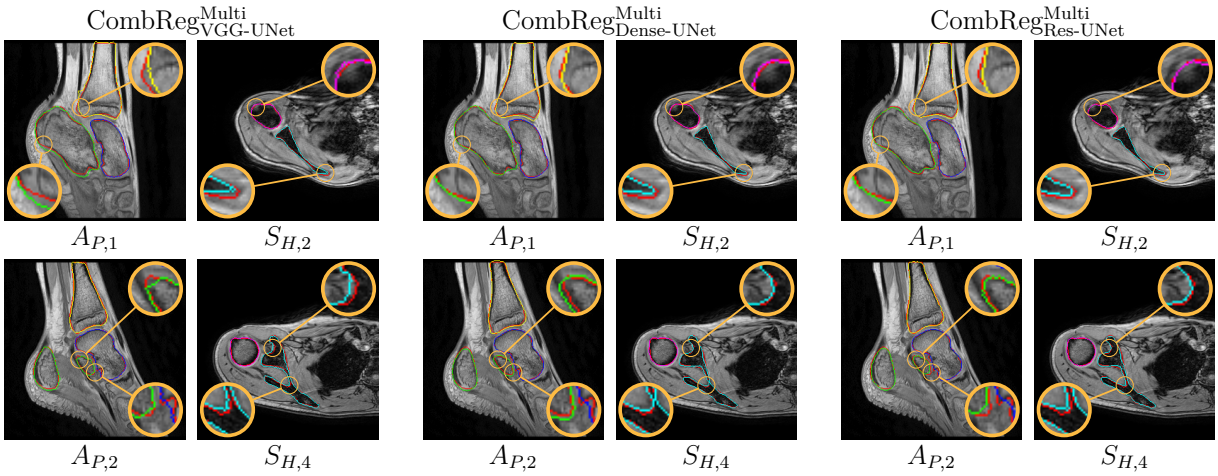


Figure 5.5 – **Visual comparison of pre-trained architectures using multi-class strategy with combined regularization.** Automatic segmentation of ankle and shoulder bones based on VGG-UNet [60], Dense-UNet [61], and Res-UNet [59] using combined regularization with multi-class strategy. Ground truth delineations are in red (–) while predicted bones comprising calcaneus, talus, tibia, humerus and scapula appear in green (–), blue (–), yellow (–), magenta (–) and cyan (–) respectively. Predicted global bone class is in orange (–).

portantly, the proposed pre-trained CombReg$^{\text{Multi}}_{\text{Res-UNet}}$ model together with regularization approaches **effectively segmented non-ossified areas in addition to ossified bones by dealing efficiently with the corresponding intensity variations within a single bone structure** ($A_{P,1}$ and $S_{H,2}$). This outcome can be seen as a crucial need for the image analysis of pediatric musculoskeletal systems (see Chapter 2).

5.5 Discussion

5.5.1 Segmentation performance

This study explored various bone segmentation strategies, regularization methods and backbone architectures and provided an insight into how combination of regularizations can improve the bone segmentation quality in a pediatric, sparse, and heterogeneous MR datasets. We analyzed the performance of each multi-structure strategy with fixed combined regularization (Figure 5.6), the impact of each regularization scheme with the multi-class scheme (Figure 5.7) and the performance of pre-trained models with fixed combined regularization (Figure 5.8) for each MRI dataset.

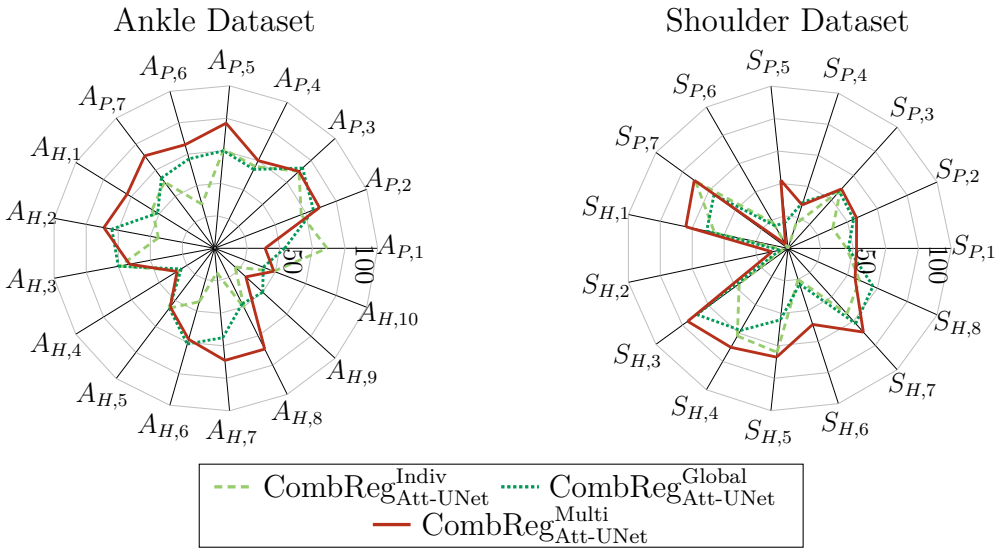


Figure 5.6 – Spider graphs showing scores obtained within ankle and shoulder datasets based on Att-UNet [42] with combined regularization using individual, global and multi strategies. Scores were computed for pathological $A_{P,1}, ..., A_{P,7}$ and healthy $A_{H,1}, ..., A_{H,10}$ ankles, as well as for pathological $S_{P,1}, ..., S_{P,7}$ and healthy $S_{H,1}, ..., S_{H,8}$ shoulders.

From the results obtained on Att-UNet models, we first observed that the multi-class strategy outperformed or at least achieved similar performance compared to individual-class and global-class approaches, on almost all ankle and shoulder examinations (Figure 5.6). However, for two subjects ($A_{P,1}$ and $S_{H,8}$), the multi-class strategy achieved the lowest scores, wherein the extremity of one bone was poorly segmented compared to the other approaches. While, class-wise segmentation provided by the multi and individual strategies yielded bone-specific meshes essential in morphological analysis (see Chapter 2), global bone tissue masks could also be transformed into class-wise predictions using positional or shape information. However, such post-processing proved difficult to implement in practice due to the fused-bone errors observed in the global scheme (Figure 5.3). Secondly, our proposed combined regularization outscored or obtained similar scores as the other regularization schemes on almost all ankle and shoulder examinations (Figure 5.7). However, for $A_{H,4}$ and $S_{P,4}$ subjects, $\text{CombReg}^{\text{Multi}}_{\text{Att-UNet}}$ ranked last and produced poor delineations in which the bone extremities were not well segmented either. From these observations, it appeared that bone extremities remained challenging to be managed by Att-UNet models. A possible explanation relies on the fact that compared to 3D or multi-view fusion models for segmentation [30], [249], our 2D slice-by-slice approaches do not benefit from 3D spatial information. Although our 2D models do not include 3D con-

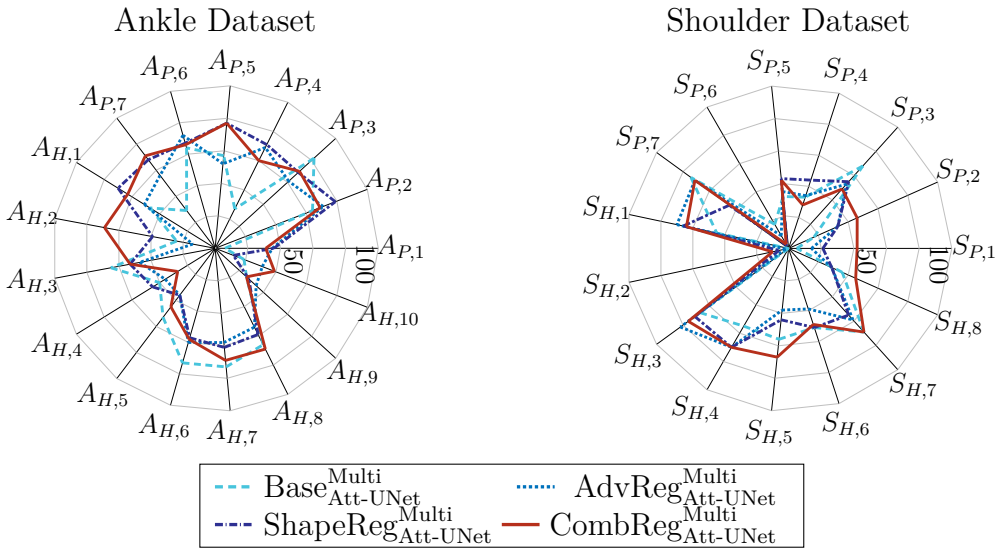


Figure 5.7 – Spider graphs showing scores obtained within ankle and shoulder datasets based on multi-class strategy using Att-UNet with baseline [42], shape priors [53], adversarial [57] and proposed combined regularizations. Scores were computed for pathological $A_{P,1}, ..., A_{P,7}$ and healthy $A_{H,1}, ..., A_{H,10}$ ankles, as well as for pathological $S_{P,1}, ..., S_{P,7}$ and healthy $S_{H,1}, ..., S_{H,8}$ shoulders.

textual information, it is less computationally expensive and requires less GPU memory consumption than 3D approaches.

We reported two outlier examinations $S_{P,6}$ and $S_{H,2}$ for which the Att-UNet models produced poor segmentation results (Figures 5.6 and 5.7). The condition of the patients did not influence the poor segmentation performance, as the two samples were of different types: one pathological ($S_{P,6}$) and one healthy ($S_{H,2}$). However, **both 3D MR images presented a higher level of noise as well as a smaller bone-muscle intensity difference than in the rest of our shoulder dataset** (Figure 5.9). **The relatively poor quality of these examinations was due to patient movements during acquisition.** Hence, the Att-UNet models did not generalize well on these samples. However, we observed that pre-trained models produced more adequate delineations (mean score of 37.5) on these outlier examinations (Figure 5.8). More generally, pre-trained models induced better overall performance than Att-UNet models on all 3D MR images, with $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ producing the best results (e.g., $A_{H,4}$ and $S_{P,4}$). From these observations, we can assume that pre-training on a large set of non-medical images attenuates the effect of noise on segmentation predictions and imposes more robust and generalizable representations. Specifically, approaches based on transfer learning (i.e., VGG-UNet,

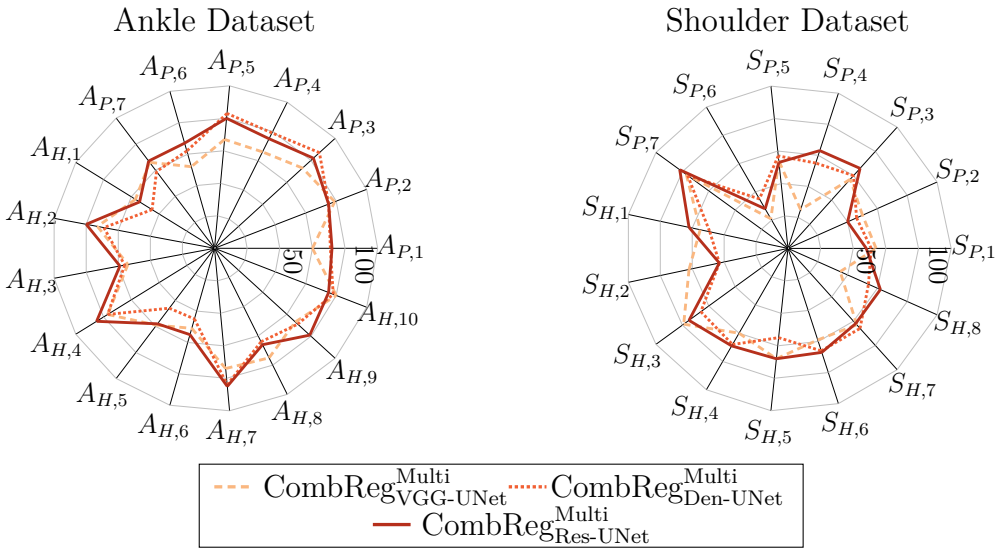
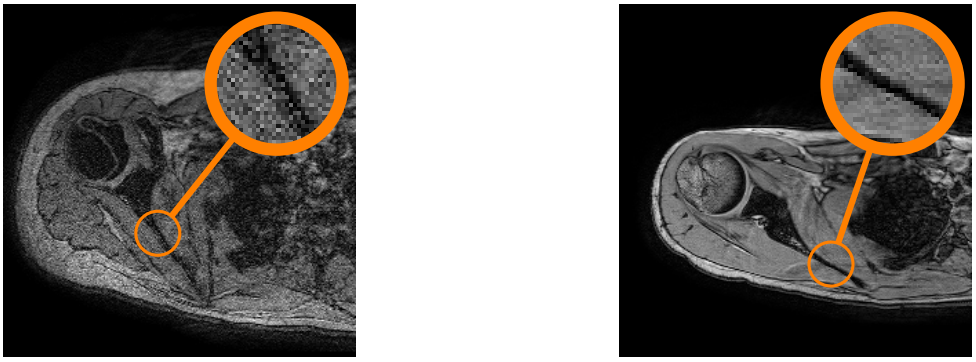


Figure 5.8 – Spider graphs showing scores obtained within ankle and shoulder datasets based on VGG-UNet [60], Dense-UNet [61] and Res-Net [59] employed with multi-class strategy and combined regularization. Scores were computed for pathological $A_{P,1}, ..., A_{P,7}$ and healthy $A_{H,1}, ..., A_{H,10}$ ankles, as well as for pathological $S_{P,1}, ..., S_{P,7}$ and healthy $S_{H,1}, ..., S_{H,8}$ shoulders.

Dense-UNet, and Res-UNet) exploit the knowledge (i.e., network's weights) previously gained while solving an image classification problem to provide better initialization for optimization and extract more robust image features that are then used by the decoder to generate segmentation masks. In this study, more complex and deeper architectures with wider convolutional layers (i.e., VGG-UNet), dense modules (i.e., Dense-UNet) and residual blocks (i.e., Res-UNet) allowed for more efficient optimization and enhanced performance compared to the standard Att-UNet architecture.

The scores obtained also demonstrated that the performance of the different approaches was not influenced by the pathological or healthy status of patients (Figures 5.6, 5.7 and 5.8). For instance, the proposed CombReg$^{\text{Multi}}_{\text{Res-UNet}}$ achieved mean scores of 74.0 and 69.6 on pathological and healthy ankle examinations respectively. As deep models were optimized using a mixture of healthy and impaired joint images, networks were therefore not biased toward any specific population. Moreover, the major differences between pathological and healthy patients was in the shape and relative positioning of the bones rather than in grayscale intensity values. Indeed, both ankle equinus and shoulder OBPP conditions result in osseous deformity and joint malformation [160], [164], while images in each joint dataset were acquired using the same acquisition proto-



$S_{P,6}$ - High noise $S_{H,3}$ - Low noise

Figure 5.9 – Comparison between image samples from $S_{P,6}$ and $S_{H,3}$ examinations. $S_{P,6}$ presented a higher level of noise as well as a smaller bone-muscle intensity due to patient movements during acquisition.

col (see Chapter 2). For example, the shoulder examination $S_{P,6}$ (Figure 5.3) exhibited a deformity of the scapular glenoid shape which resulted in an abnormal positioning of the scapula with respect to the humerus bone. It should be emphasized that the difference in bone intensity due to non-ossified areas were present in both healthy and pathological populations (Figure 5.4, $A_{P,4}$ and $S_{H,1}$), while the motion noise observed in two outlier examinations ($S_{P,6}$ and $S_{H,2}$) were not related to the joint status but were rather due to patient movement during acquisition. Hence, a unique fully automatic segmentation model could be developed for bone segmentation in pediatric MR images, regardless of the presence of bone deformity due to musculoskeletal disorders. Furthermore, **because of its generic nature, our method could be applied to other anatomical joints such as the knee or the hip, as well as on adult imaging datasets**. Such generic framework could provide new perspectives for the management of musculoskeletal disorders, by helping to evaluate treatment response and disease progression as well as being integrated into bio-mechanical models for surgery planning.

5.5.2 Perspectives

Avoiding over-fitting is one of the key problems in machine learning, especially when considering pediatric datasets whose small sample size may induce limited generalizability in deep learning models. Models with too much capacity (i.e., number of trainable parameters) are one typical cause of over-fitting as they may learn the dataset and task too well. In practice, it is therefore essential to design models with optimal capacity which

depends on the task considered and available imaging resources. In this sense, we observed that the employed Res-UNet model (13.6 million parameters) achieved better performance through the leave-one-out evaluation (Table 5.4) compared to VGG-UNet (34.7 million parameters). The multi-class segmentation strategy also allowed us to reduce the number of trainable parameters while transfer learning aimed at reducing over-fitting by using weights learned on a large scale non-medical database. Most importantly, we proposed a combined regularization methodology based on shape priors and an adversarial network to enhance the generalization capabilities of the model during optimization. **It was observed during experiments (Tables 5.3 and 5.4) that all these novel deep learning techniques (as defined in Chapter 1) led to progressive improvement in segmentation performance on unseen images.** Nevertheless, the obtained results still reflect the difficulties of developing generalizable tools without large scale datasets.

Following the performance improvements obtained by the models with a pre-trained encoder on the ImageNet dataset, this approach could be pursued by leveraging large-scale medical imaging datasets extracted from adult cohorts. Indeed, pediatric musculoskeletal imaging datasets can be assumed to share more features with adult musculoskeletal images than natural images. Therefore, transfer learning from large-scale adult musculoskeletal datasets could provide better initialization and generalization capabilities. In the context of pediatric bone segmentation, the MR image dataset from the MICCAI SKI10 knee segmentation challenge [250] could have been used, but this dataset has not been publicly available since 2018 [1]. One could also consider other publicly available large-scale MR image datasets, such as those associated with the KNOAP2020 knee osteoarthritis (OA) prediction challenge [2] or the National Institutes of Health (NIH) knee osteoarthritis initiative (OAI) [3]. Nevertheless, these challenges target the prediction of osteoarthritis severity (i.e., a classification task), and no ground truth segmentation mask is provided. Hence, with the available imaging data and OA grade labels, one can pre-train an encoder but not a full segmentation convolutional encoder-decoder. Furthermore, as mentioned in Section 1.4.2, models trained on adult imaging data would be inapplicable to analyzing pediatric images, each corresponding to a separate image domain. A fine-tuning phase is thus necessary to "adapt" the network weights to the image domain and task considered.

As already discussed in Chapter 4, the interpretability of the learned representation and the analysis of the regularization schemes remain limited, even more so when con-

1. `https://ski10.grand-challenge.org/`
2. `https://knoap2020.grand-challenge.org/`
3. `https://nda.nih.gov/oai/`

sidering an adversarial learning scheme which introduce a competitive optimization procedure between the segmentation network and discriminator. Such training procedure is also prone to numerical instability as simultaneous gradient descent on two neural networks loss functions is not guaranteed to reach an equilibrium [14], [251]. Finally, as already mentioned in Chapter 2, the proposed framework is currently limited to bone tissues segmentation. Future work may therefor aim at improving our model to detect other anatomical structures such as shoulder muscles or ankle cartilages. The severity of the pathologies will then be computed on the basis of a more complete musculoskeletal modeling.

5.6 Conclusion

In this chapter, we proposed and evaluated a partially pre-trained convolutional encoder-decoder with combined regularization from shape priors and an adversarial network, which achieved promising performance for the task of multi-structure bone segmentation on scarce heterogeneous pediatric imaging datasets of the musculoskeletal system. The generalization abilities of the segmentation model was enhanced by exploiting shape priors-based regularization which enforced globally consistent anatomical predictions and an adversarial regularization which encouraged precise delineations. We also employed a transfer learning scheme to provide more robust weight initialization and enhanced performance. In addition, the proposed method exploited specific as well as shared bone features arising from multi-class annotations in order to improve segmentation performance. Finally, we present an original score-based ranking system to simultaneously evaluate multiple architectures, bone segmentation strategies, and regularization schemes.

The obtained results bring new perspectives for the management of musculoskeletal disorders in pediatric population. Nevertheless, the development of generalizable deep learning models remains challenging. In this direction, Part III of this thesis focuses on formalizing and implementing a multi-task, multi-domain framework based on shared representations. Indeed, **while this part illustrated the benefits of multi-structure segmentation to leverage shared features between bones from the same anatomical joint, one may easily assume that bones located in distinct anatomical regions also present common characteristics**.

Part III

Generalizable multi-task, multi-domain segmentation with multi-joint shape priors and multi-scale contrastive regularization

Chapter 6

Multi-joint shape priors for multi-anatomy segmentation

6.1 Introduction

As introduced in Part I, the implementation and optimization of supervised neural networks typically requires a large amount of annotated data. However, the conception of imaging datasets is a slow and onerous process [252] that is even more challenging for pediatric databases [3]. Hence, the inherent scarcity of pediatric imaging resources can induce limited generalization capabilities in neural networks and reduce their performance on unseen images, which in turn may restrict their integration into regular clinical applications. In Part II, we illustrated the effectiveness of two regularization terms, namely shape priors based and adversarial regularizations, to avoid over-fitting issues and improve the performance of multi-structure segmentation models by imposing constraints during training. In particular, shape priors-based regularization has proven to be simple and effective in achieving more accurate and consistent outcomes for medical image segmentation [51]–[53], [55], while adversarial training led to more precise delineations [46], [56]–[58] but revealed to be prone to optimization instability [14].

In deep learning, the concept of regularization, which encompasses all methods aimed at reducing over-fitting, is not limited to techniques based on penalty terms added to the loss function (see Chapter 4). Recently, multi-task [62]–[65] and multi-domain [66]–[71] learning approaches have attracted significant interest from the medical image research community. **Intuitively, multi-task and multi-domain models benefit from parameter sharing to learn more robust and generic representations than their individual counterparts [72]–[74].** These approaches are of particular interest when targeting the segmentation of multiple sparse pediatric datasets of distinct musculoskeletal regions. Indeed, as mentioned in Chapter 2, one can easily assume that MR pediatric datasets arising from distinct anatomical joints (i.e., ankle, knee, shoulder) present shared

features, in terms of shape, pose, and intensity.

For their part, penalty terms can also leverage different prior information to alleviate over-fitting. Thus, we can distinguish between regularization terms imposing constraints on the segmentation generated by the deep model (e.g., shape, boundaries, or topological priors [50]), and those that directly penalize the weights of the network (e.g., L_1 norm to enforce weight sparsity [14]). Since the common goal of these approaches is to reduce over-fitting, it could be beneficial to combine them, as well as to design regularization terms specific to multi-task, multi-domain learning to further improve performance and to build more generalizable models. For instance, to the best of our knowledge, studies on shape priors have never proposed to simultaneously encode multiple anatomical regions in order to leverage position, orientation, size, and shape correlations between similar anatomical objects, such as pediatric bones across distinct musculoskeletal joints.

6.1.1 Multi-task and multi-domain learning

For medical image analysis, multi-task learning aims at leveraging heterogeneous forms of annotations, from global image labels (e.g., healthy versus impaired musculoskeletal joint) to finer-grained and pixel-level segmentation, to improve the performance of deep models [63]. An additional advantage of these approaches is that a variety of tasks (e.g., classification, detection, regression, segmentation) can be solved simultaneously to provide a more complete clinical diagnosis [62]. Certain frameworks have also proposed to incorporate supplementary sub-tasks (e.g., contour prediction or distance map estimation) to refine coarse, non-smooth, and discontinuous segmentation predictions from convolutional models [64]. Additionally, Chen et al. [65] designed an attention-based reconstruction task to leverage unlabeled medical images in a semi-supervised segmentation framework. Hence, two types of multi-task strategies emerge in the literature: cascade of task-specific sub-networks [62], or networks with shared encoder and task-specific decoders [63]–[65]. The former is characterized by sub-models dedicated for each task that can leverage the output of the previous network as input, while the latter defines models with partial parameters sharing between tasks. Both approaches have been reported to perform better than traditional independent models by enabling a better cooperation between tasks [62]–[65]. However, the developed pipelines remain specific to a given intensity domain.

In parallel, recent contributions have proposed to train models over multiple intensity domains (e.g., multi-modal, multi-scanner, multi-center, multi-protocol) with the same segmentation task, in order to leverage a greater amount of training data [66]–[71]. These

architectures aim at benefiting from the correlation between intensity domains to learn more robust domain-invariant feature representations and prove to be particularly useful when dealing with datasets containing a limited number of samples [67]. Numerous multi-domain schemes have been thus implemented and reported to achieve better performance than individual approaches. In particular, one can mention models exploiting transfer learning and fine-tuning between domains [67], models integrating adversarial networks to learn domain-invariant features [71], models that share their latent space only [69], [70], and models composed of domain-specific encoders and a shared decoder [70]. Following this trend to re-use and share an increasing number of parameters, Dou et al. [69] developed a single encoder-decoder segmentation network using shared convolutional kernels and domain-specific internal feature normalization parameters (i.e., batch normalization). While this highly compact architecture reaches superior performance for multi-modal segmentation, their methodology is specific to a given anatomical region of interest (e.g., abdomen or cardiac) and the segmentation task involved the same organs of interest across various intensity domains.

Furthermore, multi-task, multi-domain learning frameworks have been concurrently developed for natural image analysis. In the context of semantic scene labeling, Fourure et al. [75] proposed to train a single network over the union of multiple datasets to address the limited amount of annotated data. In their approach, each dataset is characterized by its own task (segmentation label set) and domain (intensity distribution). Hence, this framework is more generic than traditional multi-task approaches which usually focus on multiple tasks in the same domain or, traditional multi-domain techniques which consider domains containing the same set of objects. Following this, **studies on universal representations in computer vision proposed to employ a single model with agnostic kernels, as visual primitives may be shared across tasks and domains, and dataset-specific layers which enable task and domain specialization [76]–[78]**. These approaches, based on shared representations, have been reported to perform at par or superior to traditional independent models. However, to the best of our knowledge, multi-task, multi-domain learning has rarely been applied to medical image analysis, with the exception of the work of Moeskops et al. [79] which demonstrated that a single neural network can segment multiple anatomies (i.e., brain, breast, and cardiac) simultaneously. Nevertheless, instead of generating pixel-wise segmentation masks, their model relied on a triplanar patches-based approach that predicted the class of a single pixel per input patch, which proved to be computationally expensive. In particular, their architecture

did not comprise a decoder and associated skip connections as in UNet [29], to directly provide whole image segmentation leveraging the global context. Most importantly, patch-wise training lacks the efficiency of fully convolutional training to provide dense output predictions [199]. Their methodology also failed to account for the difference in intensity distribution between domains, as evidenced by the absence of internal domain-specific feature normalization.

6.1.2 Contributions

In this chapter, we propose to implement and optimize a single segmentation network over the union of multiple pediatric imaging datasets arising from separate regions of the anatomy. **Unlike previous methods (developed in Part II) that operate on individual pediatric musculoskeletal joint, our framework simultaneously learns multiple intensity domains and segmentation tasks emerging from distinct anatomical joints. This approach allows to overcome the inherent scarcity of pediatric data while benefiting from more robust shared representations.** To convert and adapt to the multi-task and multi-domain setting, we formalize a segmentation model which incorporates shared representations, domain-specific batch normalization [66]–[69], and domain-specific output layers. Furthermore, we extend the multi-task, multi-domain segmentation learning framework by incorporating multi-joint shape priors which encode the anatomical characteristics of multiple joints and further constrain the delineation tasks to avoid over-fitting. Finally, we illustrate the effectiveness of our approach on three sparse, unpaired (i.e., from different patient cohorts), and heterogeneous pediatric musculoskeletal MR imaging datasets.

The research conducted in this part has been published in the Medical Image Analysis journal [253] and substantially extends a preliminary work presented at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) [254].

The remainder of this chapter is structured as follows. Section 6.2 introduces the mathematical formalism defining multi-task, multi-domain learning. The domain-specific layers (Section 6.2.2) and multi-joint shape priors (Section 6.2.3) are subsequently presented. The experiments are explained in Section 6.3 which encompass the assessment of various multi-domain learning schemes (Section 6.3.2) and the description of the implementation details (Section 6.3.3). Finally, the results are reported and discussed in Section 6.4. Most importantly, we evaluate the proposed multi-anatomy segmentation model with multi-

joint shape priors (Section 6.4.2) and discuss the limitations of the proposed methodology (Section 6.4.3).

6.2 Deep segmentation with domain-specific layers and multi-joint shape priors

6.2.1 Multi-task, multi-domain deep segmentation

We reformulate the baseline segmentation learning framework already introduced in Chapters 3 and 4 for the novel multi-task, multi-domain setting. Let $\mathcal{D}_1, ..., \mathcal{D}_K$ be K different datasets organized such that the k^{th} dataset $\mathcal{D}_k = \{x_i^k, y_i^k\}_{i=1}^{n_k}$ contains n_k pairs of greyscale images x_i^k in intensity domain $\mathcal{I}_k$ and their corresponding class label images y_i^k in label space $\mathcal{C}_k$. Each intensity domain $\mathcal{I}_1, ..., \mathcal{I}_K$ is characterized by its own intensity distribution, while the label spaces $\mathcal{C}_1, ..., \mathcal{C}_K$ represent separate segmentation tasks constituted of different anatomical structure of interest (additionally to the background). Hence, the goal of multi-task, multi-domain deep segmentation is to learn a single mapping S between each intensity domain and its corresponding label space, formally $\forall k \in [1, ..., K]$, $S : \mathcal{I}_k \rightarrow \mathcal{C}_k$.

In what follows, the function S is approximated by a segmentation network composed of a succession of layers whose parameters must be learned during training. More specifically, $S : x_i^k \mapsto S(x_i^k; \Theta, \Gamma)$ is composed of shared parameters Θ and domain-specific weights $\Gamma = \{\Gamma_k\}_{k=1}^K$ selected based on the domain k of the input image. During training, we used the stochastic gradient descent algorithm to optimize the cross-entropy loss defined in a multi-task and multi-domain fashion:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} y_i^k \log(\hat{y}_i^k) \tag{6.1}$$

where $\hat{y}_i^k = S(x_i^k; \Theta, \Gamma)$ was the predicted segmentation. As mentioned in Section 3.2, in practice, the full expression of the loss is an average over classes and pixels. The shared parameters and domain-specific weights were simultaneously derived through this novel optimization scheme. In consequence, the network S learned to segment all structures of interest defined in label spaces $\mathcal{C}_1, ..., \mathcal{C}_K$ across all intensity domains $\mathcal{I}_1, ..., \mathcal{I}_K$. In the following, we employed the Att-UNet architecture [42] as backbone for the segmentation network S. The Att-UNet is an extension of the baseline UNet convolutional encoder-

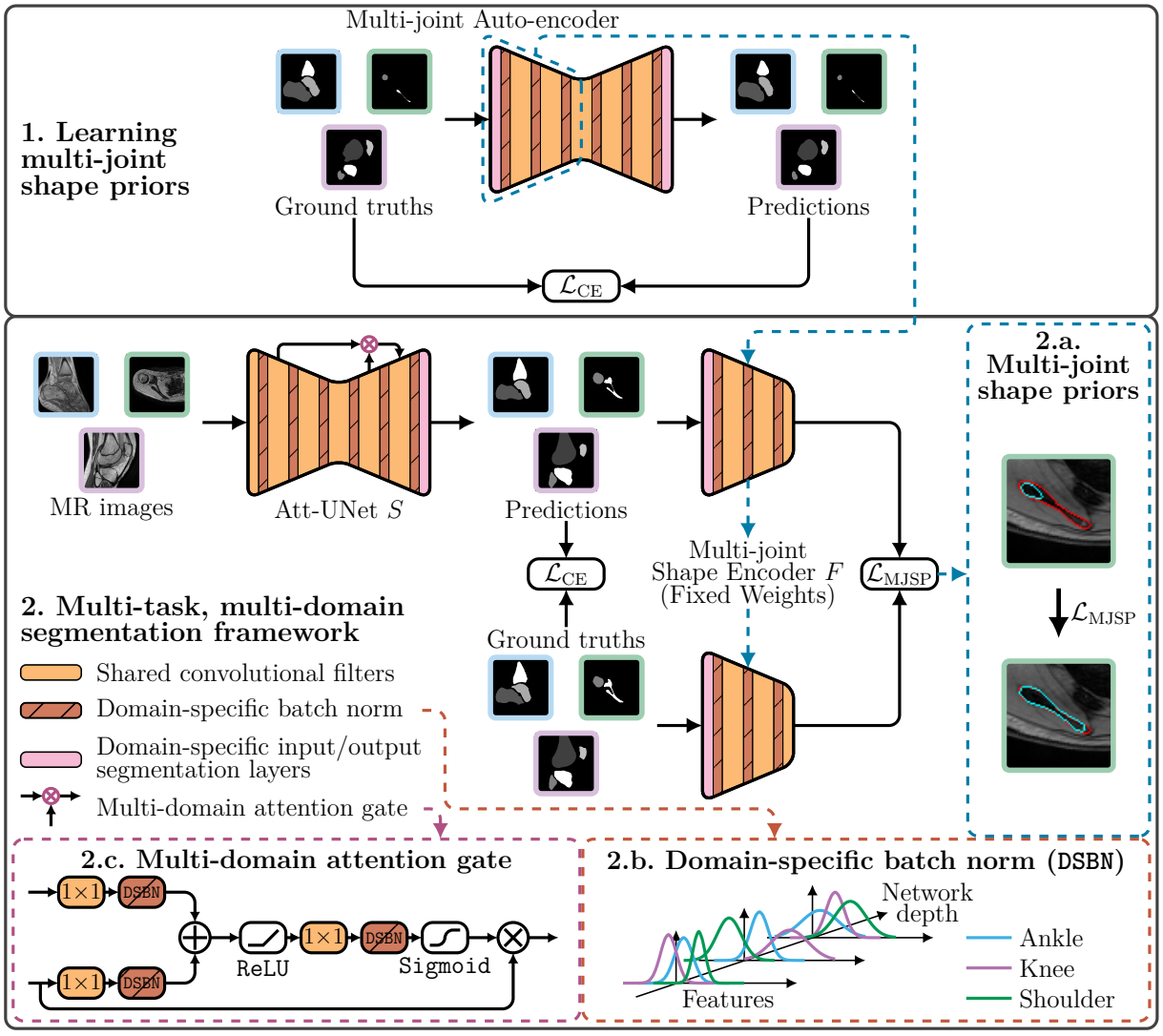


Figure 6.1 – As a first step, a multi-joint auto-encoder learns multi-joint shape priors (1) arising from ground truth segmentation of each joint. As a second step, we optimize a segmentation network S based on Att-UNet [42] in a multi-task, multi-domain framework (2) defined by imaging datasets of three pediatric joints (ankle, knee, and shoulder). The auto-encoder and segmentation networks comprise shared convolutional filters, domain-specific batch normalization (DSBN) (2.b) and domain-specific input/output segmentation layers. Their training procedures rely on the cross-entropy loss function $\mathcal{L}_{CE}$. In addition, Att-UNet incorporates multi-domain attention gates attached to its skip connections (2.c) and its optimization scheme is combined with multi-joint shape priors $\mathcal{L}_{MJSP}$ (2.a).

decoder [29] which integrates spatial attention gates into its skip connections to highlight salient features (see Section 3.4.3).

6.2.2 Domain-specific layers (DSL)

Batch normalization is an ubiquitous transformation found in deep convolutional models which aims at improving convergence speed and generalization abilities of neural networks by normalizing their internal features [41]. However, in multi-domain learning, as the individual statistics of the intensity domains $\mathcal{I}_1, ..., \mathcal{I}_K$ can be very different from each other (Figure 6.1), a domain-agnostic batch normalization layer could lead to defective features [66]–[69]. Specifically, if we consider the l^{th} layer, the mean activation over domains $K^{-1} \sum_{k=1}^{K} \mu_l^k$ could be null while the domain-specific means μ_l^k are non-zero, making a domain-agnostic normalization meaningless.

Thus, to more carefully calibrate the internal features of the model, we employed domain-specific batch normalization functions (`DSBN`) [66]–[69]:

$$\texttt{DSBN}_{\beta_l^k, \gamma_l^k}(v_{i,l}^k) = \gamma_l^k \frac{v_{i,l}^k - \mu_l^k}{\sqrt{(\sigma_l^k)^2 + \epsilon}} + \beta_l^k \tag{6.2}$$

where $v_{i,l}^k$ denoted the feature-map of the l^{th} layer produced by the i^{th} image of the k^{th} dataset, μ_l^k and σ_l^k the domain-specific mini-batch mean and standard deviation respectively. ϵ = 1e-5 was added for numerical stability. As mentioned in Section 3.4.1, batch normalization is performed for each features at layer l independently in practice. The `DSBN` weights $\Lambda_k = \{\beta_l^k, \gamma_l^k\}_l$ thus comprised the domain-specific trainable shift and scale of each feature, at each layer.

Following the definition of `DSBN`, we modified the elementary block of convolutional models (i.e., sequence of convolution, batch normalization, and activation) for multi-domain learning. This novel multi-domain block was based on shared convolution, `DSBN`, and an activation function [1]:

$$u_{i,l+1}^k = \rho(\texttt{DSBN}_{\beta_l^k, \gamma_l^k}(\Theta_l * u_{i,l}^k)) \tag{6.3}$$

Here, $u_{i,l,m}^k$ was the output activations generated by the l^{th} block with the i^{th} image of the k^{th} dataset as input, ρ was a non-linearity (e.g., `ReLU`, `SiLU`, `Sigmoid`), and $u_{i,l}^k$ was the output of the l^{th} layer. As a convention, the input image corresponded to the input of the first layer $u_{i,0}^k = x_i^k$, and we have the relation $v_{i,l}^k = \Theta_l * u_{i,l}^k$). As indicated in [41], the bias of the convolutional layer can be ignored, as its role is subsumed by the shift of

1. This notation can be easily extended to include skip connections or residual layers in which the input is a concatenation or sum of the outputs of previous layer.

the subsequent normalization transformation. Thus, the shared convolutional parameters $\Theta = \{\Theta_l\}_l$ comprised solely the convolutional filters. Based on this new multi-domain block, attention gates [42] were consequently adapted to the multi-domain setting (Figure 6.1). In practice, this corresponded to the modification of each batch normalization layer into its domain-specific equivalent. For instance, as attention gates select spatial regions based on feature activations (e.g., `Sigmoid` activation) [42], we hypothesized that their multi-domain counterpart could help highlight different areas in each domain thanks to domain-specific feature calibration (Figure 6.1).

As intensity domains and segmentation tasks were similar in nature (i.e., pediatric bone in MR images), we assumed that low-level features (e.g., edges, gradients) as well as high-level features (e.g., bone texture, bone shape) were similar across tasks and domains. **We therefore hypothesized that shared convolutional kernels would leverage features shared among tasks and domains to be more robust than their individual counterparts, while the `DSBN` would enable better generalization capabilities thanks to the domain-specific calibration of the internal features.**

Furthermore, as the K segmentation tasks were distinct, a domain-agnostic segmentation layer may predict classes from each label space $\mathcal{C}_1, ..., \mathcal{C}_K$, which is counterproductive [75] (e.g., predicting ankle bones from a shoulder image). Hence, it was essential to employ a dedicated output layer for each domain and task pair. Specifically, if u_i^k denotes the output of the penultimate layer, then:

$$\hat{y}_i^k = \texttt{Softmax}(W_k * u_i^k + b_k) \tag{6.4}$$

was a domain-specific segmentation layer which produced a segmentation mask $\hat{y}_i^k$ with $C_k + 1$ classes. Here, the weights of the domain-specific output segmentation layer $\Xi_k = \{W_k, b_k\}$ corresponded to the final 1×1 (i.e., point-wise) convolutional filter and associated bias.

To recapitulate, the domain-specific layers (DSL) $\Gamma_k = \{\Lambda_k, \Xi_k\}$ comprised the `DSBN` weights Λ_k and the weights Ξ_k of the domain-specific output segmentation layers, whereas the shared parameters Θ corresponded to the classical convolutional filters. Most notably, the domain-specific weights represented a minimal supplementary parameterization with regards to the total number of shared convolutional kernels.

6.2.3 Multi-joint shape priors

As illustrated in Chapter 4, recent works have proposed to integrate into the segmentation network a shape representation of the anatomy, which is learned from ground truth segmentation masks by a deep auto-encoder [51], [53], [55]. To summarize, an auto-encoder is a neural network composed of an encoder F which maps its input to a low-dimensional feature space that compactly encodes the characteristics of the anatomy and a decoder G which reconstructs the original input from the compact representation [51], [53], [55].

We extended the standard shape priors framework (Chapter 4) to the multi-task, multi-domain setting by designing a multi-joint auto-encoder $AE : y_i^k \mapsto G(F(y_i^k; \Theta_F, \Gamma_F); \Theta_G, \Gamma_G)$ which simultaneously learns the shape representation of multiple joints (Figure 6.1). The weights Θ_F and Θ_G corresponded to the shared convolutional kernels of F and G, whereas Γ_F and Γ_G comprised the weights of the `DSBN` and domain-specific input and output segmentation layers of F and G respectively. Similar to the design of the segmentation network, the multi-joint auto-encoder integrated `DSBN` functions to efficiently normalize its internal feature distributions, while the input and output convolutional filters operated on the distinct anatomical structures of interest.

As all segmentation tasks solely comprised pediatric bones, **we assumed that our multi-joint learning scheme would leverage shape features common between musculoskeletal joints to obtain a more robust representations of the anatomy**. Following the definition given in Chapter 4, the multi-joint auto-encoder (MJAE) training procedure was based on the cross-entropy loss function which penalizes the reconstruction of each joint to be dissimilar from the original input [53], [55]. Hence, the loss of the auto-encoder becomes:

$$\mathcal{L}_{\text{MJAE}} = \mathcal{L}_{\text{CE}} := -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k} y_i^k \log(G(F(y_i^k; \Theta_F, \Gamma_F); \Theta_G, \Gamma_G)) \tag{6.5}$$

After training the multi-joint auto-encoder, we integrated its encoder component F into the segmentation framework by computing a multi-joint shape priors term (Figure 6.1). To this end, both predictions and ground truth labels of each joint were projected onto the multi-joint latent shape space by F with learned weights Θ_F and Γ_F. Extending the definition given in Chapter 4, the multi-joint shape priors loss computed the Euclidean

distance between both latent shape representations [53], as follows:

$$\mathcal{L}_{\text{MJSP}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left\|F(y_i^k;\Theta_F,\Gamma_F) - F(\hat{y}_i^k;\Theta_F,\Gamma_F)\right\|^2 \tag{6.6}$$

The minimization of this loss enforced the predicted segmentation of each joint to be in the same low-dimensional manifold as the corresponding ground truth mask [53] and thus encouraged anatomically consistent delineations (Figure 6.1). More precisely, minimizing the Euclidean distance led to similar shape codes for each pair of segmentation masks. As stated in Chapter 4, it should be emphasized that shape codes were represented as 2D feature maps (i.e., auto-encoder bottleneck) with each value encoding a distinct feature of the anatomy. As the weights of the shape encoder remained fixed during this step, the two feature maps were in correspondence, with each value encoding the same global anatomical feature for both ground truth and predicted segmentation masks. Anatomical features typically encompass position, orientation, size, and shape information of each structure of interest as well as their respective intra- and inter-structure correlations. However, as noted in Chapter 4, due to the black-box nature of deep learning models, the interpretability of each anatomical feature remained limited in practice.

The segmentation network S was ultimately trained using the proposed loss function based on a combination of cross-entropy and multi-joint shape priors losses:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1\mathcal{L}_{\text{MJSP}} \tag{6.7}$$

where λ_1 was an empirically set weighting factor.

6.3 Multi-domain segmentation experiments

6.3.1 Imaging datasets

Experiments were conducted on the ankle, knee, and shoulder datasets presented in Chapter 2, and that comprised 20 ankle (A_1, ..., A_{20}), 17 knee (K_1, ..., K_{17}), and 15 shoulder (S_1, ..., S_{15}) 3D MR examinations respectively. It should be emphasized that compared to our previous experiments performed in Part II, we included three additional pediatric ankle examinations (two pathological and one healthy). These examinations were not available at the time of previous experiments. Furthermore, as results obtained

in Part II demonstrated that segmentation networks did not present any population-bias (i.e., better performance on either impaired or healthy examinations), we did not pursue this analysis for the current experiments. Finally, all 2D slices were downsampled to 256×256 pixels and intensities were normalized to have zero-mean and unit variance for each dataset.

6.3.2 Experimental setups

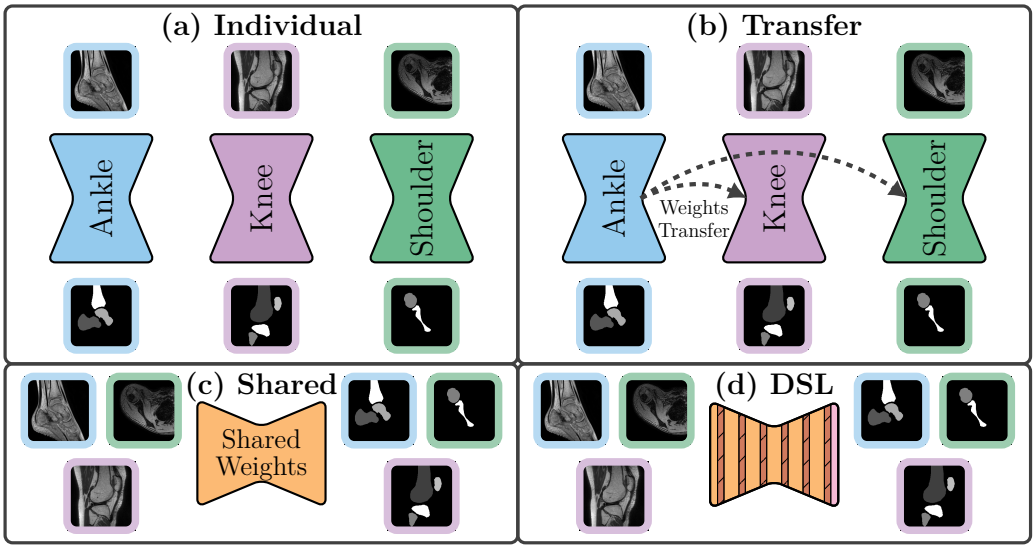


Figure 6.2 – Proposed multi-task, multi-domain segmentation strategies: (a) individual strategy constituted of domain-specific networks, (b) transfer strategy in which weights learned on one domain were transferred to other domains for initialization, (c) shared strategy comprising a single network with all parameters shared between domains, and (d) domain-specific layers (DSL) strategy based on a model with shared convolutional filters along with domain-specific batch normalization (`DSBN`) and segmentation layers. The transfer strategy encompassed all possible combinations of transfer learning between the three domains including $\text{transfer}_{\text{Ankle}}$ (as depicted here), $\text{transfer}_{\text{Knee}}$, and $\text{transfer}_{\text{Shoulder}}$ (both omitted for brevity).

In this chapter, **we investigated various multi-task, multi-domain segmentation strategies with Att-UNet [42] as backbone architecture to assess which one would provide the best segmentation results**. The compared methods built upon Att-UNet comprised four approaches (Figure 6.2): individual (trained on individual domains), transfer (pre-trained on one domain and fine-tuned on the others), shared (trained on all domains at once, with all parameters shared between domains) and DSL

(trained on all domains at once, with shared and domain-specific parameters). The shared approach differed from the DSL scheme by its domain-agnostic batch normalization and shared segmentation layer which predicted bones of interest from all domains with distinct labels (plus background). In this sense, the shared approach was analogous to that developed by [79], although their network architecture differed from Att-UNet and lacked the efficiency to provide dense segmentation predictions. In addition, all networks were trained from scratch with randomly distributed weights except in the transfer scheme in which weights learned on one domain were transferred to other domains for initialization (Figure 6.2). In the transfer scheme, models were not tested on their domain of origin because re-training on the same dataset would not have corresponded to a transfer of knowledge between domains. Hence, $\text{transfer}_{\text{Ankle}}$ denoted models pre-trained on ankle images and fine-tuned on either knee or shoulder domains. We investigated all possible combinations of transfer learning between the three datasets, and defined $\text{transfer}_{\text{Knee}}$ and $\text{transfer}_{\text{Shoulder}}$ schemes in a similar manner.

It should be emphasized that the individual approach of this part, referred to the multi bone segmentation strategies developed in Part II.

Furthermore, to evaluate the contributions of multi-joint shape priors, we performed an ablation study by setting the hyper-parameters weighting factors λ_1 to zero. Specifically, the multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ were incorporated in both shared (using a multi-joint auto-encoder with all parameters shared) and DSL (using a multi-joint auto-encoder with shared and domain-specific parameters) approaches. In this part, we did not employ the shape priors regularization $\mathcal{L}_{\text{Shape}}$ in the individual dataset scheme as performance improvement were already reported in Chapter 4.

6.3.3 Implementation details

Network	AG	Batch Size	#Epochs	Learning Rate	#Parameters		
					Individual	Shared	DSL
Auto-encoder	–	24	8	1e−4	–	7.9M	7.9M
Att-UNet	✓	18	6	5e−4	3×8.7M	8.7M	8.7M

Table 6.1 – Summary of the networks employed during experiments (i.e., multi-joint auto-encoder and Att-UNet [42]) along with their corresponding training hyper-parameter values: batch size, number of epochs and learning rate.

As previously introduced in Chapter 4, our training procedure included two steps:

first, the multi-joint auto-encoder was trained on ground-truth segmentation, and second we optimized the segmentation network to produce delineations of the desired structures of interest (Figure 6.1). The networks were optimized using the Adam optimizer with distinct batch size, number of epochs and learning rate for both (Table 6.1), and these hyper-parameters values remained fixed across all multi-task, multi-domain segmentation strategies (individual, transfer, shared, and DSL). It should be noted that the domain-specific weights introduced marginal supplementary parameterization over the shared approach, while individual schemes represented $K = 3$ times more parameters (Table 6.1). Furthermore, we employed an auto-encoder with more features as compared to Part II in order to accommodate to the multi-joint learning schemes. In shared and DSL schemes, the image batch was equally split between each dataset to prevent domain-bias during optimization. Finally, we explored various values for the multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ and found $\lambda_1 = 0.1$ to be optimal.

Implementation of the deep learning architectures was carried out in PyTorch. As mentioned in Section 3.5.1, we leverage the dynamic computation graphs of PyTorch to implement the DSL operations. Training and inference were performed using an Nvidia RTX 2080 Ti GPU with 12 GB of RAM. All the models were trained on 2D slices with extensive on-the-fly data augmentation due to limited available training data. Data augmentation comprised random rotation ($\pm 22.5°$), shifting ($\pm 10\%$), and flipping in both directions to teach the networks the desired invariance, covariance and robustness properties. Furthermore, the same post-processing as in Part II was employed after each method: first, the obtained 2D segmentation masks were stacked together to form a 3D volume, then we selected the largest connected set of each anatomical structure as final 3D predicted mask, and we finally applied morphological closing by mean of a $5 \times 5 \times 5$ spherical kernel to smooth the resulting boundaries.

6.3.4 Assessment of predicted segmentation

Assessment of the 3D delineations generated by the different methods relied on a comparison against manually annotated ground truths using the same metrics employed in Part II. For each dataset, Dice coefficient, sensitivity, specificity, maximum symmetric surface distance (MSSD), average symmetric surface distance (ASSD), and relative absolute volume difference (RAVD) metrics were computed for each bone and we reported the average scores (see Section 3.5.2). It should be noted that this is contrary to metrics evaluated in Part II, which were computed on global segmentation masks (i.e., global bone

class and background, see Section 4.3.2).

Due to the scarce amount of pediatric examinations, experiments were performed in a leave-one-out manner such that, for each dataset, one examination was retained for validation, one for test, and the remaining data were used to train the model. We iterated through the datasets simultaneously to compute the mean and standard deviation of each metric, and used each examination at maximum once for test. We did not test all combinations between datasets, as this would have introduced redundant observations in the results and drastically increased computation time (i.e., $20 \times 17 \times 15 = 5100$ possible combinations). Consequently, as the shoulder joint dataset contained the fewest number of MR image volumes, 5 ankle (A_{16}-A_{20}) and 2 knee (K_{16}-K_{17}) joint examinations were never included in the test sets since all 15 shoulder samples were already tested. Specifically, the imaging dataset with the fewest samples defined the total number of steps in the leave-one-out evaluation, as we refrained from testing examinations from this dataset multiple times to avoid redundant results and associated bias. All experiments followed the same protocol and imaging examinations with the same index (i.e., A_i, K_i, and S_i) indicated 3D samples tested in the same i^{th} fold of the leave-one-out evaluation. Following standard machine learning practice, the hyper-parameters values (τ, λ_1, λ_2, λ_3, batch size, epochs, learning rate) were selected based on the performance of the model on the validation set.

It should be emphasized that this multi-domain leave-one-out scheme is unlike experiments performed in Part II, in which the leave-one-out evaluation was achieved independently for each dataset.

Finally, we performed visual comparison of predicted segmentation masks. In particular, we evaluated the benefits in segmentation quality of the proposed multi-joint shape priors ($\mathcal{L}_{\text{MJSP}}$) using Att-UNet [42] as backbone architecture in shared and DSL schemes.

6.4 Results and discussion

6.4.1 Quantitative assessment

Assessment of the multi-task, multi-domain segmentation strategies illustrated the advantages of the proposed DSL + $\mathcal{L}_{\text{MJSP}}$ approach over its individual and shared counterparts. The method achieved first or second best performance in ankle and knee datasets, except for ankle MSSD (1.7 mm higher than the best) and knee RAVD (0.8% higher than the best). However, while the proposed method achieved performance at par with individ-

Method			Dice ↑	Sens. ↑	Spec. ↑	MSSD ↓	ASSD ↓	RAVD ↓
Att-UNet	Ankle	Individual	88.2 ± 1.9	88.1 ± 5.4	$\underline{99.8 \pm 0.1}$	17.9 ± 10.8	1.9 ± 1.1	14.1 ± 4.6
		$\text{Transfer}_{\text{Knee}}$	89.5 ± 5.7	88.3 ± 6.0	$\mathbf{99.9 \pm 0.1}$	12.6 ± 10.2	1.6 ± 1.7	14.0 ± 10.9
		$\text{Transfer}_{\text{Shoulder}}$	89.3 ± 4.2	87.5 ± 6.5	$\mathbf{99.9 \pm 0.1}$	$\underline{11.6 \pm 5.0}$	$\underline{1.3 \pm 0.6}$	12.9 ± 8.6
		Shared	88.8 ± 2.5	87.6 ± 6.3	$\mathbf{99.9 \pm 0.1}$	13.4 ± 8.1	1.5 ± 0.8	12.5 ± 7.0
		Shared + $\mathcal{L}_{\text{MJSP}}$	89.6 ± 1.6	$\mathbf{90.6 \pm 5.3}$	$\underline{99.8 \pm 0.1}$	13.4 ± 4.2	$\underline{1.3 \pm 0.3}$	13.1 ± 4.9
		DSL	$\underline{90.6 \pm 2.3}$	88.5 ± 4.6	$\mathbf{99.9 \pm 0.1}$	$\mathbf{11.0 \pm 7.4}$	$\mathbf{1.2 \pm 0.8}$	$\underline{10.9 \pm 5.6}$
		DSL + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{90.9 \pm 1.9}$	$\underline{89.1 \pm 4.6}$	$\mathbf{99.9 \pm 0.1}$	12.7 ± 9.2	$\underline{1.3 \pm 1.2}$	$\mathbf{10.5 \pm 4.4}$
	Knee	Individual	91.1 ± 3.6	88.9 ± 5.5	$\mathbf{99.9 \pm 0.1}$	16.5 ± 12.1	1.6 ± 1.5	10.7 ± 6.1
		$\text{Transfer}_{\text{Ankle}}$	92.8 ± 2.9	91.1 ± 3.3	$\mathbf{99.9 \pm 0.1}$	12.4 ± 10.3	1.0 ± 0.9	7.6 ± 5.4
		$\text{Transfer}_{\text{Shoulder}}$	92.5 ± 2.4	90.7 ± 4.0	$\mathbf{99.9 \pm 0.1}$	13.1 ± 11.3	1.0 ± 0.8	7.8 ± 4.5
		Shared	91.7 ± 3.2	88.5 ± 4.8	$\mathbf{99.9 \pm 0.1}$	12.5 ± 9.0	1.4 ± 1.4	9.5 ± 6.0
		Shared + $\mathcal{L}_{\text{MJSP}}$	$\underline{93.6 \pm 1.8}$	91.9 ± 3.1	$\mathbf{99.9 \pm 0.1}$	$\mathbf{7.9 \pm 8.4}$	$\underline{0.8 \pm 0.9}$	$\underline{6.4 \pm 3.1}$
		DSL	93.3 ± 2.5	$\underline{92.8 \pm 3.5}$	$\mathbf{99.9 \pm 0.1}$	12.8 ± 12.1	1.1 ± 1.3	$\mathbf{6.0 \pm 4.0}$
		DSL + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{93.8 \pm 2.5}$	$\mathbf{93.0 \pm 4.2}$	$\mathbf{99.9 \pm 0.1}$	$\underline{9.4 \pm 5.9}$	$\mathbf{0.7 \pm 0.4}$	6.8 ± 4.1
	Shoulder	Individual	80.9 ± 10.1	77.7 ± 14.9	$\mathbf{99.9 \pm 0.1}$	26.9 ± 14.1	2.4 ± 1.8	$\underline{15.2 \pm 16.7}$
		$\text{Transfer}_{\text{Ankle}}$	$\underline{82.6 \pm 8.8}$	$\underline{79.8 \pm 12.5}$	$\mathbf{99.9 \pm 0.1}$	26.9 ± 17.2	$\underline{2.2 \pm 1.8}$	17.0 ± 10.1
		$\text{Transfer}_{\text{Knee}}$	$\mathbf{83.3 \pm 10.1}$	$\mathbf{80.5 \pm 12.5}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{24.0 \pm 14.3}$	$\mathbf{2.1 \pm 2.4}$	$\mathbf{13.7 \pm 13.2}$
		Shared	80.1 ± 9.6	76.6 ± 12.9	$\mathbf{99.9 \pm 0.1}$	28.1 ± 12.2	2.7 ± 1.8	18.3 ± 12.4
		Shared + $\mathcal{L}_{\text{MJSP}}$	80.7 ± 9.0	79.2 ± 12.3	$\mathbf{99.9 \pm 0.1}$	$\underline{25.0 \pm 15.6}$	2.3 ± 1.8	19.4 ± 12.0
		DSL	80.9 ± 7.3	77.6 ± 11.6	$\mathbf{99.9 \pm 0.1}$	34.4 ± 19.1	3.3 ± 2.3	19.4 ± 12.2
		DSL + $\mathcal{L}_{\text{MJSP}}$	81.4 ± 9.0	79.2 ± 14.3	$\mathbf{99.9 \pm 0.1}$	31.0 ± 18.5	2.5 ± 2.2	15.7 ± 12.2

Table 6.2 – Leave-one-out quantitative assessment of Att-UNet [42] using individual, transfer, shared, and DSL strategies employed with multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ on ankle, knee, and shoulder datasets. Metrics include Dice (%), sensitivity (%), specificity (%), MSSD (mm), ASSD (mm), and RAVD (%). Mean scores and standard deviations reported in bold and underlined respectively correspond to the first and second best results obtained for each dataset.

ual and shared schemes on shoulder joint examinations, the $\text{transfer}_{\text{knee}}$ and $\text{transfer}_{\text{ankle}}$ approaches respectively ranked first and second best on nearly all metrics, with the exception of MSSD and RAVD for $\text{transfer}_{\text{ankle}}$ which reached the third and fourth ranks respectively. In particular, the performance of DSL + $\mathcal{L}_{\text{MJSP}}$ were 1.9% lower than the best in Dice, 1.3% lower in sensitivity, 7.0 mm higher in MSSD, 0.4 mm higher in ASSD, and 2.0% higher in RAVD.

6.4.2 Qualitative assessment

Visual comparison of the multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ provided visual evidence of gradual improvements in segmentation quality for both shared and DSL Att-UNet models (Figure 6.3). **Shape priors were clearly observed to promote globally more consistent and smoother contours for all anatomical joints** by forcing the model

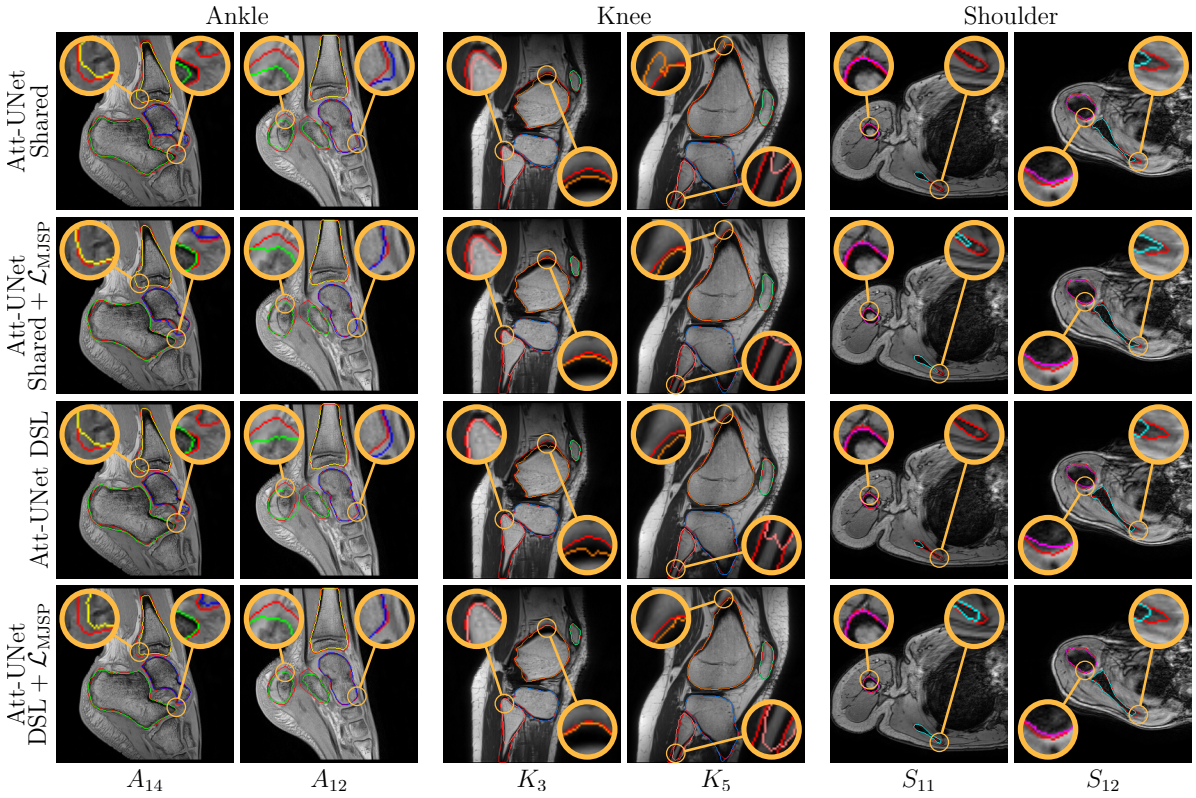


Figure 6.3 – **Visual comparison of the multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ using Att-UNet architecture.** Automatic segmentation of ankle, knee, and shoulder bones based on Att-UNet [42] employed in shared and DSL strategies. Ground truth delineations are in red (–) while predicted bones appear in green (–) for calcaneus, blue (–) for talus, yellow (–) for tibia (distal), orange (–) for femur (distal), pink (–) for fibula (proximal), light green (–) for patella, light blue (–) for tibia (proximal), magenta (–) for humerus, and cyan (–) for scapula.

to follow the learned non-linear multi-joint shape representation. More specifically, incorporation of shape priors allowed the segmentation of the complete talus (A_{14}), fibular (K_5), and scapular shapes (S_{11} and S_{12}), which were previously partially detected by both shared and DSL Att-UNet models.

6.4.3 Limitations

Regarding the performance of the multi-task, multi-domain strategies, we observed that all transfer learning schemes ($\text{Transfer}_{\text{Ankle}}$, $\text{Transfer}_{\text{Knee}}$, and $\text{Transfer}_{\text{Shoulder}}$) provided performance improvements compared to individual models on all datasets (Table 6.2), indicating a better initialization than randomly set weights by exploiting features

correlation and knowledge transfer between each task and domain pair. Compared to individual and transfer approaches, the results of shared and DSL schemes on both ankle and knee datasets indicated noticeable improvements while the results on shoulder examinations were less evident (Table 6.2). Nevertheless, **both shared and DSL schemes offer an additional advantage compared to transfer approach by learning all task and domain pairs simultaneously rather than in a sequential manner that is prone to catastrophic forgetting**. It should also be noted that in the shared segmentation scheme, predicted segmentation output that did not belong to the image task were considered as background in order to obtain a fair comparison against individual, transfer, and DSL strategies. In practice, confusion between tasks was very low, with the mean percentage of voxels per 3D examination labeled with a class foreign to the target segmentation classes (e.g., humerus identified in ankle MR images) being less than 0.001% for all tasks. A low confusion between tasks was also reported by [79] in their multi-tasks segmentation framework.

These results illustrated that learning to simultaneously segment multiple anatomical regions is a challenging setting, and the possible limitations of the Att-UNet architecture due to its low complexity, depth and width. Indeed, one could leverage more advanced state-of-the-art architecture (e.g., from the Inception, DenseNet, or EfficientNet family) to achieve better performance on all datasets. Nevertheless, we observed that the multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ improved the segmentation performance for each anatomical joint and in both shared and DSL schemes. This further demonstrated the benefits of shape priors scheme based on deep auto-encoder which were already proven effective in Part II, but within each separate anatomical joint.

6.5 Conclusion

In this chapter, we proposed a multi-task, multi-domain segmentation framework which achieved promising performance on three scarce pediatric imaging datasets of distinct musculoskeletal joints. We formalized a framework based on shared representations, domains-specific batch normalization (`DSBN`), and domain-specific output segmentation layers which allow to easily adapt any convolutional segmentation network to a multi-task, multi-domain learning setting. Finally, we incorporated multi-joint shape priors to enforced globally consistent segmentation predictions and reduce over-fitting issue.

Nevertheless, even though the proposed multi-task and multi-domain model integrate

task- and domain-specific information through specialized layers and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ constraints, domain prior knowledge could be further exploited to improve the generalizability of learned shared representations. To address this issue, we propose in Chapter 7 a contrastive regularization which impose domain-specific clusters in the shared representations of the model.

Chapter 7

Enhanced generalizability via multi-scale contrastive regularization

7.1 Introduction

In this chapter, we extend our approach to reduce over-fitting issues through multi-task, multi-domain learning for the segmentation of sparse pediatric imaging datasets originating from separate musculoskeletal joints. The methodology previously presented in Chapter 6 is expanded following two regularization schemes: through transfer learning from non-medical images, as already illustrated in Chapter 5, and through a penalty term on the learned shared representation, tailored specifically to multi-task, multi-domain learning framework. In particular, transfer learning leverages low-level features shared between image types to improve generalization capabilities, especially when targeting small datasets [44]. Hence, a standard practice in medical image analysis relies on exploiting the weights of state-of-the-art neural networks (e.g., EfficientNet [191] or Transformers [255]) trained on the ImageNet large-scale natural images database [242]. On the other hand, even though multi-task and multi-domain models can integrate task- and domain-specific information through specialized layers, task and domain prior knowledge could be further exploited to improve the generalizability of learned shared representations.

In this direction, the work of Dou et al. [69] introduced a knowledge distillation regularization loss whose goal is to constrain the prediction distributions of their multi-modal segmentation model to be similar across domains. Similarly, Zhu et al. [80] imposed a Gaussian mixture distribution on the shared latent representation of their image translation network to preserve fine structures between domains. However, such a hypothesis may be too restrictive. Indeed, in representation learning, a good representation can be characterized by the presence of natural clusters corresponding to the classes of the problem

(i.e., disentangled representation) [81]. Hence, a number of self-supervised representation learning techniques focus on pulling together data points from the same class and pushing apart negative samples in embedded space using a contrastive metric [82]–[84]. A recent contribution extended this idea to fully-supervised image classification setting by leveraging the label information and considering many positive anchors simultaneously [85]. Thus, the contrastive regularization maximizes the performance of the classifier by imposing intra-class cohesion and inter-class separation in latent space. In the context of semi-supervised medical image segmentation, Hu et al. [256] exploited unannotated data by designing a contrastive loss forcing pixels from the same class to assemble in embedded space. Unlike [80], in these non-parametric contrastive approaches, it is not necessary to define a prior distribution (e.g., Gaussian, Poisson) for the latent variables. Hence, **contrastive regularization techniques appear more generic and appropriate to impose domain-specific clusters in the shared representations of deep multi-task, multi-domain models**.

7.1.1 Contributions

In this chapter, we propose a multi-task, multi-domain segmentation framework to address the scarcity issue associated with pediatric imaging data by leveraging multiple anatomical regions. Our method learns multiple intensity domains and segmentation tasks arising from separate musculoskeletal joints. We extend the multi-anatomy learning framework by integrating a multi-scale contrastive regularization during optimization to improve the generalization capabilities of neural networks. **Following classical contrastive approaches that operate on image classes, we leverage dataset label information to enhance intra-domain similarity and impose inter-domain margins.** However, compared to standard contrastive learning, our contrastive regularization is applied in a multi-scale fashion, in the same spirit as deep supervision [257]. In addition, we leverage a pre-trained Efficient encoder [191] as backbone for the segmentation network to further reduce data scarcity limitations. Finally, we extend the evaluation of the multi-task, multi-domain learning framework initiated in Chapter 6.

As previously mentioned, the research conducted in this part has been published in the Medical Image Analysis journal [253] and substantially extends a preliminary work presented at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) [254].

7.2 Efficient segmentation network with multi-scale contrastive regularization

In this section, we first describe the proposed multi-task, multi-domain segmentation network built upon Efficient-UNet (Section 7.2.1). We then incorporate the multi-scale contrastive regularization into our model (Section 7.2.2).

7.2.1 Efficient segmentation network with pre-trained encoder

We briefly recall the multi-task, multi-domain segmentation framework developed in Chapter 6 which incorporates a segmentation network S and a multi-joint shape encoder F parameterized by shared convolutional filters Θ and domains specific layers Γ including domain-specific batch normalization (`DSBN`) and domain-specific input/output segmentation layers. The optimization of S minimizes a loss based on cross-entropy defined in a multi-task, multi-domain setting and multi-joint shape priors computed by F with fixed weights, as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{MJSP}} \tag{7.1}$$

The framework developed in Chapter 6 relied on the Att-UNet [42] as backbone architecture for the segmentation network S, nevertheless our training strategy is architecture independent. The previous methodologies developed in Chapter 5 illustrated that leveraging transfer learning and fine-tuning from the ImageNet database [242] lead to improved segmentation outcomes, especially in the low data regime specific to pediatric imaging resources [40], [44]. Hence, as in Chapter 5, the architecture of the neural network S was based on UNet [29] whose encoder branch was replaced by a classification network with weights previously trained on ImageNet classification task (Figure 7.2). We further advanced this strategy by integrating an Efficient classification network from the EfficientNet family as encoder [191]. Specifically, we employed the `EfficientNetB3` encoder which incorporates mobile inverted bottlenecks convolutional blocks (MBConv [258]) to simultaneously balance the network depth, width, and resolution while improving predictive performance [191]. The `EfficientNetB3` represents a good compromise between complexity and performance compared to other model versions (from `EfficientB0` to `EfficientB7`).

To fit the `EfficientNetB3` image dimensions, we concatenated three copies of each MR slice to extend them from single greyscale channel to three channels. The encoder branch was then built on classical convolution, batch normalization, and sigmoid linear

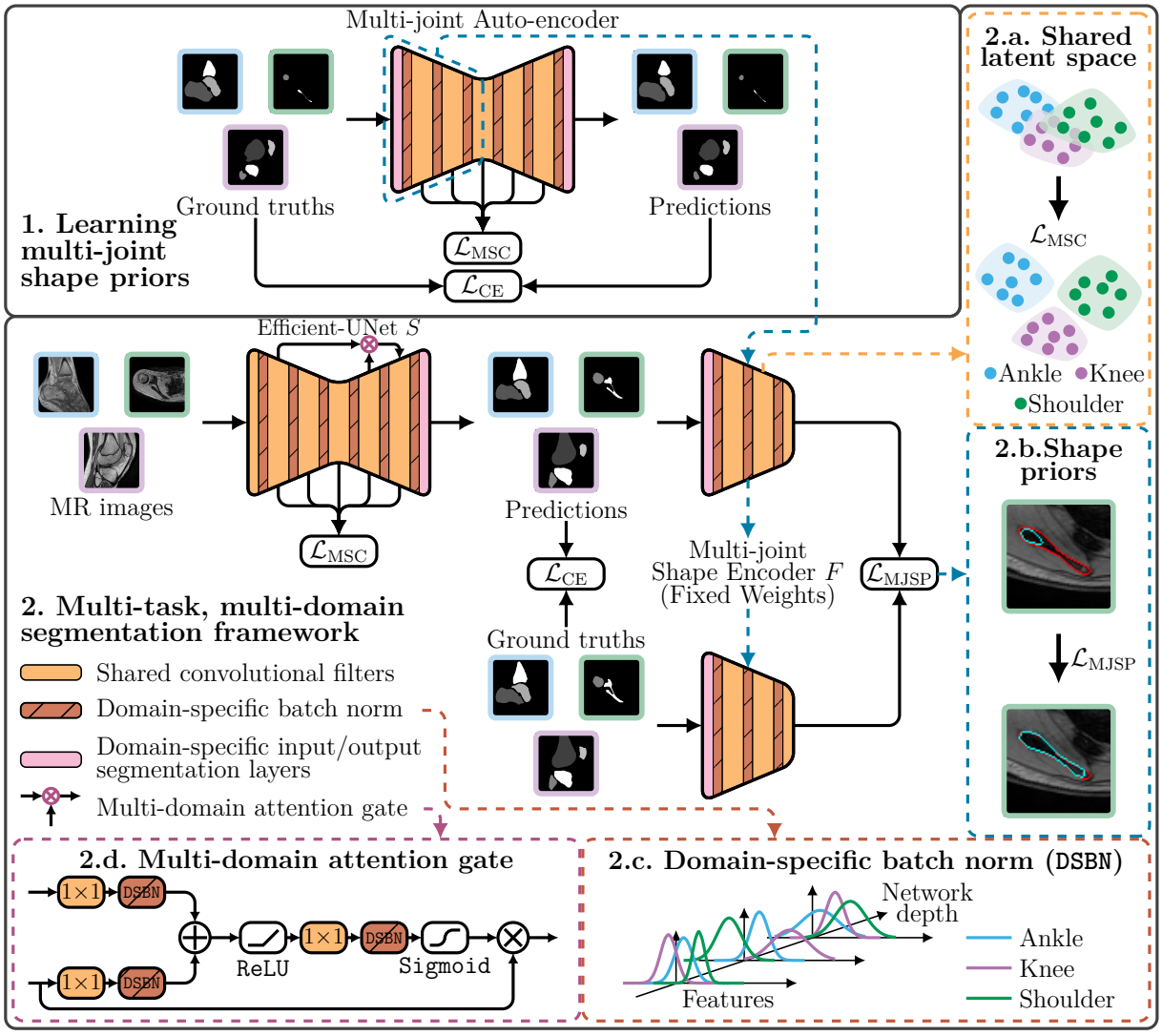


Figure 7.1 – The first step of the proposed method involves a multi-joint auto-encoder that learns multi-joint shape priors (1) arising from ground truth segmentation. As a second step, we optimize a segmentation network S based on Efficient-UNet [191] in a multi-task, multi-domain framework (2) defined by imaging datasets of three pediatric joints. The auto-encoder and segmentation networks comprise shared convolutional filters, domain-specific batch normalization (DSBN) calibrating the internal features statistics (2.c) and domain-specific input/output segmentation layers delineating distinct anatomical regions. Their training procedures rely on the cross-entropy loss function $\mathcal{L}_{CE}$ and integrate a multi-scale contrastive regularization $\mathcal{L}_{MSC}$ to promote inter-domain separation in the shared representations (2.a). In addition, Efficient-UNet incorporates multi-domain attention gates (2.d) and multi-joint shape priors $\mathcal{L}_{MJSP}$ computed by the multi-joint shape encoder F to enforce anatomically consistent predictions (2.b).

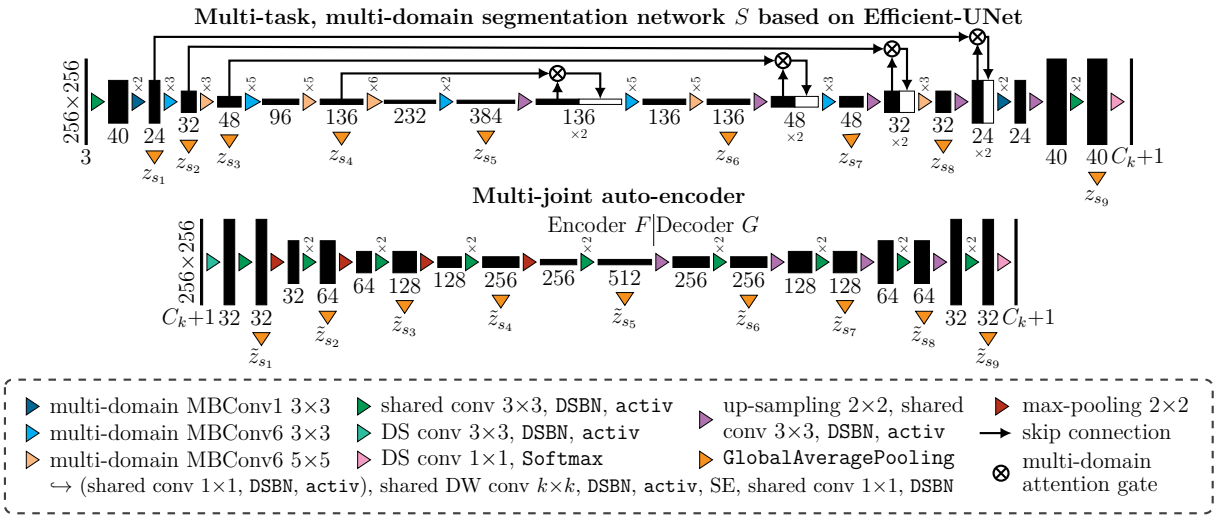


Figure 7.2 – Proposed neural network architectures: multi-task, multi-domain segmentation network S based on Efficient-UNet (top) [191] and multi-joint auto-encoder (bottom) comprising encoder F and decoder G. The multi-scale embedding $(z_{s_1}, ..., z_{s_9})$ and $(\tilde{z}_{s_1}, ..., \tilde{z}_{s_9})$ are obtained via `GlobalAveragePooling`. C_k+1 denotes the number of classes (plus background) in the k^{th} segmentation task while activations (`activ`) correspond to either `SiLU` (Efficient-UNet) or `ReLU` (auto-encoder) functions. The multi-domain MBConv block integrates shared point-wise (1×1) and depth-wise (DW) convolutions, domain-specific batch normalization (`DSBN`) and squeeze-and-excite (SE) modules [191].

unit (`SiLU`) non-linearity along with MBConv blocks (MBConv1-6, Figure 7.2) consisting of point-wise and depth-wise convolutions, as well as additional squeeze-and-excite modules [191]. Specifically, combination of point-wise and depth-wise convolutions layers allows to reduce the number of parameters by leveraging the decoupling of cross-channel correlations and spatial correlations [259]. For their parts, squeeze-and-excite modules aim at improving performance by adaptively recalibrating channel-wise features through explicit modeling of interdependencies between channels [260]. The overall architecture (i.e., depth, width, and resolution) of `EfficientNetB3` encoder is then defined in a principled way using a compound scaling coefficient [191]. Ultimately, `EfficientNetB3` produced a 384 dimensional output and the resulting feature-map corresponded to the central part between the contracting and expanding paths of S (Figure 7.2). Next, we constructed a symmetrical decoder branch with up-sampling layers, classical convolutions and MBConv blocks (Figure 7.2). Contrary to encoder weights that are pre-trained on ImageNet [242], the decoder weights were randomly initialized. Finally, to improve both model interpretability and performance, we employed spatial attention gates to implicitly suppress irrelevant regions of the input images while highlighting salient features [42]. These

modules attached to the skip connections selected important features using contextual information from the decoding branch (Figure 7.1).

As illustrated in Chapter 6, we adapted the segmentation network S, including MB-conv modules and spatial attention gates (Figure 7.2), to the multi-domain setting by modifying each batch normalization layer into its domain-specific equivalent (i.e., `DSBN`).

7.2.2 Multi-scale contrastive regularization

We consider the multi-domain convolutional block introduced in Chapter 6, which was based on shared convolution filters Θ, `DSBN` transformations with domain-specific trainable shift β_l^k and scale γ_l^k, and an activation function ρ:

$$u_{i,l+1}^k = \rho(\texttt{DSBN}_{\beta_l^k,\gamma_l^k}(\Theta_l * u_{i,l}^k)) \tag{7.2}$$

The multi-domain block mapped its inputs to a shared representation in which features were shifted and scaled according to their domain before applying a non-linear activation. Here, we hypothesized that learning shared representations with domain-specific clusters would enhance the generalization capabilities of the model (i.e., segmentation network or multi-joint auto-encoder) and improve the accuracy of the predicted delineations. More precisely, we assumed that a local variation in the output of each multi-domain block should preserve the category of the domain [81]. Hence, **we designed a novel regularization term aimed at disentangling domain representations by conserving intra-domain cohesion and inter-domain separation in the shared latent space** (Figure 7.1). The proposed contrastive regularization was adapted from image classification [85] to multi-task, multi-domain segmentation using the known domains labels.

However, rather than applying the contrastive regularization after each multi-domain block (i.e., after each non-linearity), we imposed the clusterization constraints at each scale of the model (i.e., in a multi-scale manner) to reduce computational complexity. To this end, we considered an ensemble of layers indices $\mathcal{S}$ corresponding to the different spatial scale of the segmentation network, which were symmetrically distributed between the encoder and the decoder (Figure 7.2). Our multi-scale approach untangled the domain representations at each stage of the encoder and decoder modules in a deeply-supervised manner. Since the semantic information extracted and captured by the neural network differed at each scale as well as across scales, we hypothesized that it was necessary to enforce a multi-scale regularization to achieve better generalization capabilities compared

to the single scale constraint.

Let $z_{i,s}^k = \texttt{GlobalAveragePooling}(u_{i,s}^k)$ be the embedding of x_i^k at scale $s \in \mathcal{S}$ to which we applied `GlobalAveragePooling` to project the data in a lower-dimensional space $\mathbb{R}^d$ invariant to spatial transformations (e.g., rotation, translation, flipping), allowing global comparison of image representations originating from different domains (Figure 7.2). The dimensionality d of the representations were thus distinct at each scale and $z_{i,s}^k$ was then normalized to lie on the unit hyper-sphere, which enabled to measure distances by using an inner product [85].

We note $\mathcal{P}_i^k = \{j \in [1, ..., n_k] : j \neq i\}$ the set of indexes of all images from the same domain as x_i^k (i.e., positive pairs) and $n = \sum_{k=1}^K n_k$ the total number of images across domains. The multi-scale contrastive loss was defined as follows:

$$\mathcal{L}_{\text{MSC}} = -\frac{1}{n|\mathcal{S}|} \sum_{s \in \mathcal{S}} \sum_{\substack{1 \leq k \leq K \\ 1 \leq i \leq n_k}} \frac{1}{|\mathcal{P}_i^k|} \sum_{j \in \mathcal{P}_i^k} \log \left(\frac{\exp(z_{i,s}^k \cdot z_{j,s}^k / \tau)}{\sum\limits_{\substack{(k',i') \\ \neq (k,i)}} \exp(z_{i,s}^k \cdot z_{i',s}^{k'} / \tau)} \right) \tag{7.3}$$

where $z_{i,s}^k \cdot z_{j,s}^k$ denoted the inner product between two L^2 normalized representations (i.e., cosine similarity) and τ was the temperature hyper-parameter which controlled the smoothness of the loss as well as imposed hard negative/positive predictions [83], [85]. As the cosine similarity was bounded in the interval $[-1, 1]$ regardless of the dimensionality of the representations, we assumed that the temperature τ should be constant over scales. **Optimization of $\mathcal{L}_{\text{MSC}}$ encouraged the model to produce, at each scale, closely aligned representations for all pairs from the same domain and orthogonal representations for negative couples.** Thus, the multi-scale contrastive regularization gathered the embedding from the same domain, while simultaneously separating clusters from different domains (Figure 7.1).

Based on our multi-scale contrastive regularization, we imposed a clusterization constraint on the shared representations of both the multi-joint auto-encoder ($\tilde{z}_{s_1}, ..., \tilde{z}_{s_9}$ as denoted in Figure 7.2) and segmentation network ($z_{s_1}, ..., z_{s_9}$ as denoted in Figure 7.2). In particular, the loss of the auto-encoder integrated the contrastive term to promote separated low-dimensional manifolds for each anatomical joint:

$$\mathcal{L}_{\text{MJAE}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{MSC}} \tag{7.4}$$

with λ_1 as an empirically set weighting factor.

For its part, the segmentation network S was ultimately trained using the proposed loss function based on a combination of cross-entropy, multi-scale contrastive regularization and multi-joint priors losses:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_2 \mathcal{L}_{\text{MSC}} + \lambda_3 \mathcal{L}_{\text{MJSP}} \tag{7.5}$$

where the weighting factors λ_2 and λ_3 were empirically set.

7.3 Experiments

Experiments were performed on ankle, knee, and shoulder pediatric MR datasets (see Section 2.4.2) following the same setting as described in Chapter 6. Furthermore, we extended the ablation study conducted using Att-UNet backbone architecture to include the single scale $\mathcal{L}_{\text{SSC}}$ and multi scale $\mathcal{L}_{\text{MSC}}$ contrastive regularizations. Specifically, as the intensity domains $\mathcal{I}_1, ..., \mathcal{I}_K$ were not differentiated in the shared approach, the multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$ could only be integrated in the DSL (i.e., domain-specific layers) scheme. We also assessed the advantages of the multi-scale contrastive over a single-scale contrastive ($\mathcal{L}_{\text{SSC}}$) method which only constrains the network bottleneck (i.e. encoder output). The ablation study was performed by setting the hyper-parameters weighting factors λ_1, λ_2, and λ_3 to zero respectively.

7.3.1 Pre-trained architectures

In the next experiment, we evaluated the performance of our method based on Efficient-UNet with pre-trained `EfficientNetB3` encoder [191], DSL, multi-scale contrastive regularization and multi-joint shape priors against Inception-Net [86] and Dense-Net [61] backbone architectures similarly pre-trained on large natural image database [242]. Specifically, the pre-trained models Inception-UNet [86], Dense-UNet [61] and Efficient-UNet [191] were compared using individual, shared with multi-joint shape priors (shared + $\mathcal{L}_{\text{MJSP}}$), and DSL with multi-scale contrastive regularization and multi-joint shape priors (DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$) schemes. For the shared and DSL strategies, we only retained the best approach observed within each during Att-UNet experiments (i.e., shared + $\mathcal{L}_{\text{MJSP}}$ and DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$, Section 7.4.1). Finally, the transfer scheme was discarded in this experimental setup as networks were already partially pre-trained on the ImageNet

database [242].

Following the same design as in Chapter 5, the compared Inception and Dense-UNet architectures referred to UNet models with encoder respectively replaced by either an `InceptionV3` [86] or a `DenseNet121` [61] classifier network pre-trained on a natural image classification task (Table 7.1). Similarly to Efficient-UNet, the decoder components of both Inception-UNet and Dense-UNet were designed to be symmetrical from their respective encoder branches. Consequently, their decoders were extended from original UNet design by adding convolutional filters and more features, as well as Inception modules [86] and dense blocks [61] respectively. In addition, as for Efficient-UNet, spatial attention gates were incorporated to the skip connections of both Inception-UNet and Dense-UNet pre-trained architectures (Table 7.1).

7.3.2 Implementation details

Network	Pre-trained Encoder	AG	Batch Size	#Epochs	Learning Rate	#Parameters		
						Individual	Shared	DSL
Auto-encoder	–	–	24	8	1e−4	–	7.9M	7.9M
Att-UNet	–	✓	18	6	5e−4	3×8.7M	8.7M	8.7M
Inception-UNet	`InceptionV3`	✓	18	6	5e−4	3×48.1M	48.1M	48.3M
Dense-UNet	`DenseNet121`	✓	12	4	1e−4	3×23.3M	23.3M	23.6M
Efficient-UNet	`EfficientNetB3`	✓	12	4	5e−4	3×14.6M	14.6M	14.8M

Table 7.1 – Summary of the networks employed during experiments and their corresponding architecture design, including: pre-trained encoder [61], [86], [191], attention gate (AG) [42] and number of trainable parameters in individual, shared and DSL learning schemes; along with their corresponding training hyper-parameter values: batch size, number of epochs and learning rate.

Each of the networks employed through the experiments was characterized by specific implementation and architecture designs (Table 7.1). As previously indicated, all networks integrated attention gates with the exception of the auto-encoder due to the lack of skip connections. Moreover, `ReLU` and `SiLU` non-linear activation functions were implemented, in accordance with the original design of the employed models [42], [61], [86], [191]. As in Chapter 6, all networks were optimized using the Adam optimizer with distinct batch size, number of epochs and learning rate for each (Table 7.1), and these hyper-parameters values remained fixed across all multi-task, multi-domain segmentation strategies (individual, transfer, shared, and DSL). In shared and DSL schemes, the image batch was equally split between each dataset to prevent domain-bias during optimization.

The number of scales employed in the multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$ remained fixed across the networks with $|\mathcal{S}| = 9$. Meanwhile, the single-scale contrastive constraint $\mathcal{L}_{\text{SSC}}$ only involved the 5^{th} spatial scale corresponding to the network bottleneck (i.e., encoder output). Additionally, model complexity (i.e., number of trainable parameters) varied across architectures and learning schemes with a maximum of $3 \times 48.1\text{M}$ (millions) parameters for individual Inception-UNet (Table 7.1). Most notably, domain-specific weights represented at maximum 3.0% of the total of trainable parameters and DSL frameworks were highly compact compared to individual schemes which required $K = 3$ times more parameters.

Finally, we explored various values for the hyper-parameters of the multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$ and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$, and found $\tau = 0.1$, $\lambda_1 = 0.05$, and $\lambda_3 = 0.1$ to be optimal. The optimal value of λ_2 varied between architecture with $\lambda_2 = 0.5$ for Att-UNet and Inception-UNet whereas $\lambda_2 = 0.05$ for Dense-UNet and Efficient-UNet. For its part, the single scale contrastive regularization $\mathcal{L}_{\text{SSC}}$ was weighted by an hyper-parameter set to 0.1. As mentioned in Chapter 6, all networks were trained on 2D slices with substantial data augmentation, and we employed the same post-processing based on the selection of largest connected set to which we applied morphological closing.

7.3.3 Quantitative and qualitative assessments

The quantitative assessment of the generated segmentation relied on the metrics described in Chapter 6 (Dice, sensitivity, specificity, MSSD, ASSD, and RAVD) and experiments followed the same leave-one-out evaluation design within multiple datasets.

Similarly to Chapter 5, the limited amount of 3D examinations forced us to perform the statistical analysis between methods on the 2D MR images. To compare the multi-task, multi-domain strategies, we concatenated the 2D scores obtained on each dataset to create a unique distribution per metric. Specifically, we employed the Kolmogorov-Smirnov non-parametric test [261], [262] using Dice, sensitivity, and specificity scores obtained from the 2649 ankle, 3041 knee, and 3682 shoulder 2D slices which corresponded to the 45 MR image volumes in the test sets. Nevertheless, to avoid distorting the scores distributions, we retained only the scores obtained from the 1294 ankle, 2283 knee, and 3357 shoulder 2D images with at least one anatomical structure of interest. The non-normality of the 2D results distributions was preliminary verified using D'Agostino and Pearson normality test [245], [246]. Moreover, due to the skew of the non-normal distributions, we reported their mean and the distances from the mean to the upper and lower bound of the 68%

confidence interval, which corresponds to the 16 and 84 percentiles, as in [228]. Since transfer models ($\text{transfer}_{\text{Ankle}}$, $\text{transfer}_{\text{Knee}}$, and $\text{transfer}_{\text{Shoulder}}$) were not tested on their original domain, we used the 2D scores obtained in the individual scheme as substitute. For each backbone architecture (Att-UNet, Inception-UNet, Dense-UNet, and Efficient-UNet) we evaluated the statistical significance of the performance obtained by our methodology based on DSL with multi-scale contrastive regularization and multi-joint shape priors (DSL+$\mathcal{L}_{\text{MSC}}$+$\mathcal{L}_{\text{MJSP}}$) compared to other multi-task, multi-domain strategies and reported the results in Table 7.4.

Finally, we performed visual comparison of predicted segmentation masks at two levels. First, we extend the visualization conducted in Chapter 6 to assess the benefits in segmentation quality of the proposed multi-scale contrastive regularization ($\mathcal{L}_{\text{MSC}}$) along with multi-joint shape priors ($\mathcal{L}_{\text{MJSP}}$) using Att-UNet as backbone architecture in shared and DSL schemes. Second, we compared the segmentation obtained by the proposed Efficient-UNet pre-trained architecture in individual, shared + $\mathcal{L}_{\text{MJSP}}$, and DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ optimization schemes. We also provide attention maps computed by multi-domain attention gates to assess the interpretability of the proposed multi-task, multi-domain deep learning architectures (Att-UNet, Inception-UNet, Dense-UNet, and Efficient-UNet in DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ learning scheme). Specifically, we visualized the attention maps extracted by the spatial attention gate with highest resolution, which were up-sampled to original image resolution (i.e., 256×256) for Inception-UNet, Dense-UNet, and Efficient-UNet models.

7.3.4 Multi-joint ranking system

To simultaneously compare the performance of each segmentation strategy across multiple metrics and datasets, we defined a multi-joint ranking system inspired by the one proposed in Section 5.3.3. In particular, we individually compared the methods on ankle, knee, and shoulder datasets, and to assess the best multi-task, multi-domain method, we used a multi-joint ranking defined as the average score over all three joints. As opposed to the rankings obtained in Part II which were based on global-class metrics, the rankings of the present part relied on multi-class bone scores. Moreover, since transfer models ($\text{transfer}_{\text{Ankle}}$, $\text{transfer}_{\text{Knee}}$, and $\text{transfer}_{\text{Shoulder}}$) were not tested on their original domain, we used the scores obtained in the individual scheme as substitute. Finally, to assess the robustness of the multi-joint ranking system, we analyzed the effect of the modification of the threshold values (each resulting in a different ranking system) as in 5.

7.3.5 Assessment of learned shared representations

To assess the benefits of the proposed multi-scale contrastive regularization on the internal features of multi-domain neural networks, we compared the shared representations learned by Att-UNet and the multi-joint auto-encoder in shared, DSL, and DSL + $\mathcal{L}_{\text{MSC}}$ schemes. First, we computed the multi-scale embeddings $z_{s_1}, ..., z_{s_9}$ of Att-UNet (respectively $\tilde{z}_{s_1}, ..., \tilde{z}_{s_9}$ of the multi-joint auto-encoder, Figure 7.2) using ankle, knee, and shoulder 2D MR images (respectively 2D segmentation masks) originating from the training and validation sets. The 2D segmentation masks consisting of solely background were discarded during the process. Then, we applied the dimensionality reduction procedure recommended in [87], to visualize the high dimensional feature vectors belonging to $\mathbb{R}^d$ with d ranging from 32 to 512 (Figure 7.2). For vector space dimension $d > 50$, we first employed principal component analysis to reduce the representations to 50 dimensional feature vectors. We ultimately used the t-SNE algorithm with perplexity and learning rate respectively set to 30 and 200, to embed the data into a 2D space (see Section 4.4.2 for additional details on the t-SNE algorithm).

Finally, to provide a quantitative validation of the multi-scale contrastive regularization, we computed and compared the mean inter- and intra-domain cosine similarity of Att-UNet representations learned in shared, DSL + $\mathcal{L}_{\text{SSC}}$, and DSL + $\mathcal{L}_{\text{MSC}}$ schemes. As evaluating the similarity measure of each possible data points pairs was too computationally expensive, we randomly selected 10^5 pairs within and between each domain, and reported their respective mean cosine similarity and standard deviation in Table 7.8.

7.4 Results

The proposed method based on Efficient-UNet with pre-trained encoder, DSL, multi-scale contrastive regularization, and multi-joint shape priors was evaluated on three pediatric imaging domains and segmentation tasks. In this section, we report the quantitative results (Section 7.4.1) and qualitative comparisons (Section 7.4.3) of the multi-task, multi-domain strategies with different backbone architectures.

7.4.1 Quantitative assessment

Assessment of the multi-task, multi-domain segmentation strategies using Att-UNet architecture as backbone demonstrated that the segmentation method based on DSL with

		Method	Dice ↑	Sens. ↑	Spec. ↑	MSSD ↓	ASSD ↓	RAVD ↓
Att-UNet	Ankle	Individual	88.2 ± 1.9	88.1 ± 5.4	$\underline{99.8 \pm 0.1}$	17.9 ± 10.8	1.9 ± 1.1	14.1 ± 4.6
		$\text{Transfer}_{\text{Knee}}$	89.5 ± 5.7	88.3 ± 6.0	$\mathbf{99.9 \pm 0.1}$	12.6 ± 10.2	1.6 ± 1.7	14.0 ± 10.9
		$\text{Transfer}_{\text{Shoulder}}$	89.3 ± 4.2	87.5 ± 6.5	$\mathbf{99.9 \pm 0.1}$	11.6 ± 5.0	1.3 ± 0.6	12.9 ± 8.6
		Shared	88.8 ± 2.5	87.6 ± 6.3	$\mathbf{99.9 \pm 0.1}$	13.4 ± 8.1	1.5 ± 0.8	12.5 ± 7.0
		Shared + $\mathcal{L}_{\text{MJSP}}$	89.6 ± 1.6	$\underline{90.6 \pm 5.3}$	$\underline{99.8 \pm 0.1}$	13.4 ± 4.2	1.3 ± 0.3	13.1 ± 4.9
		DSL	90.6 ± 2.3	88.5 ± 4.6	$\mathbf{99.9 \pm 0.1}$	11.0 ± 7.4	1.2 ± 0.8	10.9 ± 5.6
		DSL + $\mathcal{L}_{\text{MJSP}}$	90.9 ± 1.9	89.1 ± 4.6	$\mathbf{99.9 \pm 0.1}$	12.7 ± 9.2	1.3 ± 1.2	10.5 ± 4.4
		DSL + $\mathcal{L}_{\text{SSC}}$	90.6 ± 2.1	87.7 ± 4.9	$\mathbf{99.9 \pm 0.1}$	$\mathbf{9.0 \pm 3.0}$	$\underline{1.0 \pm 0.3}$	11.3 ± 4.7
		DSL + $\mathcal{L}_{\text{MSC}}$	$\underline{91.5 \pm 2.0}$	$\mathbf{90.7 \pm 4.7}$	$\mathbf{99.9 \pm 0.1}$	$\underline{9.7 \pm 3.7}$	$\underline{1.0 \pm 0.3}$	$\underline{9.6 \pm 4.2}$
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{91.8 \pm 1.8}$	$\mathbf{90.7 \pm 4.8}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{9.0 \pm 2.9}$	$\mathbf{0.9 \pm 0.3}$	$\mathbf{8.8 \pm 4.7}$
	Knee	Individual	91.1 ± 3.6	88.9 ± 5.5	$\mathbf{99.9 \pm 0.1}$	16.5 ± 12.1	1.6 ± 1.5	10.7 ± 6.1
		$\text{Transfer}_{\text{Ankle}}$	92.8 ± 2.9	91.1 ± 3.3	$\mathbf{99.9 \pm 0.1}$	12.4 ± 10.3	1.0 ± 0.9	7.6 ± 5.4
		$\text{Transfer}_{\text{Shoulder}}$	92.5 ± 2.4	90.7 ± 4.0	$\mathbf{99.9 \pm 0.1}$	13.1 ± 11.3	1.0 ± 0.8	7.8 ± 4.5
		Shared	91.7 ± 3.2	88.5 ± 4.8	$\mathbf{99.9 \pm 0.1}$	12.5 ± 9.0	1.4 ± 1.4	9.5 ± 6.0
		Shared + $\mathcal{L}_{\text{MJSP}}$	93.6 ± 1.8	91.9 ± 3.1	$\mathbf{99.9 \pm 0.1}$	$\underline{7.9 \pm 8.4}$	0.8 ± 0.9	6.4 ± 3.1
		DSL	93.3 ± 2.5	92.8 ± 3.5	$\mathbf{99.9 \pm 0.1}$	12.8 ± 12.1	1.1 ± 1.3	6.0 ± 4.0
		DSL + $\mathcal{L}_{\text{MJSP}}$	$\underline{93.8 \pm 2.5}$	$\underline{93.0 \pm 4.2}$	$\mathbf{99.9 \pm 0.1}$	9.4 ± 5.9	$\underline{0.7 \pm 0.4}$	6.8 ± 4.1
		DSL + $\mathcal{L}_{\text{SSC}}$	93.7 ± 1.6	92.8 ± 2.3	$\mathbf{99.9 \pm 0.1}$	11.3 ± 8.9	1.1 ± 1.2	6.4 ± 3.7
		DSL + $\mathcal{L}_{\text{MSC}}$	$\mathbf{94.3 \pm 2.0}$	$\mathbf{93.2 \pm 3.9}$	$\mathbf{99.9 \pm 0.1}$	8.9 ± 9.7	0.8 ± 0.7	$\mathbf{5.5 \pm 4.0}$
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{94.3 \pm 1.4}$	92.5 ± 2.5	$\mathbf{99.9 \pm 0.1}$	$\mathbf{5.6 \pm 2.1}$	$\mathbf{0.5 \pm 0.2}$	$\underline{5.9 \pm 3.0}$
	Shoulder	Individual	80.9 ± 10.1	77.7 ± 14.9	$\mathbf{99.9 \pm 0.1}$	26.9 ± 14.1	2.4 ± 1.8	15.2 ± 16.7
		$\text{Transfer}_{\text{Ankle}}$	82.6 ± 8.8	79.8 ± 12.5	$\mathbf{99.9 \pm 0.1}$	26.9 ± 17.2	2.2 ± 1.8	17.0 ± 10.1
		$\text{Transfer}_{\text{Knee}}$	$\underline{83.3 \pm 10.1}$	$\underline{80.5 \pm 12.5}$	$\mathbf{99.9 \pm 0.1}$	$\underline{24.0 \pm 14.3}$	$\underline{2.1 \pm 2.4}$	$\underline{13.7 \pm 13.2}$
		Shared	80.1 ± 9.6	76.6 ± 12.9	$\mathbf{99.9 \pm 0.1}$	28.1 ± 12.2	2.7 ± 1.8	18.3 ± 12.4
		Shared + $\mathcal{L}_{\text{MJSP}}$	80.7 ± 9.0	79.2 ± 12.3	$\mathbf{99.9 \pm 0.1}$	25.0 ± 15.6	2.3 ± 1.8	19.4 ± 12.0
		DSL	80.9 ± 7.3	77.6 ± 11.6	$\mathbf{99.9 \pm 0.1}$	34.4 ± 19.1	3.3 ± 2.3	19.4 ± 12.2
		DSL + $\mathcal{L}_{\text{MJSP}}$	81.4 ± 9.0	79.2 ± 14.3	$\mathbf{99.9 \pm 0.1}$	31.0 ± 18.5	2.5 ± 2.2	15.7 ± 12.2
		DSL + $\mathcal{L}_{\text{SSC}}$	81.3 ± 9.1	78.1 ± 12.2	$\mathbf{99.9 \pm 0.1}$	25.4 ± 13.7	2.7 ± 2.7	17.1 ± 13.3
		DSL + $\mathcal{L}_{\text{MSC}}$	82.1 ± 8.0	79.8 ± 9.1	$\mathbf{99.9 \pm 0.1}$	27.5 ± 11.6	2.7 ± 1.8	15.6 ± 13.1
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{84.9 \pm 6.3}$	$\mathbf{82.9 \pm 9.2}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{17.6 \pm 8.0}$	$\mathbf{1.5 \pm 1.1}$	$\mathbf{13.5 \pm 10.5}$

Table 7.2 – Leave-one-out quantitative assessment of Att-UNet [42] using individual, transfer, shared, and DSL strategies employed with single-scale contrastive regularization $\mathcal{L}_{\text{SSC}}$, multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$, and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ on ankle, knee, and shoulder datasets. Metrics include Dice (%), sensitivity (%), specificity (%), MSSD (mm), ASSD (mm), and RAVD (%). Mean scores and standard deviations reported in bold and underlined respectively correspond to the first and second best results obtained for each dataset.

multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$ and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ achieved the best results on all metrics, except for sensitivity (0.7% lower than the best) and RAVD (0.4% higher than the best) on the knee dataset (Table 7.2). For ankle examinations, the method outperformed other approaches in Dice (+0.3%), MSSD (−0.7 mm), ASSD (−0.1 mm) and RAVD (−0.8%), while reaching sensitivity performance (90.7%) comparable to DSL+$\mathcal{L}_{\text{MSC}}$ strategy. With respect to the scores obtained for knee bone segmentation, our approach improved MSSD (−2.3 mm) and ASSD (−0.2 mm), while achieving same Dice

		Method	Dice ↑	Sens. ↑	Spec. ↑	MSSD ↓	ASSD ↓	RAVD ↓
Inception-UNet	Ankle	Individual	$\underline{91.4 \pm 2.6}$	$\mathbf{92.2 \pm 3.3}$	$\mathbf{99.9 \pm 0.1}$	$\underline{8.5 \pm 4.2}$	$\underline{0.9 \pm 0.4}$	$\underline{9.7 \pm 5.9}$
		Shared + $\mathcal{L}_{\text{MJSP}}$	91.3 ± 2.2	$\underline{91.6 \pm 5.1}$	$\mathbf{99.9 \pm 0.1}$	9.8 ± 4.1	1.0 ± 0.3	10.0 ± 5.0
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{93.2 \pm 1.5}$	$\mathbf{92.2 \pm 3.8}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{6.5 \pm 2.0}$	$\mathbf{0.7 \pm 0.2}$	$\mathbf{7.4 \pm 3.4}$
	Knee	Individual	93.9 ± 2.2	92.1 ± 3.7	$\mathbf{99.9 \pm 0.1}$	$\mathbf{5.5 \pm 2.6}$	$\mathbf{0.5 \pm 0.2}$	6.9 ± 4.5
		Shared + $\mathcal{L}_{\text{MJSP}}$	$\underline{94.2 \pm 2.1}$	$\underline{93.0 \pm 3.7}$	$\mathbf{99.9 \pm 0.1}$	6.4 ± 3.0	$\mathbf{0.5 \pm 0.2}$	$\underline{6.0 \pm 3.4}$
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{94.5 \pm 1.1}$	$\mathbf{93.6 \pm 2.0}$	$\mathbf{99.9 \pm 0.1}$	$\underline{6.3 \pm 2.2}$	$\mathbf{0.5 \pm 0.2}$	$\mathbf{5.2 \pm 2.7}$
	Shoulder	Individual	82.8 ± 7.3	79.6 ± 9.3	$\mathbf{99.9 \pm 0.1}$	21.8 ± 10.9	2.1 ± 1.5	$\underline{15.9 \pm 10.7}$
		Shared + $\mathcal{L}_{\text{MJSP}}$	$\underline{83.1 \pm 5.8}$	$\mathbf{81.2 \pm 10.6}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{20.0 \pm 11.2}$	$\underline{1.7 \pm 1.5}$	16.4 ± 10.5
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{84.5 \pm 5.8}$	$\underline{80.5 \pm 9.2}$	$\mathbf{99.9 \pm 0.1}$	$\underline{20.2 \pm 6.5}$	$\mathbf{1.6 \pm 0.7}$	$\mathbf{14.7 \pm 9.7}$
Dense-UNet	Ankle	Individual	$\underline{92.4 \pm 1.7}$	91.4 ± 4.7	$\mathbf{99.9 \pm 0.1}$	7.4 ± 2.4	$\underline{0.8 \pm 0.2}$	$\underline{7.9 \pm 4.5}$
		Shared + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{93.4 \pm 1.5}$	$\mathbf{92.8 \pm 4.4}$	$\mathbf{99.9 \pm 0.1}$	$\underline{6.9 \pm 2.1}$	$\mathbf{0.7 \pm 0.2}$	$\mathbf{6.6 \pm 4.6}$
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{93.4 \pm 1.3}$	$\underline{92.5 \pm 4.0}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{6.4 \pm 1.7}$	$\mathbf{0.7 \pm 0.2}$	$\mathbf{6.6 \pm 4.1}$
	Knee	Individual	$\underline{94.3 \pm 1.3}$	92.6 ± 2.5	$\mathbf{99.9 \pm 0.1}$	5.2 ± 2.2	$\mathbf{0.5 \pm 0.1}$	5.6 ± 2.9
		Shared + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{95.1 \pm 1.6}$	$\mathbf{94.9 \pm 2.4}$	$\mathbf{99.9 \pm 0.1}$	$\underline{4.6 \pm 1.8}$	$\mathbf{0.5 \pm 0.2}$	$\mathbf{4.5 \pm 3.2}$
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{95.1 \pm 1.2}$	$\underline{93.7 \pm 2.5}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{4.5 \pm 1.4}$	$\mathbf{0.5 \pm 0.1}$	$\underline{4.7 \pm 2.2}$
	Shoulder	Individual	82.5 ± 9.2	79.5 ± 12.4	$\mathbf{99.9 \pm 0.1}$	22.1 ± 11.6	1.9 ± 1.5	16.2 ± 14.4
		Shared + $\mathcal{L}_{\text{MJSP}}$	$\underline{84.9 \pm 4.8}$	$\underline{85.2 \pm 8.1}$	$\mathbf{99.9 \pm 0.1}$	$\underline{20.2 \pm 13.0}$	$\underline{1.4 \pm 1.0}$	$\underline{14.9 \pm 8.2}$
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{86.6 \pm 4.3}$	$\mathbf{87.0 \pm 5.3}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{16.0 \pm 6.5}$	$\mathbf{1.2 \pm 0.6}$	$\mathbf{11.5 \pm 6.5}$
Efficient-UNet	Ankle	Individual	$\underline{92.3 \pm 1.5}$	$\underline{92.0 \pm 3.8}$	$\mathbf{99.9 \pm 0.1}$	7.0 ± 2.1	$\underline{0.8 \pm 0.2}$	8.2 ± 4.1
		Shared + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{93.8 \pm 0.9}$	$\mathbf{93.5 \pm 2.8}$	$\mathbf{99.9 \pm 0.1}$	$\underline{6.5 \pm 1.6}$	$\mathbf{0.6 \pm 0.1}$	$\mathbf{5.9 \pm 2.2}$
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{93.8 \pm 1.3}$	$\mathbf{93.5 \pm 4.0}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{5.6 \pm 1.8}$	$\mathbf{0.6 \pm 0.2}$	$\underline{6.9 \pm 3.7}$
	Knee	Individual	94.1 ± 1.3	93.0 ± 2.9	$\mathbf{99.9 \pm 0.1}$	$\underline{4.7 \pm 1.2}$	$\underline{0.5 \pm 0.1}$	5.7 ± 2.5
		Shared + $\mathcal{L}_{\text{MJSP}}$	$\underline{95.0 \pm 1.2}$	$\underline{94.3 \pm 2.4}$	$\mathbf{99.9 \pm 0.1}$	4.8 ± 1.7	$\underline{0.5 \pm 0.2}$	$\underline{4.1 \pm 2.3}$
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{95.4 \pm 1.1}$	$\mathbf{95.0 \pm 2.0}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{4.2 \pm 1.3}$	$\mathbf{0.4 \pm 0.1}$	$\mathbf{3.8 \pm 1.6}$
	Shoulder	Individual	$\underline{87.7 \pm 4.0}$	86.8 ± 5.7	$\mathbf{99.9 \pm 0.1}$	16.0 ± 5.4	$\underline{1.0 \pm 0.5}$	$\underline{8.4 \pm 6.1}$
		Shared + $\mathcal{L}_{\text{MJSP}}$	86.9 ± 4.1	$\mathbf{89.0 \pm 4.8}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{14.3 \pm 5.1}$	$\mathbf{0.9 \pm 0.3}$	10.7 ± 8.2
		DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{87.9 \pm 3.8}$	$\underline{87.4 \pm 4.8}$	$\mathbf{99.9 \pm 0.1}$	$\underline{15.6 \pm 5.5}$	$\underline{1.0 \pm 0.5}$	$\mathbf{7.3 \pm 5.0}$

Table 7.3 – Leave-one-out quantitative assessment of the pre-trained architectures: Inception-UNet [86], Dense-UNet [61], and Efficient-UNet [191] on ankle, knee, and shoulder datasets. Individual, shared, and DSL strategies are employed with multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$ and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$. Metrics include Dice (%), sensitivity (%), specificity (%), MSSD (mm), ASSD (mm), and RAVD (%). Mean scores and standard deviations reported in bold and underlined respectively correspond to the first and second best results obtained for each dataset and each backbone.

results (94.3%) as DSL+$\mathcal{L}_{\text{MSC}}$ scheme. Additionally, for the shoulder dataset, our method outperformed other approaches in Dice (+1.6%), sensitivity (+2.4%), MSSD (−6.4 mm), ASSD (−0.6 mm), and RAVD (−0.2%). All methods achieved excellent specificity scores on all datasets (> 99.8%, Table 7.2). Moreover, the statistical analysis performed on 2D slices using Dice, sensitivity and specificity metrics indicated that the proposed method (DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$) produced significant improvements in segmentation performance (p-values < 0.01, Table 7.4). The 2D results also confirmed the overall performance improvements produced by our approach on Dice (+2.1%) and sensitivity (+0.8%) scores.

	Method	Dice 2D ↑	p-value	Sens. 2D ↑	p-value	Spec. 2D ↑	p-value
Att-UNet	Individual	$84.1^{+13.9}_{-14.6}$	$<\mathbf{1\times10^{-6}}$	$84.8^{+13.7}_{-14.2}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.1}$	$<\mathbf{1\times10^{-6}}$
	$\text{Transfer}_{\text{Ankle}}$	$85.1^{+13.1}_{-12.5}$	$<\mathbf{1\times10^{-6}}$	$85.8^{+12.5}_{-13.5}$	$<\mathbf{1\times10^{-6}}$	$99.8^{+0.2}_{-0.1}$	$<\mathbf{1\times10^{-6}}$
	$\text{Transfer}_{\text{Knee}}$	$84.7^{+13.3}_{-13.8}$	$<\mathbf{1\times10^{-6}}$	$85.4^{+13.3}_{-13.7}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	$\text{Transfer}_{\text{Shoulder}}$	$84.8^{+13.4}_{-15.4}$	$<\mathbf{1\times10^{-6}}$	$85.0^{+13.6}_{-15.7}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	Shared	$83.6^{+14.3}_{-15.0}$	$<\mathbf{1\times10^{-6}}$	$83.3^{+15.1}_{-15.9}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	Shared + $\mathcal{L}_{\text{MJSP}}$	$85.4^{+12.8}_{-12.7}$	$<\mathbf{1\times10^{-6}}$	$87.5^{+11.4}_{-13.8}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	DSL	$84.2^{+13.9}_{-16.8}$	$<\mathbf{1\times10^{-6}}$	$84.3^{+14.4}_{-16.6}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	DSL + $\mathcal{L}_{\text{MJSP}}$	$84.8^{+13.6}_{-16.2}$	$<\mathbf{1\times10^{-6}}$	$86.6^{+12.3}_{-14.5}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	DSL + $\mathcal{L}_{\text{SSC}}$	$85.1^{+13.1}_{-15.2}$	$<\mathbf{1\times10^{-6}}$	$84.9^{+13.6}_{-15.9}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	DSL + $\mathcal{L}_{\text{MSC}}$	$86.1^{+12.2}_{-12.8}$	$<\mathbf{1\times10^{-6}}$	$86.2^{+12.3}_{-13.4}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{88.2^{+10.2}_{-10.9}}$	–	$\mathbf{88.3^{+10.4}_{-12.9}}$	–	$\mathbf{99.9^{+0.1}_{-0.1}}$	–
Incept.	Individual	$86.4^{+11.8}_{-12.1}$	$<\mathbf{1\times10^{-6}}$	$86.4^{+12.1}_{-12.2}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	Shared + $\mathcal{L}_{\text{MJSP}}$	$86.5^{+11.8}_{-11.8}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{87.8^{+10.9}_{-13.2}}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{88.2^{+10.3}_{-11.0}}$	–	$87.7^{+11.0}_{-12.6}$	–	$\mathbf{99.9^{+0.1}_{-0.1}}$	–
Dense	Individual	$87.7^{+10.6}_{-9.8}$	$<\mathbf{1\times10^{-6}}$	$87.4^{+11.1}_{-12.2}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	Shared + $\mathcal{L}_{\text{MJSP}}$	$89.0^{+9.4}_{-8.9}$	$\mathbf{2.3\times10^{-4}}$	$\mathbf{90.7^{+8.3}_{-8.6}}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{89.5^{+8.8}_{-6.9}}$	–	$89.8^{+8.9}_{-8.4}$	–	$\mathbf{99.9^{+0.1}_{-0.1}}$	–
Efficient	Individual	$89.6^{+8.7}_{-6.0}$	$<\mathbf{1\times10^{-6}}$	$90.0^{+8.5}_{-7.9}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	1.2×10^{-1}
	Shared + $\mathcal{L}_{\text{MJSP}}$	$90.0^{+8.3}_{-6.5}$	$\mathbf{3.1\times10^{-5}}$	$\mathbf{91.7^{+6.9}_{-5.3}}$	$<\mathbf{1\times10^{-6}}$	$\mathbf{99.9^{+0.1}_{-0.1}}$	$<\mathbf{1\times10^{-6}}$
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{90.1^{+8.4}_{-6.9}}$	–	$90.6^{+8.2}_{-8.7}$	–	$\mathbf{99.9^{+0.1}_{-0.1}}$	–

Table 7.4 – Statistical analysis between the proposed methods using the four backbone architectures: Att-UNet [42], Inception-UNet [86], Dense-UNet [61], and Efficient-UNet [191]. Multi-task, multi-domain strategies include: individual, transfer, shared, and DSL employed with single-scale contrastive regularization $\mathcal{L}_{\text{SSC}}$, multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$, and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$. Statistical analysis performed through Kolmogorov-Smirnov non-parametric test using Dice (%), sensitivity (%), and specificity (%) computed on 2D slices from ankle, knee and shoulder datasets. Bold p-values (< 0.01) highlight statistically significant results for each metric and each backbone, while best 2D results are reported in bold. Mean 2D scores and the distances from the mean to the upper and lower bound of the 68% confidence interval are reported.

We then evaluated the performance of the backbone architectures with an encoder pre-trained on ImageNet using individual, shared + $\mathcal{L}_{\text{MJSP}}$, and DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ learning schemes (Table 7.3). Results obtained with Inception-UNet, Dense-UNet, and Efficient-UNet models further illustrated the benefits of the proposed learning scheme based on DSL, multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$, and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$. In Inception-UNet experiments, the DSL+$\mathcal{L}_{\text{MSC}}$+$\mathcal{L}_{\text{MJSP}}$ scheme ranked best in all metrics and in all datasets except for knee MSSD (0.8 mm higher than the best), shoulder sensitivity (0.7% lower than the best), and shoulder MSSD (0.2 mm higher than the best). Similarly, Dense-UNet backbone with DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ approach ranked best in all metrics and in all datasets except for ankle sensitivity (1.3% lower than the best), knee

sensitivity (1.2% lower than the best), and knee RAVD (0.2% higher than the best). For its part, the proposed Efficient-UNet with DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ achieved the best performance in all metrics and in all datasets except for ankle RAVD (1.0% higher than the best), shoulder sensitivity (1.6% lower than the best), and shoulder MSSD (1.3 mm higher than the best). Moreover, with respect to the 2D results, the proposed DSL+$\mathcal{L}_{\text{MSC}}$+$\mathcal{L}_{\text{MJSP}}$ scheme consistently reached the best Dice performance while Shared+$\mathcal{L}_{\text{MJSP}}$ achieved the best sensitivity within each backbone (Table 7.4). The obtained p-values indicated that the proposed DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ produced statistically significant different results (p-values < 0.01), except compared with the individual scheme using the Efficient backbone on the sensitivity metric. In this particular case, the difference between the 2D scores distributions was not statistically significant. However, as DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ produced statistically significant improvements on the remaining 2D metrics, we considered the overall improvements to be statistically significant.

Finally, when comparing the four backbone architectures (Att-UNet, Inception-UNet, Dense-UNet, and Efficient-UNet) with fixed DSL+$\mathcal{L}_{\text{MSC}}$+$\mathcal{L}_{\text{MJSP}}$ learning scheme (Tables 7.2 and 7.3), we observed that the proposed Efficient-UNet DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ reached the best performance in all metrics and in all datasets except for ankle RAVD (0.3% higher than Dense-UNet DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$).

7.4.2 Multi-joint rankings

Efficient-UNet DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ ranked first in performance (Table 7.5) on the knee dataset (mean score of 77.8) and on the multi-joint segmentation task (mean score of 64.4), while achieving the second best performance on the ankle (mean score of 67.6) and shoulder (mean score of 47.8) datasets. The proposed framework was marginally outperformed by Shared + $\mathcal{L}_{\text{MJSP}}$ model on ankle (+0.3 mean score) and shoulder (+0.3 mean score) segmentation tasks. Individual Att-UNet ranked last on ankle (mean score of 36.7), knee (mean score of 47.4), shoulder (mean score of 30.3), and multi-joint (mean score of 38.2) datasets. It was observed in the experiments based on Att-UNet architecture that transfer learning between pediatric datasets consistently outperformed the standard approach. For their part, the results of shared and DSL schemes on both ankle and knee datasets indicated noticeable score improvements while the results on shoulder examinations were less evident. Additionally, the ranks achieved by the pre-trained architectures (Inception-UNet, Dense-UNet, and Efficient-UNet) further demonstrated that the proposed DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ learning scheme promoted better performance as compared

	Method	Ankle		Knee		Shoulder		Multi-joint	
		Score	Rank	Score	Rank	Score	Rank	Score	Rank
Att-UNet	Individual	36.7 ± 12.0	19	47.4 ± 22.3	19	30.3 ± 20.5	11	38.2 ± 20.1	20
	$\text{Transfer}_{\text{Ankle}}$	–	–	58.2 ± 19.0	16	27.1 ± 16.2	15	40.6 ± 20.6	18
	$\text{Transfer}_{\text{Knee}}$	48.3 ± 22.0	15	–	–	34.4 ± 21.8	7	43.4 ± 22.9	17
	$\text{Transfer}_{\text{Shoulder}}$	47.4 ± 17.7	16	56.9 ± 19.0	17	–	–	44.9 ± 22.0	15
	Shared	44.4 ± 15.7	18	51.0 ± 21.4	18	22.6 ± 19.3	18	39.4 ± 22.5	19
	Shared + $\mathcal{L}_{\text{MJSP}}$	46.9 ± 8.5	17	64.3 ± 15.1	13	29.0 ± 23.7	13	46.7 ± 22.2	13
	DSL	50.1 ± 16.2	12	62.3 ± 19.3	14	18.9 ± 14.4	19	43.8 ± 24.8	16
	DSL + $\mathcal{L}_{\text{MJSP}}$	49.7 ± 12.9	14	64.8 ± 14.9	12	27.5 ± 19.0	14	47.4 ± 22.0	12
	DSL + $\mathcal{L}_{\text{SSC}}$	50.0 ± 11.9	13	61.8 ± 17.1	15	25.8 ± 20.0	17	45.9 ± 22.4	14
	DSL + $\mathcal{L}_{\text{MSC}}$	54.8 ± 11.0	11	68.5 ± 16.9	10	26.2 ± 20.2	16	49.8 ± 24.1	11
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	57.2 ± 10.9	9	68.8 ± 11.6	8	37.8 ± 20.9	5	54.6 ± 19.8	7
Incept.	Individual	58.2 ± 11.0	8	67.6 ± 13.7	11	30.1 ± 21.9	12	52.0 ± 22.7	10
	Shared + $\mathcal{L}_{\text{MJSP}}$	55.8 ± 12.3	10	68.7 ± 13.0	9	32.0 ± 18.7	9	52.2 ± 21.3	9
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	63.7 ± 10.3	5	70.5 ± 9.9	5	31.7 ± 14.1	10	55.3 ± 20.5	6
Dense	Individual	61.2 ± 11.3	7	70.1 ± 9.2	7	32.1 ± 21.2	8	54.5 ± 22.0	8
	Shared + $\mathcal{L}_{\text{MJSP}}$	66.9 ± 12.2	3	75.8 ± 10.7	2	35.4 ± 18.0	6	59.3 ± 22.3	5
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	66.0 ± 11.0	4	73.9 ± 8.9	4	41.3 ± 17.6	4	60.4 ± 19.0	3
Efficient	Individual	61.3 ± 8.5	6	70.4 ± 8.5	6	46.4 ± 14.8	3	59.4 ± 14.8	4
	Shared + $\mathcal{L}_{\text{MJSP}}$	$\mathbf{67.9 \pm 7.5}$	**1**	75.5 ± 9.3	3	$\mathbf{48.1 \pm 14.3}$	**1**	63.8 ± 15.8	2
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	67.6 ± 10.5	2	$\mathbf{77.8 \pm 7.0}$	**1**	47.8 ± 16.7	2	$\mathbf{64.4 \pm 17.3}$	**1**

Table 7.5 – Multi-joint scores and rankings of the four backbone architectures: Att-UNet [42], Inception-UNet [86], Dense-UNet [61], and Efficient-UNet [191] on ankle, knee, and shoulder datasets. Multi-task, multi-domain strategies include: individual, transfer, shared, and DSL employed with single-scale contrastive regularization $\mathcal{L}_{\text{SSC}}$, multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$, and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$. Results encompass mean, standard deviation and associated rank. Methods were ranked according to their mean score. Best results are in bold.

to individual training. Finally, assessment of the robustness of the multi-joint ranking further confirmed the efficiency performance of the multi-task, multi-domain proposed model. Indeed, Efficient-UNet DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ ranked first whatever the selected threshold values (Table 7.6).

7.4.3 Qualitative assessment

Visual comparison of the multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$ and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$ completed the evaluation performed in Chapter 6 and provided visual evidence of gradual improvements in segmentation quality for both shared and DSL Att-UNet models (Figure 7.3). As noted in Chapter 6, we observed that shape priors enforce globally more consistent delineations for all targeted anatomical structures. Additionally, **the contrastive regularization encouraged more precise bone extraction in all domains (A_{12}, K_3, and S_{11}) through more robust shared representations with**

	Method	Multi-joint rankings									
		Dice_{75}	Dice_{85}	Sens_{75}	Sens_{85}	MSSD_{20}	MSSD_{40}	ASSD_{3}	ASSD_{5}	RAVD_{5}	RAVD_{15}
Att-UNet	Individual	20	20	20	20	20	20	20	20	20	20
	$\text{Transfer}_{\text{Ankle}}$	18	18	18	18	18	18	18	18	18	18
	$\text{Transfer}_{\text{Knee}}$	17	17	17	17	17	17	17	17	17	17
	$\text{Transfer}_{\text{Shoulder}}$	15	15	15	15	15	15	15	15	15	15
	Shared	19	19	19	19	19	19	19	19	19	19
	Shared + $\mathcal{L}_{\text{MJSP}}$	13	13	13	13	13	13	13	13	**12**	13
	DSL	16	16	16	16	16	16	16	16	16	16
	DSL + $\mathcal{L}_{\text{MJSP}}$	12	12	12	12	12	12	12	12	**13**	12
	DSL + $\mathcal{L}_{\text{SSC}}$	14	14	14	14	14	14	14	14	14	14
	DSL + $\mathcal{L}_{\text{MSC}}$	11	11	11	11	11	11	11	11	11	11
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	7	7	7	7	**8**	7	**8**	7	7	7
Incept.	Individual	10	10	10	10	**9**	10	10	10	10	10
	Shared + $\mathcal{L}_{\text{MJSP}}$	9	9	9	9	**10**	9	9	9	9	9
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	6	6	6	6	6	6	6	6	6	6
Dense	Individual	8	8	8	8	**7**	8	**7**	8	8	8
	Shared + $\mathcal{L}_{\text{MJSP}}$	5	**4**	5	**4**	**4**	5	**4**	**4**	**4**	5
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	3	3	3	3	3	3	3	3	3	3
Efficient	Individual	4	**5**	4	**5**	**5**	4	**5**	**5**	**5**	4
	Shared + $\mathcal{L}_{\text{MJSP}}$	2	2	2	2	2	2	2	2	2	2
	DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$	1	1	1	1	1	1	1	1	1	1

Table 7.6 – Transformed multi-joint rankings of the four backbone architectures: Att-UNet [42], Inception-UNet [86], Dense-UNet [61], and Efficient-UNet [191] on ankle, knee and shoulder datasets. Multi-task, multi-domain strategies include: individual, transfer, shared, and DSL employed with single-scale contrastive regularization $\mathcal{L}_{\text{SSC}}$, multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$, and multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$. Rankings were computed using different threshold values: Dice = 75 or 85%, Sensitivity = 75 or 85%, MSSD = 20 or 40 mm, ASSD = 3 or 5 mm and RAVD = 5 or 15%. Modified ranks are in bold.

domain-specific clusters. Meanwhile, the proposed DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ approach fostered the benefits of both previous terms and generated smoother and more realistic bone delineations (A_{14}, K_5 and S_{11}).

We then visually compared the pre-trained Efficient-UNet models employed in individual, shared + $\mathcal{L}_{\text{MJSP}}$, and DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ learning strategies (Figure 7.4). First, the qualitative comparison demonstrated that models with pre-trained encoder benefited from transfer learning to achieve robust feature extraction and produce highly accurate delineations in the three considered anatomical regions (A_2, K_{11}, and S_8). However, we observed that individual models produced segmentation errors in several imaging examinations, for instance, by over-segmenting the femoral shape in knee joint (K_{11}) or under-segmenting the scapular bone in shoulder joint (S_3). Specifically, because the boundary between bone and ligament was not detected by the individual model, ligamentous tissues were erroneously classified as femur bone (K_{11}). Furthermore, the thin structure of scapu-

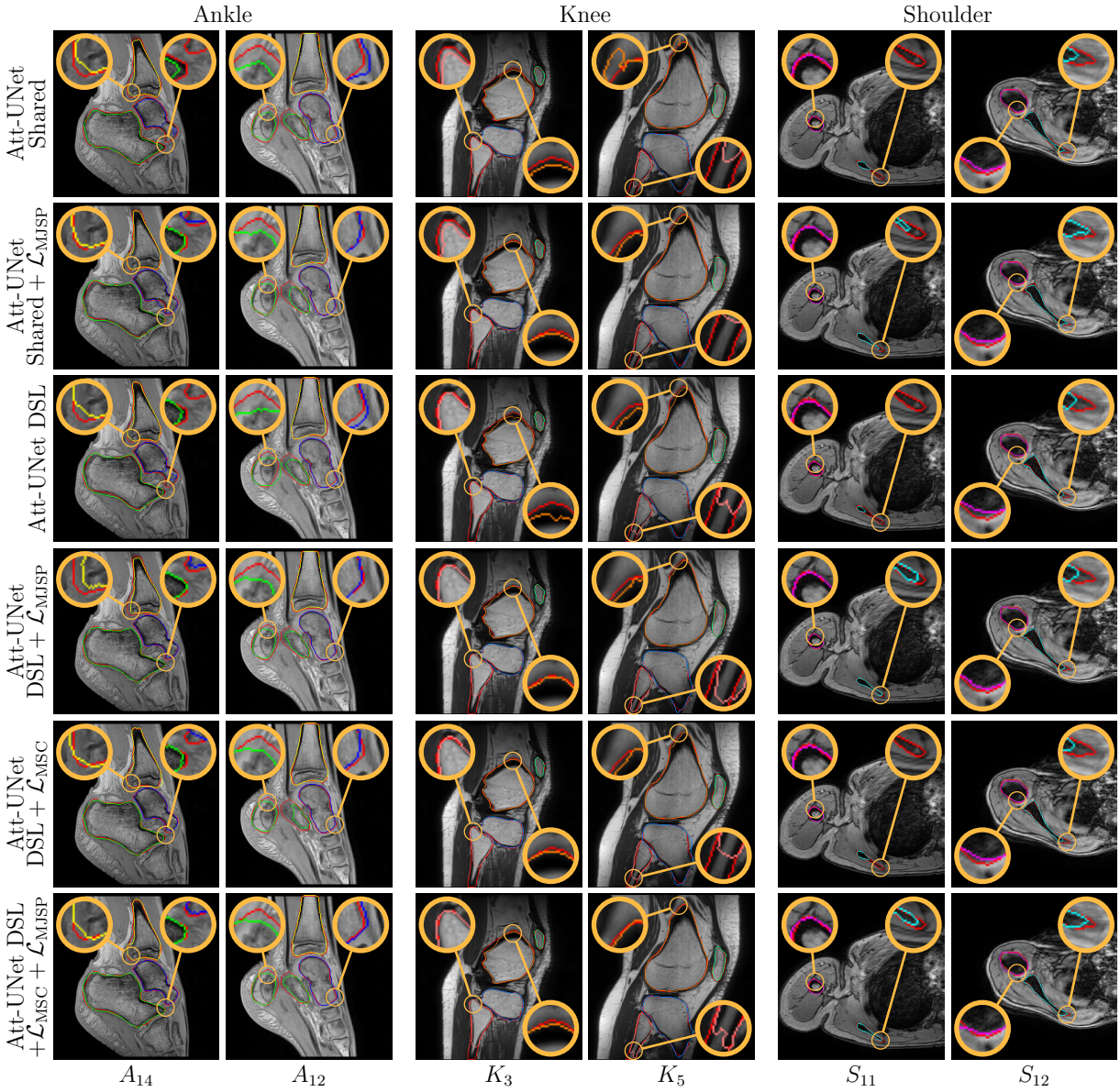


Figure 7.3 – **Visual comparison of the multi-scale contrastive regularization $\mathcal{L}_{\mathbf{MSC}}$ and multi-joint shape priors $\mathcal{L}_{\mathbf{MJSP}}$ using Att-UNet architecture.** Automatic segmentation of ankle, knee, and shoulder bones based on Att-UNet [42] employed in shared and DSL strategies. Ground truth delineations are in red (–) while predicted bones appear in green (–) for calcaneus, blue (–) for talus, yellow (–) for tibia (distal), orange (–) for femur (distal), pink (–) for fibula (proximal), light green (–) for patella, light blue (–) for tibia (proximal), magenta (–) for humerus, and cyan (–) for scapula.

lar bone led to its partial misclassification as background (S_3). Additionally, the calcaneus shape was also under-segmented due to intensity difference within the bone (A_{11}). While

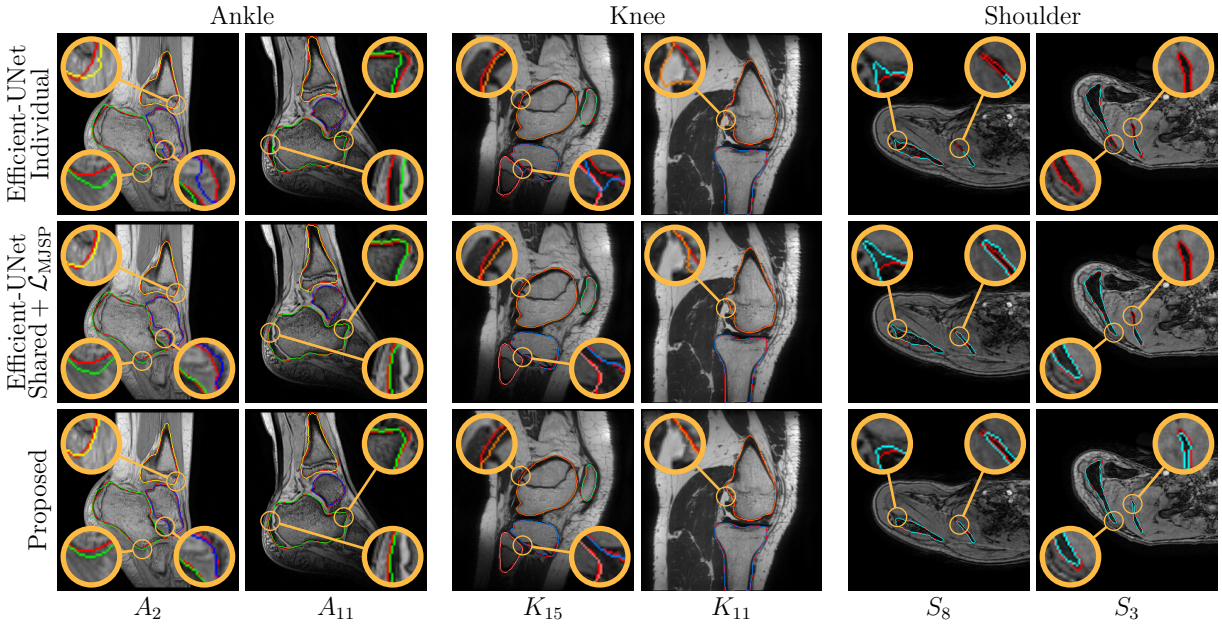


Figure 7.4 – **Visual comparison of the pre-trained Efficient-UNet models employed in individual, shared + $\mathcal{L}_{\text{MJSP}}$, and DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ strategies on ankle, knee, and shoulder bones.** Automatic segmentation of ankle, knee, and shoulder bones based on Efficient-UNet [191] employed in individual, shared + $\mathcal{L}_{\text{MJSP}}$, and DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ strategies. Ground truth delineations are in red (–) while predicted bones appear in green (–) for calcaneus, blue (–) for talus, yellow (–) for tibia (distal), orange (–) for femur (distal), pink (–) for fibula (proximal), light green (–) for patella, light blue (–) for tibia (proximal), magenta (–) for humerus, and cyan (–) for scapula.

the shared + $\mathcal{L}_{\text{MJSP}}$ model produced segmentation improvements over its individual counterparts, it was essential to employ the DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ model incorporating layer specialization along with multiple regularizers to learn robust shared representations and achieve precise bone shape predictions on unseen images (A_{11}, K_{15}, and S_8).

Finally, we provide visualization of the attention maps computed by the multi-domain attention gates of the Att-UNet, Inception-UNet, Dense-UNet, and Efficient-UNet architectures employed in DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ learning scheme (Figure 7.5). As indicated in Chapter 4, these attention maps were crucial in interpreting the inference process of deep neural networks. **This visualization confirmed that the segmentation models exploited the spatial and contextual information from the encoder branch to focus on the bone of interest in each anatomical joint.** Indeed, knee attention maps clearly equally highlighted each bone of interest (femur, fibula, patella, and tibia), and suppressed most of the irrelevant regions. In some cases, background elements were

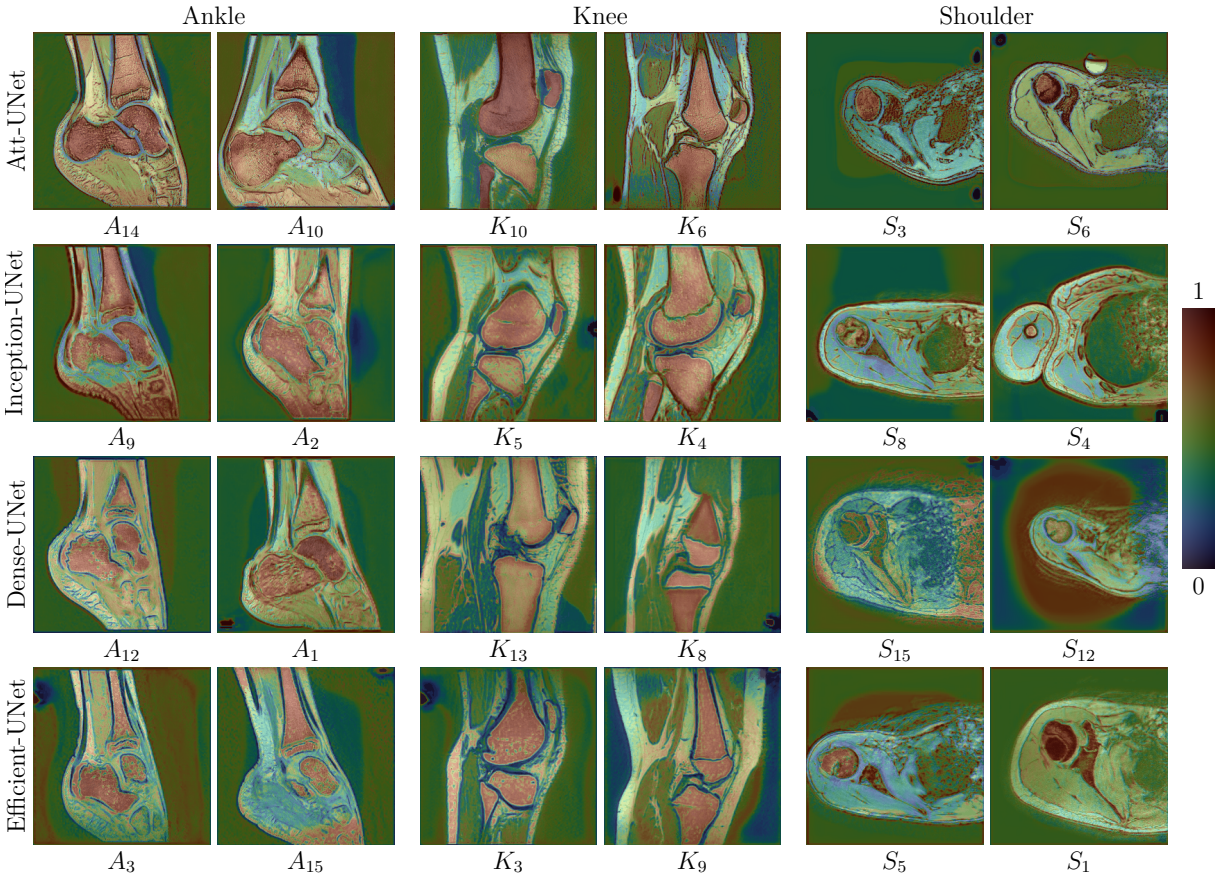


Figure 7.5 – **Visualization of the attention maps computed by the multi-domain attention gates using DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ learning scheme.** Architectures encompassed Att-UNet [42], Inception-UNet [86], Dense-UNet [61], and Efficient-UNet [191] employed on ankle, knee, and shoulder joint images. Pixel-wise coefficients ranging from 0 in blue to 1 in red indicated low to high attention.

also included (e.g., A_9 with Inception-UNet and S_{12} with Dense-UNet) and may help the inference process which remains difficult to interpret. We can note that attention maps computed on shoulder joint images highlighted the scapula less than the humerus bone. Meanwhile, ankle joint attention maps focused on the calcaneus, talus, and tibia bones, with some background structures also being highlighted. Finally, for each bone of interest, we observed a discontinuity in the attention coefficients at the bone borders (e.g., K_3 with Efficient-UNet), that allowed the network to effectively distinguish and extract their shape from the rest of the image.

7.5 Discussion

In this chapter, we developed and evaluated a novel multi-task, multi-domain deep segmentation framework with multi-scale contrastive regularization and multi-joint shape priors. To the best of our knowledge, the proposed multi-task, multi-domain segmentation method is the first illustration to optimize a single neural network over multiple pediatric musculoskeletal joints. Experiments performed on the ankle, knee, and shoulder joint imaging datasets demonstrated improved bone segmentation performance compared to individual, transfer, and shared learning schemes. The statistical analysis validated the significance of the results, while visual comparison of the predicted delineations further confirmed the enhancements in segmentation quality of the proposed framework. The proposed methodology could provide significant benefits to the management of pediatric imaging resources and can have a major impact for any deep learning-based medical image analysis framework.

7.5.1 Segmentation performance

From the extension of the Att-UNet experiments performed in Chapter 6, it appeared essential to employ both $\mathcal{L}_{\text{MJSP}}$ and $\mathcal{L}_{\text{MSC}}$ terms to benefit from the shared representation and layer specialization, and reach performance improvements over independent and transfer models on all datasets. This outcome was also supported by the results obtained on Inception-UNet, Dense-UNet, and Efficient-UNet models (Tables 7.3 and 7.5). It is also worth emphasizing that the multi-scale contrastive $\mathcal{L}_{\text{MSC}}$ regularization outperformed its single-scale $\mathcal{L}_{\text{SSC}}$ counterpart on all datasets (Tables 7.2 and 7.5), indicating that disentangling representations at each scale provided better generalization performance than focusing only on the features within the network's bottleneck. For instance, the ankle Dice score increased from 90.6% to 91.5%, while knee and shoulder Dice metrics improved by 0.7% and 0.8% respectively. Finally, we observed that, within Att-UNet models, the proposed DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ scheme achieved important improvements in MSSD and ASSD metrics (Table 7.2), indicating lower surface errors. Qualitative assessment (Figure 7.3) further confirmed this observation as compared methods were reported to partially segment the talus, fibular, and scapular shapes. In contrast, our method provided complete bone segmentation resulting in substantial surface metric (i.e., MSSD and ASSD) improvements.

When comparing the performance of the four employed backbone architectures in

the individual learning scheme (Tables 7.2, 7.3, and 7.5), we observed that Inception-UNet, Dense-UNet, and Efficient-UNet outperformed Att-UNet in all metrics and in all datasets. As also highlighted in Chapter 5, this clearly indicated that designing a segmentation model with a pre-trained encoder resulted in better initialization through features learned on ImageNet and better segmentation performance by mean of a more complex and deeper CNN architecture. Indeed, compared to the complexity of the Att-UNet model, the number of trainable parameters in Inception-UNet, Dense-UNet and Efficient-UNet corresponded to an increase by a factor of five, three, and two respectively (Table 7.1). However, to avoid over-fitting, it is also crucial to limit the number of trainable parameters, as models with too much capacity may learn the dataset and task too well. In practice, the optimal model capacity depends on the considered task and available imaging resources which are limited in the context of sparse pediatric datasets. In this sense, we observed step-wise performance improvements from Inception-UNet (48.3M parameters) to Dense-UNet (23.6M parameters) and ultimately Efficient-UNet (14.8M parameters) networks (Table 7.3). For instance, shoulder Dice score increased from 84.5% for Inception-UNet to 86.6% for Dense-UNet and ultimately to 87.9% using Efficient-UNet. Finally, the proposed multi-task, multi-domain approach also allowed us to reduce the number of learnable parameters by a factor of $K = 3$, and to consequently minimize over-fitting and improve generalizability. Meanwhile, the supplementary parameterization introduced by the domain-specific layers was considered marginal (i.e., less than 3.0%).

As demonstrated through our experiments, the proposed DSL$+\mathcal{L}_{\text{MSC}}+\mathcal{L}_{\text{MJSP}}$ learning scheme is architecture-independent, and thus can be effortlessly integrated into various existing CNN models and can improve the overall performance in all datasets. Indeed, **although the obtained Dice and ASSD performance gains can be considered limited, these improvements are consistent and robust (i.e., lower standard deviation)**. Most importantly, experiments performed on Inception-UNet, Dense-UNet, and Efficient-UNet illustrated that the multi-scale contrastive regularization can be computed from the internal representations of networks composed of distinct building blocks (i.e., Inception, dense, or MBConv blocks) and diverse feature transformation operation (i.e., classical, point-wise, depth-wise, or asymmetrical convolutions and `ReLU` or `SiLU` non-linearity functions).

As indicated by the high variance in shoulder results (Tables 7.2 and 7.3), **the shoulder dataset was more challenging to segment than those from ankle and knee joints, due to more complex bone shapes (i.e thin scapular blade), higher vari-**

	Method	Dice ↑	Sens. ↑	Spec. ↑	MSSD ↓	ASSD ↓	RAVD ↓
A	$\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ †	$\mathbf{94.1 \pm 1.1}$	$\mathbf{93.5 \pm 3.1}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{5.3 \pm 2.3}$	$\mathbf{0.6 \pm 0.2}$	$\mathbf{6.2 \pm 2.4}$
	Efficient-UNet DSL++	93.8 ± 1.3	$\mathbf{93.5 \pm 4.0}$	$\mathbf{99.9 \pm 0.1}$	5.6 ± 1.8	$\mathbf{0.6 \pm 0.2}$	6.9 ± 3.7
K	$\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$	–	–	–	–	–	–
	Efficient-UNet DSL++	$\mathbf{95.4 \pm 1.1}$	$\mathbf{95.0 \pm 2.0}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{4.2 \pm 1.3}$	$\mathbf{0.4 \pm 0.1}$	$\mathbf{3.8 \pm 1.6}$
S	$\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ †	$\mathbf{89.5 \pm 3.3}$	$\mathbf{89.3 \pm 4.0}$	$\mathbf{99.9 \pm 0.1}$	18.5 ± 16.9	1.2 ± 1.5	$\mathbf{6.1 \pm 3.3}$
	Efficient-UNet DSL++	87.9 ± 3.8	87.4 ± 4.8	$\mathbf{99.9 \pm 0.1}$	$\mathbf{15.6 \pm 5.5}$	$\mathbf{1.0 \pm 0.5}$	7.3 ± 5.0

Table 7.7 – Quantitative comparison of $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ framework proposed in Part II and Efficient-UNet DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ (DSL++) pipeline of Part III on ankle (A), knee (K), and shoulder (S) datasets. Metrics encompass Dice (%), sensitivity (%), specificity (%), MSSD (mm), ASSD (mm) and RAVD (%). Bold results correspond to the best performance for each dataset and for each metric. † indicates that results previously reported using global-class masks (Table 5.4) have been transformed into multi-class scores.

ability among pediatric patients (i.e., different age groups), and the presence of examinations with a higher level of noise due to patient movements during acquisition. A similar comment was noted in Chapter 5 with respect to the presence of two outliers examinations, one healthy and one pathological, in the shoulder dataset. Interestingly, the attention maps (Figure 7.5) could explain the lower performance for segmenting the scapular shape which appeared more challenging to detect than the humerus bone. Finally, compared to our previous experiments performed in Chapters 4 and 5, we incorporated three additional ankle pediatric examinations with a higher level of noise in the test sets which led to a marginal drop in performance for ankle bone segmentation. Nevertheless, we still observed that for the ankle joint segmentation, the DSL scheme outperformed the shared approach, which in turn outranked the individual scheme.

Finally, we compared the performance of the frameworks proposed in Parts II and III (respectively $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ and Efficient-UNet DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$) on ankle, knee, and shoulder datasets (Table 7.7). To obtain a fair comparison, we transformed the score of $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ previously reported in Table 5.4 into multi-class bone metrics. From this comparison, it appeared that both frameworks achieved similar results on all ankle metrics, whereas a larger difference in performance was reported on the shoulder dataset. Indeed, $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ reached better Dice (+1.6%), sensitivity (+1.9%), and RAVD (−1.2%) scores, while Efficient-UNet DSL + $\mathcal{L}_{\text{MSC}}$ + $\mathcal{L}_{\text{MJSP}}$ scored best in MSSD (−2.9 mm) and ASSD (−0.2 mm), with lower variance for both metrics. With respect to the knee dataset, only Efficient-UNet DSL+$\mathcal{L}_{\text{MSC}}$+$\mathcal{L}_{\text{MJSP}}$ could be evaluated, as access to the pediatric knee imaging database was granted solely to develop our multi-task, multi-domain framework. It should be emphasized that experiments performed in Part III included three additional

pediatric ankle examinations (two pathological and one healthy) which were not available at the time of the experiments of Part II. Moreover, the leave-one-out evaluation schemes differed between experiments, it was achieved independently for each dataset in Part II, whereas Part III employed a multi-domain leave-one-out scheme (see Sections 4.3.4 and 6.3.4). Although this comparison is not completely fair as all implementation details were not strictly identical, the following conclusion can be formulated.

Ultimately, as $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ leveraged a combination of shape priors and adversarial regularization, one could expect performance gain by integrating a discriminator network in the Efficient-UNet $\text{DSL} + \mathcal{L}_{\text{MSC}} + \mathcal{L}_{\text{MJSP}}$ multi-anatomy framework. However, this proved difficult due to the unstable simultaneous optimization of the segmentation and adversarial networks. But **most importantly, the multi-task, multi-domain Efficient-UNet $\text{DSL} + \mathcal{L}_{\text{MSC}} + \mathcal{L}_{\text{MJSP}}$ model simultaneously learned to segment all three datasets, whereas $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ was limited to segment one dataset at a time.**

7.5.2 Assessment of learned shared representations

Similarly to the experiments performed in Chapter 4, the present visualization of the shared representation provided an indirect analysis of the inference process of deep neural networks and a qualitative validation of the benefits of the additional multi-scale contrastive regularization on both intra-domain cohesion and inter-domain separation (Figure 7.6). In both Att-UNet and auto-encoder networks, the shared representation learned using shared and DSL schemes did not present margins between domains. More specifically, shared models presented mixed features with most discriminative domain disentanglement in the network bottleneck (s_5) which corresponded to the higher dimensional vector space ($d = 512$) allowing more robust differentiation between domains. On the contrary, the addition of the contrastive regularization led to distinctive domain-specific clusters at each scale of both networks. Hence, **the shared representations of our proposed neural networks were invariant to local variations and preserved the category of the input domain through the different scales of the models**. Moreover, the generalization capabilities of the networks were visually attested as validation data points were located inside their respective domain clusters.

The quantitative evaluation (Table 7.8) further supported the visualizations obtained through the t-SNE algorithm (Figure 7.6). Indeed, the shared Att-UNet representations presented inter-domain cosine similarity measures with high mean (> 0.58) and standard

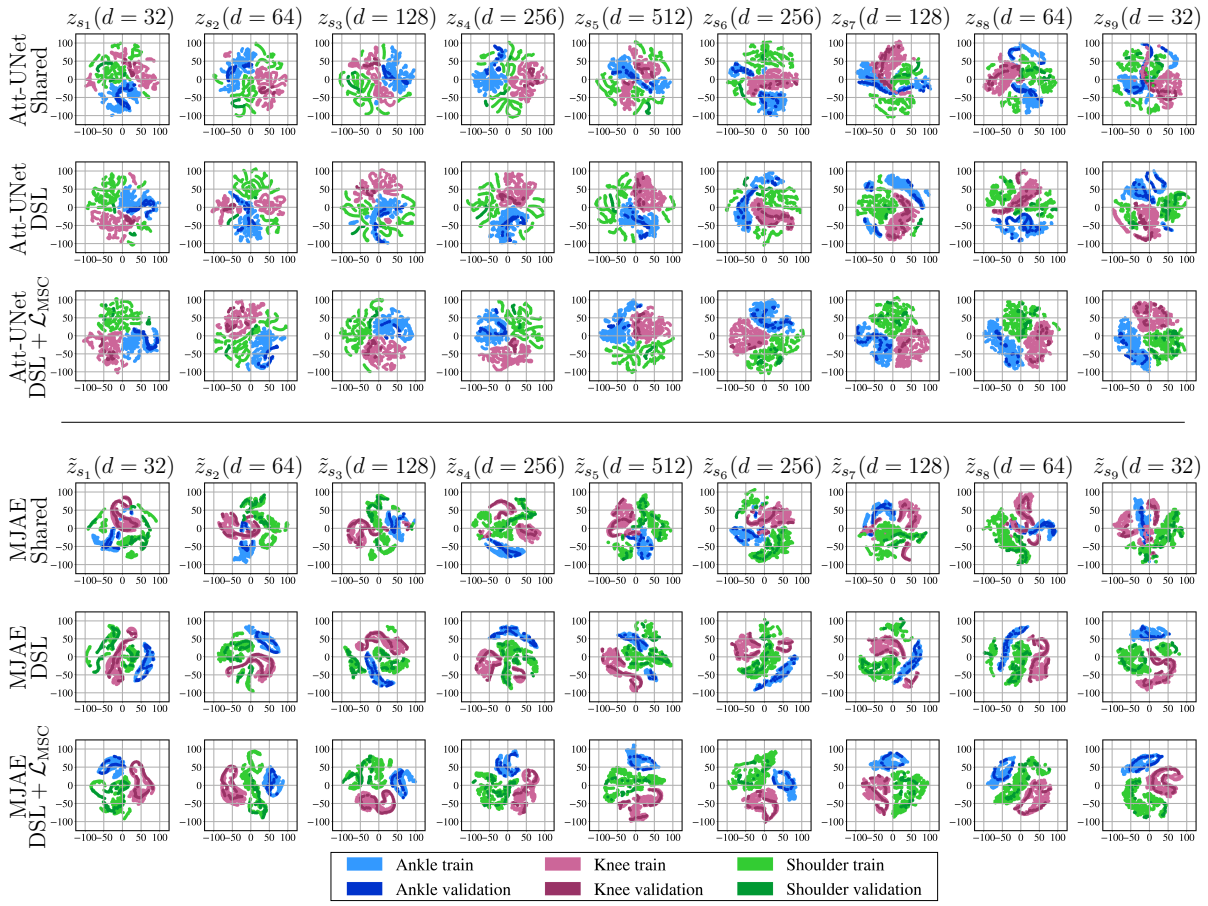


Figure 7.6 – **Visual comparison of the shared representations learned in shared, DSL and DSL + $\mathcal{L}_{\text{MSC}}$ learning schemes**. Architectures encompassed Att-UNet [42] and the multi-joint auto-encoder (MJAE). The multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$ promoted intra-domain cohesion and inter-domain margins in embedded spaces at each scale. This visualization was obtained using the t-SNE algorithm [87] in which each colored dot represented a 2D MR slice or segmentation mask from the training or validation set of the ankle, knee, or shoulder datasets.

deviation (> 0.06) suggesting entangled domain representations with low cohesion. Moreover, as previously mentioned, the network bottleneck corresponding to the representation z_{s_5} (Figure 7.6) presented better domain disentanglement due to higher dimensionality. Additionally, the multi-scale contrastive regularization expectedly led to, at each scale, an increase in intra-domain similarity (> 0.98) indicating more closely aligned representations from the same domain and a decrease in inter-domain similarity (< 0.47) reflecting more discriminative (i.e., orthogonal) representations between different domains. However, we observed that domain representations were less disentangled at scale s_9 (inter-domain

Cosine similarity			z_{s_1} ($d = 32$)			z_{s_5} ($d = 512$)			z_{s_9} ($d = 32$)		
			Ankle	Knee	Shoulder	Ankle	Knee	Shoulder	Ankle	Knee	Shoulder
Att-UNet	Shared	Ankle	0.84(0.16)	0.84(0.11)	0.83(0.10)	0.82(0.16)	0.58(0.20)	0.69(0.12)	0.86(0.14)	0.73(0.17)	0.86(0.09)
		Knee	0.84(0.11)	0.93(0.07)	0.93(0.06)	0.58(0.20)	0.88(0.12)	0.71(0.11)	0.73(0.17)	0.90(0.10)	0.83(0.12)
		Shoulder	0.83(0.10)	0.93(0.06)	0.97(0.03)	0.69(0.12)	0.71(0.11)	0.93(0.05)	0.86(0.09)	0.83(0.12)	0.95(0.05)
	DSL $+\mathcal{L}_{\text{SSC}}$	Ankle	0.84(0.17)	0.83(0.11)	0.84(0.08)	0.99(0.01)	0.25(0.02)	0.21(0.02)	0.90(0.11)	0.82(0.10)	0.83(0.08)
		Knee	0.83(0.11)	0.93(0.07)	0.93(0.05)	0.25(0.02)	0.99(0.01)	0.21(0.02)	0.82(0.10)	0.90(0.10)	0.85(0.08)
		Shoulder	0.84(0.08)	0.93(0.05)	0.98(0.02)	0.21(0.02)	0.21(0.02)	0.99(0.01)	0.83(0.08)	0.85(0.08)	0.97(0.04)
	DSL $+\mathcal{L}_{\text{MSC}}$	Ankle	0.99(0.01)	0.31(0.05)	0.29(0.04)	0.99(0.01)	0.29(0.02)	0.22(0.03)	0.98(0.02)	0.47(0.05)	0.39(0.03)
		Knee	0.31(0.05)	0.99(0.01)	0.34(0.03)	0.29(0.02)	0.99(0.01)	0.27(0.02)	0.47(0.05)	0.98(0.01)	0.47(0.05)
		Shoulder	0.29(0.04)	0.34(0.03)	0.99(0.01)	0.22(0.03)	0.27(0.02)	0.99(0.01)	0.39(0.03)	0.47(0.05)	0.99(0.01)

Table 7.8 – Quantitative analysis based on cosine similarity of the shared representations learned by Att-UNet in shared, DSL + $\mathcal{L}_{\text{SSC}}$, and DSL + $\mathcal{L}_{\text{MSC}}$ strategies using ankle, knee, and shoulder datasets. The included scales correspond to the encoder first layer (s_1), network bottleneck (s_5), and decoder last layer (s_9). Mean and standard deviation similarity measures are reported.

similarity greater than 0.39) than at scales s_1 and s_5 (inter-domain similarity lower than 0.34). Therefore, the effectiveness of contrastive learning to disentangle domain representations varies at each scale, as we observed a quantitative difference in the cosine similarity of the learned representations. Finally, we also assessed the representations learned with the single-scale contrastive regularization which only constrained the network bottleneck (i.e., encoder output or z_{s_5}). Compared with $\mathcal{L}_{\text{MSC}}$, only the representation associated with the 5^{th} scale was disentangled while z_{s_5} and z_{s_9} were not affected by the single-scale contrastive constraint. This further supported the necessity to employ a multi-scale contrastive $\mathcal{L}_{\text{MSC}}$ regularization to disentangle representations at each layer, as opposed to $\mathcal{L}_{\text{SSC}}$ term.

7.5.3 Benefits for clinical practice

As mentioned in Part I, current deep learning models are typically specific to anatomical region of interest and may suffer from the limited availability of imaging data, which is exacerbated in pediatric clinical workflows. **Our approach demonstrated that designing a collaborative framework incorporating multi-anatomy datasets with close intensity domains and related segmentation tasks can lead to performance improvements on each dataset.** In turn, this could lead to a more efficient use of imaging resources (pediatric or adult), most notably for the treatment of musculoskeletal disorders affecting different anatomical joints (see Chapter 2). Several patient cohorts impaired by distinct pathologies could be leveraged to optimize a single model

with enhanced generalization capabilities, thus reducing the overall cost of medical image acquisition. More generally, our approach could be transposed to other sets of anatomical structures sharing common characteristics, such as blood vessels in brain, liver, and retina images [263]. Additionally, the multi-scale contrastive regularization could be integrated to enhance vascular segmentation by imposing domain-specific clusters in the embedded spaces of the shared neural network.

Similarly to earlier studies employing highly compact multi-domain models [66]–[69], our work demonstrated that deep neural networks can easily learn related segmentation tasks across multiple intensity domains. Specifically, **this further confirmed the usefulness of employing `DSBN` functions for multi-domain learning, which were previously successfully applied for multi-modal, multi-scanner, multi-center, or multi-protocol segmentation [66]–[69] and have now proven to be equally effective in a multi-anatomy scenario**. Furthermore, when dealing with pediatric patients, it may be beneficial to define domains corresponding to different age groups, as anatomy is significantly modified during child development. However, in this thesis, we were unable to explore such multi-age setting due to the limited amount of imaging resources per age group. Finally, as opposed to previous plain UNet models developed in [66]–[69], our model relied on a more complex architecture based on an pre-trained `EfficientNetB3` encoder to achieve more accurate segmentation and integrated multi-domain spatial attention gates to improve its interpretability.

7.5.4 Limitations

This work has certain limitations which are categorically listed in this section. First, as previously mentioned in Chapter 4, although the coarse localization of the anatomical structures of interest computed by attention gates and the t-SNE visualizations of the learned shared representations provide some interpretability of the network inference process, these approaches do not fully explain the features learned by the segmentation model. Similarly, even though incorporating regularization through the loss function successfully constrains the network parameters and promotes the desired generalizable characteristics during training, the optimization procedure of deep neural networks remains difficult to analyse. Specifically, **while the constraints computed by the multi-scale contrastive regularization are explicit, the interpretability of the multi-joint shape priors, on its part, is limited as it is based on a deep auto-encoder**. This limitation of the deep auto-encoder based shape priors method was already described in

Chapter 4. It is thus essential to develop more interpretable models allowing a finer analysis of the internal behavior of the framework during training and inference. In Chapter 4, we briefly referenced the work of Zhang et al. on interpretable CNN which provides a clear semantic representation by assigning to each filter a specific object part to explicitly memorize during the learning process [237]. One can also mention the ExplAIn [264] framework which introduces an intermediate pixel-level labeling task to directly explain the final image-level lesion classification prediction. Such interpretable and explainable models could therefore be of great interest for medical image analysis applications, as it would allow a better analysis of the network failures. It should be mentioned, that the terms of interpretability and explainability are usually used interchangeably within the community, as these concepts are not rigorously mathematically defined [265].

Second, as introduced in Chapter 1, while a common hypothesis in machine learning is that the training and test data originate from the same data distribution, an emerging field (i.e., domain generalization or out-of-distribution generalization) has proposed to address the more challenging setting in which the goal is to learn a model that can generalize to an unseen test domain [72]–[74]. **In this thesis, although our model managed multiple domains, we only addressed plain generalizability within each domain (i.e., unseen test image from the same distribution as the training data). While the performance improvements obtained during the leave-one-out evaluation indicated better generalization abilities within each domain, our model is currently unable to generalize well on new unseen domains (e.g., new modality or anatomical joint).** In the context of life-long learning in which a single model continuously learns new domains, Karani et al. [67] have demonstrated that `DSBN` parameters could be fine-tuned with limited amount of training data from the novel domain, while the convolutional filters remained fixed. We assumed that our model could be similarly fine-tuned on a new domain without forgetting the knowledge learned on the previous domains. However, this approach still requires access to labeled imaging data from the new domain, unlike domain generalization frameworks in which data from the new test domain is assumed to be unavailable. Out-of-distribution generalization is thus more generic than traditional domain adaption techniques or life-long learning schemes. Therefore, domain generalization appears crucial for medical image segmentation, where each anatomical region and acquisition protocol defines a new domain in which imaging resources are not necessarily available for network training or fine-tuning purposes.

7.5.5 Perspectives

As our experiments were conducted on only one imaging modality (i.e., T1-weighted MR), we were unable to evaluate the genericity of our approach over multiple modalities (e.g., T2-weighted MR, CT) because of the lack of available data. However, previous studies have already demonstrated that a single neural network incorporating shared convolutional filters and `DSBN` functions can effectively process both CT and MR modalities simultaneously [69]. So, we assumed that our model could be easily extended to multiple modalities. Similarly, we limited our experiments to bone tissue segmentation without considering other musculoskeletal tissues such as muscles, ligaments, or cartilages due to the unavailability of annotations. We also hypothesized that our framework could be upgraded to multi-tissue segmentation since previous works [31], [32], [40] have already demonstrated that deep learning models can effectively segment knee cartilages, knee muscles, and shoulder muscles, respectively. More generally, whereas methods developed on natural images employed up to ten domains [77], our experiments involved only three imaging domains due to the scarcity of pediatric imaging resources and the lack of open access pediatric databases. Hence, **future studies are aimed at incorporating supplementary MR imaging sequences to further promote generic features during optimization and segmenting additional tissues to provide a more complete description of the musculoskeletal system.**

In this direction of including an increasing number of imaging datasets, it may be beneficial to adapt our framework to federated learning schemes or annotations efficient approaches (see Figure 1.1). Indeed, a federated learning scheme similar to the one developed by Shen et al. [266] for multi-task pancreas segmentation, would allow optimization of a single model using training data from multiple institutions without centralizing imaging resources. This would consequently prevent data privacy and security issues, which is crucial in medical workflows [266]. For their part, annotation-efficient approaches such as [267], would allow to include imaging datasets with weak labels (e.g., scribbles) and large amount of unlabeled data. In turn, this would reduce the burden of producing large-scale and high-quality annotated segmentation dataset, which is laborious to obtain.

Furthermore, as previously noted noted in Chapter 5, we did not consider 3D architectures in our experiments due to their higher computational complexity and GPU memory consumption compared to their 2D counterparts [30]. Although our models did not integrate a third spatial dimension, we observed smooth delineations in all directions, indicating continuous segmentation predictions between adjacent 2D slices. Addi-

tionally, our experiments were performed using only four neural network architectures (Att-UNet, Inception-UNet, Dense-UNet, and Efficient-UNet), hence it would be beneficial to include supplementary comparisons based on additional deep learning models including Transformers-based ones [268] to further evaluate the genericity of our contributions. Interestingly, the architecture of vision Transformers initially originates from natural language processing and, unlike every model presented in this thesis, does not rely on convolutional layers [255]. One could also consider an ensemble approach integrating all backbone architectures to combine the advantages of each model.

With respect to the contrastive learning, we assumed that the temperature hyper-parameter τ should be constant at each scale, as the cosine similarity between representations was bounded in $[-1, 1]$ regardless of scale. However, we observed in Table 7.8 that contrastive learning was less efficient at certain scales. Specifically, in the DSL+$\mathcal{L}_{\text{MSC}}$ learning scheme, the shared representation of the 9^{th} scale was less disentangled than in 1^{st} and 5^{th} scales. Hence, one could also propose to employ different temperatures at each scale and to learn such parameters during training, so that the contrastive metric be more sensitive at each scale and better disentangle representation between domains. Nevertheless, such a training procedure might be more challenging to optimize due the numerical instability associated with learnable temperature parameters.

7.6 Conclusion

Developing generalizable deep segmentation model is fundamental to provide accurate and reliable delineations on unseen images for clinical and morphological evaluation of the pediatric musculoskeletal system. We introduced a multi-task, multi-domain learning framework for pediatric bone segmentation in sparse MR imaging datasets acquired on separate anatomical joints. This multi-anatomy approach simultaneously benefited from robust shared representations and specialized layers that fitted to the domain-specific intensity distributions and task-specific segmentation label sets. Furthermore, the generalization capabilities of the segmentation model were enhanced by exploiting a multi-scale contrastive regularization to enforce domain clustering in the shared representations and multi-joint shape priors which encouraged anatomically consistent shape predictions.

An important perspective from this thesis is that collaborative utilization of pediatric resources and intelligent design of deep learning models can improve the segmentation performance on small musculoskeletal imaging datasets. Nev-

ertheless, our framework currently provides an incomplete description of the pediatric musculoskeletal system which solely encompass bone tissues. Hence, future work is aimed at improving our model to segment other anatomical structures (e.g., ankle cartilages, knee ligaments, or shoulder muscles). Thus, morphological and functional analysis will rely on a more complete modeling of the musculoskeletal system, towards a better management of pediatric disorders.

CONCLUSION

General conclusion

The research conducted in this thesis aimed to address the generalization gap and data scarcity issues encountered when developing deep learning methods for pediatric musculoskeletal image segmentation. **We proposed and evaluated frameworks based on emerging deep learning paradigms which achieved promising performance for the task of bone segmentation on scarce and heterogeneous pediatric MR imaging datasets of the ankle, knee, and shoulder joints.** In particular, the generalization performances of the segmentation models were enhanced by exploiting state-of-the-art architectures, transfer learning schemes, multi-anatomy approaches, and regularization techniques. The contributions of this work were categorized into two research objectives, as summarized below:

- **Research objective 1.** We developed and evaluated a partially pre-trained convolutional encoder-decoder with combined regularization from shape priors and an adversarial network, which improved performance for multi-structure bone segmentation on pediatric imaging datasets of the musculoskeletal system. The framework benefited from the proposed combined regularization to reduce the data scarcity issue while improving model generalizability. In particular, the shape priors-based regularization, derived from a non-linear shape representation learned by an auto-encoder, guided the segmentation network to make anatomically consistent predictions (Chapter 4). For its part, the adversarial regularization computed by a discriminator network encouraged more precise delineations with limited imaging resources. Additionally, the framework leveraged a state-of-the-art residual encoder and a transfer learning scheme from the ImageNet database to further alleviate data scarcity limitations (Chapter 5). The proposed $\text{CombReg}^{\text{Multi}}_{\text{Res-UNet}}$ achieved excellent performance on both ankle and shoulder datasets, with 94.1% and 89.5% Dice scores respectively (Table 7.7).
- **Research objective 2.** We implemented and assessed a multi-task, multi-domain learning framework for pediatric bone segmentation in sparse MR imaging datasets

acquired on separate anatomical joints. This multi-anatomy approach simultaneously benefited from robust shared representations and specialized layers that fitted to the domain-specific intensity distributions and task-specific segmentation label sets to mitigate the scarcity issue of pediatric resources (Chapter 6). In particular, the multi-task, multi-domain segmentation model integrated a pre-trained Efficient encoder, shared convolutional filters, multi-domain attention gates, domain-specific batch normalization, and domain-specific output layers. Furthermore, the generalization capabilities of the segmentation model were enhanced by exploiting a multi-scale contrastive regularization and multi-joint shape priors. The multi-scale contrastive regularization leveraged dataset label information to enhance intra-domain similarity and impose inter-domain margins, while the multi-joint shape priors encoded the anatomical characteristics of multiple joints to constrain the segmentation task (Chapter 7). The proposed Efficient-UNet $\text{DSL} + \mathcal{L}_{\text{MSC}} + \mathcal{L}_{\text{MJSP}}$ achieved excellent multi-anatomy segmentation performance, reaching 93.8%, 95.4%, and 87.9% Dice scores on ankle, knee, and shoulder datasets (Table 7.7).

In this thesis, we could not evaluate whether the generalization performances achieved by the two proposed frameworks ($\text{CombReg}^{\textbf{Multi}}_{\textbf{Res-UNet}}$ and Efficient-UNet $\textbf{DSL} + \mathcal{L}_{\textbf{MSC}} + \mathcal{L}_{\textbf{MJSP}}$) are sufficient to enable the deployment of fully-automatic bone segmentation methods for pediatric MR imaging in clinical practice. Nevertheless, the obtained results illustrate that collaborative utilization of pediatric resources and intelligent design of deep learning models can improve the segmentation performance on small musculoskeletal imaging datasets.

General limitations

This research work presented certain limitations, which are categorically listed in this section. First, a limitation intrinsic to current deep learning models is the limited interpretability of both the training and inference steps (see Section 1.3.4). Although the regularization terms presented in this thesis (shape priors $\mathcal{L}_{\text{Shape}}$, adversarial $\mathcal{L}_{\text{Adv}}$, multi-joint shape priors $\mathcal{L}_{\text{MJSP}}$, and multi-scale contrastive $\mathcal{L}_{\text{MSC}}$) successfully constrained the network's weights and promoted enhanced generalization capabilities for robust bone extraction, these advanced training methods did not provide a better comprehension of the optimization process. **In particular, at this time, one cannot understand the ef-**

fects of the regularization terms on the highly-dimensional landscape of the loss function and the repercussions on the gradient descent algorithm. Ideally, the regularizations should promote smoother loss functions and prevent the presence of local minima leading to over-fitting issues, but this is extremely challenging to verify in such a highly-dimensional setting. Moreover, since the computation of the $\mathcal{L}_{\text{Shape}}$, $\mathcal{L}_{\text{Adv}}$, and $\mathcal{L}_{\text{MJSP}}$ regularization losses, was itself based on deep learning models (i.e., an auto-encoder or a discriminator), the interpretability of these regularizations was even more limited. For instance, it remains difficult to assess which shape features were constrained by $\mathcal{L}_{\text{Shape}}$ or $\mathcal{L}_{\text{MJSP}}$ during optimization. For their part, the constraints computed by the multi-scale contrastive regularization $\mathcal{L}_{\text{MSC}}$ were explicit and their effects on the learned representations were easier to validate (see Section 7.5.2). In addition, attention gates allowed us to gain some insights into the inference process of segmentation models, but remained insufficient in understanding networks' failures. Finally, it should be emphasized that the "black-box" nature of deep learning also presents ethical issues for potential deployment in clinical practice. Indeed, following the European Union General Data Protection Regulation [1], patients have a right to an explanation of how an automated medical system reached a clinical decision. While the extent of such an "explanation" is not clearly defined, this could limit the deployment of current deep learning models as it remains unclear whether neural networks learn clinically relevant features [269]. In this context, interpretable and explainable artificial intelligence methods [237], [264] seem all the more essential to enable the explanation of neural network decisions and to understand the possible failure of network predictions.

Second, the experiments performed in this thesis had limitations, some of which have already been mentioned in Parts II and III. For instance, all of our frameworks relied on 2D slice-by-slice approaches that do not benefit from 3D spatial information. We favored 2D approaches due to their lower computational complexity and GPU memory consumption over their 3D counterparts (e.g., VNet [30] or 3D UNet [208]). Although our models did not integrate complete spatial context, we observed continuous segmentation predictions between adjacent 2D slices, indicating smooth delineations in all directions. Nevertheless, our models presented a bias along a certain acquisition direction, and it remains unclear whether 3D models could improve segmentation performance. Additionally, it should be noted that combining transfer learning schemes from the 2D ImageNet database with 3D models is challenging due to the difference in data dimensionality. To circumvent this

1. `https://gdpr-info.eu/`

problem, it would be beneficial to employ 3D models pre-trained on 3D computer vision or medical image analysis tasks. In this direction, initiatives such as the ModelZoo [2] aim to curate and provide platforms to easily find pre-trained models for various deep learning software and applications. For its part, the open-source MONAI framework [3], which is specific to medical image applications, also offers access to pre-trained 3D models. Leveraging knowledge (i.e., weights) obtained from large-scale medical databases could be highly relevant in the context of small-scale pediatric imaging dataset segmentation. Nevertheless, fine-tuning these 3D models with limited GPU computational capacity remains challenging. Ultimately, experiments performed in this thesis were restricted by the computational capacity of the available GPU, and we could not evaluate whether models implemented on multi-GPUs may achieve more stable optimization or better results at inference.

Third, our experiments presented additional limitations directly related to the nature of the pediatric imaging resources available for this thesis. In particular, due to the scarcity of pediatric imaging resources and the lack of open access pediatric databases, we only employed three sparse MR image datasets of the ankle, knee, and shoulder joints. Hence, we limited our experiments to only one imaging modality (i.e., T1-weighted MR), without considering other modalities such as T2-weighted MR, X-ray, or CT. However, previous studies have illustrated that deep learning can successfully segment musculoskeletal structures in CT scans [206], [224], [226]–[228], X-ray radiographs [178], [231], ultrasound [179], [225], or PET/CT scans [230]. Similarly, our experiments were only conducted for the segmentation of bone tissues, and we were unable to evaluate the genericity of our approach to other musculoskeletal tissues (i.e., muscles, ligaments, or cartilages) due to the unavailability of annotations. Nevertheless, neural networks have also been demonstrated to efficiently extract multiple tissues simultaneously, such as knee bones, muscles, cartilage, and ligaments [31]–[36]. So, we assumed that our models developed in Parts II and III could be easily extended to other modalities and upgraded to multi-tissue segmentation due to the versatility of deep learning. **Ultimately, the clinical impact of the methods presented in this thesis could be considered limited as our experiments only targeted the segmentation of pediatric bones and consequently failed to provide a complete description of the musculoskeletal system.**

2. `https://modelzoo.co/`
3. `https://monai.io/`

General perspectives

Based on the methods developed and results obtained in this thesis, the following future research perspectives can be considered. First, as mentioned in the previous section, a straightforward extension of our models lies in the integration of other musculoskeletal tissues to provide a more detailed assessment of the pediatric anatomy. Ideally, the extracted 3D meshes of bones, muscles, cartilages, and ligaments could be employed in the subsequent morphological and functional analyses to better manage pediatric disorders. For instance, volumetric or anatomical information provided by the generated segmentation could help clinicians assess the patient's level of impairment [166]. For its part, the multi-task, multi-domain learning framework had specific benefits and perspectives for clinical practice. In particular, the results obtained in Part III illustrated that the collaborative utilization of pediatric resources and intelligent design of deep learning models could improve the segmentation performance on small musculoskeletal imaging datasets. Hence, **future studies could leverage additional patient cohorts impaired by distinct pathologies to optimize a single model with enhanced generalization capabilities, thus reducing the overall cost of medical image acquisition**. In this direction of including an increasing number of imaging datasets, it may be beneficial to consider federated learning approaches [21]. This would allow optimization of a single model using training data from multiple institutions without centralizing imaging resources, thus allowing to prevent data privacy and security issues, which is crucial in medical workflows. However, it should be mentioned that training models over large datasets spread across multiple imaging centers can also raise ecological concerns. Indeed, optimizing deep learning models requires GPUs whose energy consumption and associated carbon footprint can be substantial when considering large datasets. Interestingly, the carbon emissions can vary drastically depending on the location and time at which the training is performed. Recent studies have thus proposed guidelines for reducing carbon emissions during model development [270].

Second, while our experiments included several state-of-the-art pre-trained encoders (`VGG19` [60], `DenseNet121` [61], `ResNet50` [59], `InceptionV3` [86], `EfficientNetB3` [191]), Transformers-based architectures have recently shown promising results in computer vision [255] and medical image analysis tasks [268]. Similarly, the self-configuring method nn-UNet has also attracted the attention of the medical image research community due to its ability to configure some hyper-parameters automatically and provide robust perfor-

mance on diverse segmentation challenges [207]. **It would therefore be beneficial to perform supplementary comparisons and further evaluation of the genericity of our contributions** by employing Transformer or nn-UNet models as backbones. Interestingly, unlike every model proposed in this thesis, vision Transformers are not based on convolutional layers but employ attention mechanisms originating from natural language processing. Hence, vision Transformers lack the inductive bias resulting from convolutional operators (i.e., equivariant representations) and can capture global and wider range relations in the image, but at the cost of a more onerous training and complex network architecture [255]. Indeed, the baseline vision Transform encompasses 86 million learnable parameters while its "huge" equivalent necessitates 632 million parameters. It can thus be impractical to train or fine-tune such models using a single GPU. Nevertheless, recent works have proposed hybrid pipelines combining vision Transformers and CNNs, to learn long-range dependencies and effectively capture global contextual representation at multiple scales while reducing memory consumption [268], [271]. This research direction could thus be highly relevant to design more efficient models for medical image analysis tasks.

Third, while our approaches targeted the generalization gap and data scarcity issues, the paucity of ground truth annotations remains an important challenge for deep learning-based medical image analysis. **It would thus be beneficial to incorporate annotations-efficient approaches into our frameworks to reduce the burden of producing labels for pediatric examinations, which are laborious to obtain.** Hence, semi- or self-supervision would allow leveraging a large amount of unlabeled data, while weakly-supervised schemes would enable the inclusion of weak segmentation labels, such as contours scribbles. For its part, the combination of interactive segmentation tools and few-shot learning methods has the potential to reduce the annotation burden by enabling the user to make minor corrections interactively [21]. In this direction, one could also consider out-of-distribution generalization schemes to reduce the need for annotated imaging resources [72]–[74]. Indeed, while the networks developed in this thesis only addressed plain generalizability within each domain, out-of-distribution generalization aims to learn a model that can generalize to an unseen test domain. This appears crucial for medical image segmentation, where each anatomical region and acquisition protocol defines a new domain in which imaging resources are not necessarily available for network training or fine-tuning purposes. Ultimately, an ideal medical image segmentation system should be an intelligent and domain-agnostic model capable of automatically delineating

any anatomical structures. Such a system could be integrated seamlessly into any imaging device and provide morphological and functional information, regardless of the clinical application, with little to no user interaction.

References

[1] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, « A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises », *Proceedings of the IEEE*, vol. 109, *5*, pp. 820–838, 2021, ISSN: 0018-9219. DOI: 10.1109/JPROC.2021.3054390.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, « A survey on deep learning in medical image analysis », *Medical Image Analysis*, vol. 42, pp. 60–88, 2017, ISSN: 13618415. DOI: 10.1016/j.media.2017.07.005.

[3] A. Hirschmann, J. Cyriac, B. Stieltjes, T. Kober, J. Richiardi, and P. Omoumi, « Artificial intelligence in musculoskeletal imaging: review of current literature, challenges, and trends », *Seminars in Musculoskeletal Radiology*, vol. 23, *3*, pp. 304–311, 2019, ISSN: 1098-898X. DOI: 10.1055/s-0039-1684024.

[4] J. E. Burns, J. Yao, and R. M. Summers, « Artificial intelligence in musculoskeletal imaging: a paradigm shift », *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research*, vol. 35, *1*, pp. 28–35, 2020, ISSN: 1523-4681. DOI: 10.1002/jbmr.3849.

[5] N. Gorelik and S. Gyftopoulos, « Applications of artificial intelligence in musculoskeletal imaging: from the request to the report », *Canadian Association of Radiologists Journal = Journal l'Association Canadienne Des Radiologistes*, vol. 72, *1*, pp. 45–59, 2021, ISSN: 1488-2361. DOI: 10.1177/0846537120947148.

[6] C. Pons, « Analyse morphologique et biomécanique de l'épaule et du membre supérieur des enfants avec une paralysie obstétricale du plexus brachial : impact sur les thérapeutiques », These de doctorat, Brest, 2018.

[7] J. S. Meyer and D. Jaramillo, « Musculoskeletal MR imaging at 3T », *Magnetic Resonance Imaging Clinics of North America*, vol. 16, *3*, pp. 533–545, vi, 2008, ISSN: 1064-9689. DOI: 10.1016/j.mric.2008.04.004.

[8] D. Jaramillo and T. Laor, « Pediatric musculoskeletal MRI: basic principles to optimize success », *Pediatric Radiology*, vol. 38, *4*, pp. 379–391, 2008, ISSN: 0301-0449. DOI: 10.1007/s00247-007-0645-4.

[9] C. Balassy and M. Hörmann, « Role of MRI in paediatric musculoskeletal conditions », *European Journal of Radiology*, vol. 68, *2*, pp. 245–258, 2008, ISSN: 0720-048X. DOI: 10.1016/j.ejrad.2008.07.018.

[10] J. S. Duncan, M. F. Insana, and N. Ayache, « Biomedical imaging and analysis in the age of big data and deep learning [scanning the issue] », *Proceedings of the IEEE*, vol. 108, *1*, pp. 3–10, 2020, ISSN: 1558-2256. DOI: 10.1109/JPROC.2019.2956422.

[11] D. Rueckert and J. A. Schnabel, « Model-based and data-driven strategies in medical image computing », *Proceedings of the IEEE*, vol. 108, *1*, pp. 110–124, 2020, ISSN: 1558-2256. DOI: 10.1109/JPROC.2019.2943836.

[12] A. S. Lundervold and A. Lundervold, « An overview of deep learning in medical imaging focusing on MRI », *Zeitschrift für Medizinische Physik*, Special Issue: Deep Learning in Medical Physics, vol. 29, *2*, pp. 102–127, 2019, ISSN: 0939-3889. DOI: 10.1016/j.zemedi.2018.11.002.

[13] Y. LeCun, Y. Bengio, and G. Hinton, « Deep learning », *Nature*, vol. 521, *7553*, pp. 436–444, 2015, ISSN: 1476-4687. DOI: 10.1038/nature14539.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, red. by F. Bach, ser. Adaptive Computation and Machine Learning series. Cambridge, MA, USA: MIT Press, 2016, 800 pp., ISBN: 978-0-262-03561-3. URL: http://www.deeplearningbook.org.

[15] C. Fang, S. Bai, Q. Chen, Y. Zhou, L. Xia, L. Qin, S. Gong, X. Xie, C. Zhou, D. Tu, C. Zhang, X. Liu, W. Chen, X. Bai, and P. H. S. Torr, « Deep learning for predicting COVID-19 malignant progression », *Medical Image Analysis*, vol. 72, p. 102 096, 2021, ISSN: 1361-8423. DOI: 10.1016/j.media.2021.102096.

[16] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, « Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs »,

JAMA, vol. 316, *22*, pp. 2402–2410, 2016, ISSN: 1538-3598. DOI: 10.1001/jama.2016.17216.

[17] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, « Key challenges for delivering clinical impact with artificial intelligence », *BMC Medicine*, vol. 17, *1*, p. 195, 2019, ISSN: 1741-7015. DOI: 10.1186/s12916-019-1426-2.

[18] S. Benjamens, P. Dhunnoo, and B. Meskó, « The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database », *npj Digital Medicine*, vol. 3, *1*, pp. 1–8, 2020, ISSN: 2398-6352. DOI: 10.1038/s41746-020-00324-0.

[19] D. Lyell, E. Coiera, J. Chen, P. Shah, and F. Magrabi, « How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices », *BMJ Health & Care Informatics*, vol. 28, *1*, e100301, 2021, ISSN: 2632-1009. DOI: 10.1136/bmjhci-2020-100301.

[20] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, « Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis », *Medical Image Analysis*, vol. 54, pp. 280–296, 2019, ISSN: 1361-8423. DOI: 10.1016/j.media.2019.03.009.

[21] N. Tajbakhsh, H. Roth, D. Terzopoulos, and J. Liang, « Guest editorial annotation-efficient deep learning: the holy grail of medical imaging », *IEEE Transactions on Medical Imaging*, vol. 40, *10*, pp. 2526–2533, 2021, ISSN: 1558-254X. DOI: 10.1109/TMI.2021.3089292.

[22] A. Takalkar and M. Hernandez Pampaloni, « Pediatric cardiac PET/CT imaging », *PET clinics*, vol. 15, *3*, pp. 371–380, 2020, ISSN: 1879-9809. DOI: 10.1016/j.cpet.2020.03.011.

[23] N. S. Kwatra, R. Lim, M. S. Gee, L. J. States, A. Vossough, and E. Y. Lee, « PET/MR imaging: current updates on pediatric applications », *Magnetic Resonance Imaging Clinics of North America*, vol. 27, *2*, pp. 387–407, 2019, ISSN: 1557-9786. DOI: 10.1016/j.mric.2019.01.012.

[24] S. Gatidis, C. la Fougère, and J. F. Schaefer, « Pediatric oncologic imaging: a key application of combined PET/MRI », *RoFo: Fortschritte Auf Dem Gebiete Der*

Rontgenstrahlen Und Der Nuklearmedizin, vol. 188, *4*, pp. 359–364, 2016, ISSN: 1438-9010. DOI: 10.1055/s-0041-109513.

[25] A. Berg and G. Greve, « Trends in pediatric imaging: ultrasound », *Acta Radiologica (Stockholm, Sweden: 1987)*, vol. 54, *9*, pp. 1096–1105, 2013, ISSN: 1600-0455. DOI: 10.1177/0284185113501808.

[26] G. R. Schooler, J. T. Davis, H. E. Daldrup-Link, and D. P. Frush, « Current utilization and procedural practices in pediatric whole-body MRI », *Pediatric Radiology*, vol. 48, *8*, pp. 1101–1107, 2018, ISSN: 1432-1998. DOI: 10.1007/s00247-018-4145-5.

[27] M. B. K. Sammer, A. C. Sher, L. J. States, A. T. Trout, and V. J. Seghers, « Current trends in pediatric nuclear medicine: a society for pediatric radiology membership survey », *Pediatric Radiology*, vol. 50, *8*, pp. 1139–1147, 2020, ISSN: 1432-1998. DOI: 10.1007/s00247-020-04670-9.

[28] S. Mallat, « Understanding deep convolutional networks », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, *2065*, p. 20 150 203, 2016. DOI: 10.1098/rsta.2015.0203.

[29] O. Ronneberger, P. Fischer, and T. Brox, « U-Net: convolutional networks for biomedical image segmentation », *in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

[30] F. Milletari, N. Navab, and S.-A. Ahmadi, « V-Net: fully convolutional neural networks for volumetric medical image segmentation », *in 2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

[31] F. Ambellan, A. Tack, M. Ehlke, and S. Zachow, « Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: data from the osteoarthritis initiative », *Medical Image Analysis*, vol. 52, pp. 109–118, 2019, ISSN: 1361-8423. DOI: 10.1016/j.media.2018.11.009.

[32] Z. Zhou, G. Zhao, R. Kijowski, and F. Liu, « Deep convolutional neural network for segmentation of knee joint anatomy », *Magnetic Resonance in Medicine*, vol. 80, *6*, pp. 2759–2770, 2018, ISSN: 1522-2594. DOI: 10.1002/mrm.27229.

[33] S. Ebrahimkhani, M. H. Jaward, F. M. Cicuttini, A. Dharmaratne, Y. Wang, and A. G. S. de Herrera, « A review on segmentation of knee articular cartilage: from conventional methods towards deep learning », *Artificial Intelligence in Medicine*, vol. 106, p. 101 851, 2020, ISSN: 1873-2860. DOI: 10.1016/j.artmed.2020.101851.

[34] E. Panfilov, A. Tiulpin, M. T. Nieminen, S. Saarakkala, and V. Casula, « Deep learning-based segmentation of knee MRI for fully automatic subregional morphological assessment of cartilage tissues: data from the osteoarthritis initiative », *Journal of Orthopaedic Research: Official Publication of the Orthopaedic Research Society*, vol. 40, *5*, pp. 1113–1124, 2022, ISSN: 1554-527X. DOI: 10.1002/jor.25150.

[35] A. M. Schmidt, A. D. Desai, L. E. Watkins, H. A. Crowder, M. S. Black, V. Mazzoli, E. B. Rubin, Q. Lu, J. W. MacKay, R. D. Boutin, F. Kogan, G. E. Gold, B. A. Hargreaves, and A. S. Chaudhari, « Generalizability of deep learning segmentation algorithms for automated assessment of cartilage morphology and MRI relaxometry », *Journal of magnetic resonance imaging: JMRI*, 2022, ISSN: 1522-2586. DOI: 10.1002/jmri.28365.

[36] S. W. Flannery, A. M. Kiapour, D. J. Edgar, M. M. Murray, and B. C. Fleming, « Automated magnetic resonance image segmentation of the anterior cruciate ligament », *Journal of Orthopaedic Research: Official Publication of the Orthopaedic Research Society*, vol. 39, *4*, pp. 831–840, 2021, ISSN: 1554-527X. DOI: 10.1002/jor.24926.

[37] X. He, C. Tan, Y. Qiao, V. Tan, D. Metaxas, and K. Li, « Effective 3D humerus and scapula extraction using low-contrast and high-shape-variability MR data », *in Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10953, SPIE, 2019, pp. 118–124. DOI: 10.1117/12.2513107.

[38] E. Brui, A. Y. Efimtcev, V. A. Fokin, R. Fernandez, A. G. Levchuk, A. C. Ogier, A. A. Samsonov, J. P. Mattei, I. V. Melchakova, D. Bendahan, and A. Andreychenko, « Deep learning-based fully automatic segmentation of wrist cartilage in MR images », *NMR in biomedicine*, vol. 33, *8*, e4320, 2020, ISSN: 1099-1492. DOI: 10.1002/nbm.4320.

[39] J. Ding, P. Cao, H.-C. Chang, Y. Gao, S. H. S. Chan, and V. Vardhanabhuti, « Deep learning-based thigh muscle segmentation for reproducible fat fraction quantification using fat-water decomposition MRI », *Insights into Imaging*, vol. 11, *1*, p. 128, 2020, ISSN: 1869-4101. DOI: 10.1186/s13244-020-00946-8.

[40] P.-H. Conze, S. Brochard, V. Burdin, F. T. Sheehan, and C. Pons, « Healthy versus pathological learning transferability in shoulder muscle MRI segmentation using deep convolutional encoder-decoders », *Computerized Medical Imaging and Graphics*, vol. 83, p. 101 733, 2020, ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2020.101733.

[41] S. Ioffe and C. Szegedy, « Batch normalization: accelerating deep network training by reducing internal covariate shift », *in Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, PMLR, 2015, pp. 448–456. URL: https://proceedings.mlr.press/v37/ioffe15.html.

[42] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, « Attention U-Net: learning where to look for the pancreas », *in Medical Imaging with Deep Learning*, 2018. URL: https://openreview.net/forum?id=Skft7cijM.

[43] J. Kukačka, V. Golkov, and D. Cremers, *Regularization for deep learning: a taxonomy*, 2017. DOI: 10.48550/arXiv.1710.10686. arXiv: 1710.10686[cs,stat].

[44] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, « Transfusion: understanding transfer learning for medical imaging », *in Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91-Abstract.html.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, « ImageNet: a large-scale hierarchical image database », *in 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[46] P.-H. Conze, A. E. Kavur, E. Cornec-Le Gall, N. S. Gezer, Y. Le Meur, M. A. Selver, and F. Rousseau, « Abdominal multi-organ segmentation with cascaded convolutional and adversarial deep networks », *Artificial Intelligence in Medicine*, vol. 117, p. 102 109, 2021, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2021.102109.

[47] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, « DCAN: deep contour-aware networks for object instance segmentation from histology images », *Medical Image Analysis*, vol. 36, pp. 135–146, 2017, ISSN: 1361-8415. DOI: 10.1016/j.media.2016.11.004.

[48] K. Josephson, A. Ericsson, and J. Karlsson, « Segmentation of medical images using three-dimensional active shape models », *in Image Analysis*, H. Kalviainen, J. Parkkinen, and A. Kaarna, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2005, pp. 719–728, ISBN: 978-3-540-31566-7. DOI: 10.1007/11499145_73.

[49] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, « Multi-organ localization with cascaded global-to-local regression and shape prior », *Medical Image Analysis*, vol. 23, *1*, pp. 70–83, 2015, ISSN: 1361-8415. DOI: 10.1016/j.media.2015.04.007.

[50] M. S. Nosrati and G. Hamarneh, *Incorporating prior knowledge in medical image segmentation: a survey*, 2016. DOI: 10.48550/arXiv.1607.01092. arXiv: 1607.01092[cs].

[51] A. V. Dalca, J. Guttag, and M. R. Sabuncu, « Anatomical priors in convolutional networks for unsupervised biomedical segmentation », *in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9290–9299. DOI: 10.1109/CVPR.2018.00968.

[52] A. Myronenko, « 3D MRI brain tumor segmentation using autoencoder regularization », *in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 311–320, ISBN: 978-3-030-11726-9. DOI: 10.1007/978-3-030-11726-9_28.

[53] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, « Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation », *IEEE Transactions on Medical Imaging*, vol. 37, *2*, pp. 384–395, 2018, ISSN: 1558-254X. DOI: 10.1109/TMI.2017.2743464.

[54] D. D. Pham, G. Dovletov, S. Warwas, S. Landgraeber, M. Jäger, and J. Pauli, « Deep learning with anatomical priors: imitating enhanced autoencoders in latent space for improved pelvic bone segmentation in MRI », *in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 1166–1169. DOI: 10.1109/ISBI.2019.8759221.

[55] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, « Learning and incorporating shape models for semantic segmentation », *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2017, pp. 203–211, ISBN: 978-3-319-66182-7. DOI: 10.1007/978-3-319-66182-7_24.

[56] D. Nie and D. Shen, « Adversarial confidence learning for medical image segmentation and synthesis », *International Journal of Computer Vision*, vol. 128, *10*, pp. 2494–2513, 2020, ISSN: 0920-5691. DOI: 10.1007/s11263-020-01321-2.

[57] V. K. Singh, H. A. Rashwan, S. Romani, F. Akram, N. Pandey, M. M. K. Sarker, A. Saleh, M. Arenas, M. Arquez, D. Puig, and J. Torrents-Barrena, « Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network », *Expert Systems with Applications*, vol. 139, p. 112 855, 2020, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2019.112855.

[58] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, « SegAN: adversarial network with multi-scale L1 loss for medical image segmentation », *Neuroinformatics*, vol. 16, *3*, pp. 383–392, 2018, ISSN: 1559-0089. DOI: 10.1007/s12021-018-9377-x.

[59] K. He, X. Zhang, S. Ren, and J. Sun, « Deep residual learning for image recognition », *in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[60] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. DOI: 10.48550/arXiv.1409.1556. arXiv: 1409.1556[cs].

[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, « Densely connected convolutional networks », *in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

[62] L. Song, J. Lin, Z. J. Wang, and H. Wang, « An end-to-end multi-task deep learning framework for skin lesion analysis », *IEEE Journal of Biomedical and Health Informatics*, vol. 24, *10*, pp. 2912–2921, 2020, ISSN: 2168-2208. DOI: 10.1109/JBHI.2020.2973614.

[63] T.-L.-T. Le, N. Thome, S. Bernard, V. Bismuth, and F. Patoureaux, « Multitask classification and segmentation for cancer diagnosis in mammography », *in Medical Imaging with Deep Learning*, 2019. URL: https://openreview.net/forum?id=r1xDM5DGcV.

[64] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, J. Joseph, and M. Sivaprakasam, « Psi-Net: shape and boundary aware joint multi-task deep network for medical image segmentation », *in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 7223–7226. DOI: 10.1109/EMBC.2019.8857339.

[65] S. Chen, G. Bortsova, A. García-Uceda Juárez, G. van Tulder, and M. de Bruijne, « Multi-task attention-based semi-supervised learning for medical image segmentation », *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 457–465, ISBN: 978-3-030-32248-9. DOI: 10.1007/978-3-030-32248-9_51.

[66] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, « MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data », *IEEE Transactions on Medical Imaging*, vol. 39, *9*, pp. 2713–2724, 2020, ISSN: 0278-0062. DOI: 10.1109/TMI.2020.2974574.

[67] N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu, « A lifelong learning approach to brain MR segmentation across scanners and protocols », *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 476–484, ISBN: 978-3-030-00928-1. DOI: 10.1007/978-3-030-00928-1_54.

[68] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, « Domain-specific batch normalization for unsupervised domain adaptation », *in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7346–7354. DOI: 10.1109/CVPR.2019.00753.

[69] Q. Dou, Q. Liu, P. A. Heng, and B. Glocker, « Unpaired multi-modal segmentation via knowledge distillation », *IEEE Transactions on Medical Imaging*, vol. 39, *7*, pp. 2415–2425, 2020, ISSN: 1558-254X. DOI: 10.1109/TMI.2019.2963882.

[70] V. V. Valindria, N. Pawlowski, M. Rajchl, I. Lavdas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker, « Multi-modal learning from unpaired images: application to multi-organ segmentation in CT and MRI », *in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 547–556. DOI: 10.1109/WACV.2018.00066.

[71] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, « Unsupervised domain adaptation in brain lesion segmentation with adversarial networks », *in Information Processing in Medical Imaging*, M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap, and D. Shen, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2017, pp. 597–609, ISBN: 978-3-319-59050-9. DOI: 10.1007/978-3-319-59050-9_47.

[72] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, « Generalizing to unseen domains: a survey on domain generalization », *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2022, ISSN: 1558-2191. DOI: 10.1109/TKDE.2022.3178128.

[73] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, « Domain generalization: a survey », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2022.3195549.

[74] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin, « Generalizing to unseen domains: a survey on domain generalization », *in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed., International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4627–4635. DOI: 10.24963/ijcai.2021/628.

[75] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Trémeau, and C. Wolf, « Semantic segmentation via multi-task, multi-domain learning », *in Structural, Syntactic, and Statistical Pattern Recognition*, A. Robles-Kelly, M. Loog, B. Biggio, F. Escolano, and R. Wilson, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 333–343, ISBN: 978-3-319-49055-7. DOI: 10.1007/978-3-319-49055-7_30.

[76] H. Bilen and A. Vedaldi, *Universal representations:the missing link between faces, text, planktons, and cat breeds*, 2017. DOI: 10.48550/arXiv.1701.07275. arXiv: 1701.07275[cs,stat].

[77] S.-A. Rebuffi, A. Vedaldi, and H. Bilen, « Efficient parametrization of multi-domain deep neural networks », *in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8119–8127. DOI: 10.1109/CVPR.2018.00847.

[78] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, « Learning multiple visual domains with residual adapters », *in Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/e7b24b112a44fdd9ee93bdf998c6ca0e-Abstract.html.

[79] P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, « Deep learning for multi-task medical image segmentation in multiple modalities », *in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 478–486, ISBN: 978-3-319-46723-8. DOI: 10.1007/978-3-319-46723-8_55.

[80] Y. Zhu, Y. Tang, Y. Tang, D. C. Elton, S. Lee, P. J. Pickhardt, and R. M. Summers, « Cross-domain medical image translation by shared latent gaussian mixture model », *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 379–389, ISBN: 978-3-030-59713-9. DOI: 10.1007/978-3-030-59713-9_37.

[81] Y. Bengio, A. Courville, and P. Vincent, « Representation learning: a review and new perspectives », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, *8*, pp. 1798–1828, 2013, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2013.50.

[82] R. Hadsell, S. Chopra, and Y. LeCun, « Dimensionality reduction by learning an invariant mapping », *in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.

[83] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, « A simple framework for contrastive learning of visual representations », *in International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html.

[84] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, « Domain generalization via model-agnostic learning of semantic features », *in Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/2974788b53f73e7950e8aa49f3a306db-Abstract.html.

[85] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, « Supervised contrastive learning », *in Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 18 661–18 673. URL: https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

[86] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, « Rethinking the inception architecture for computer vision », *in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

[87] L. v. d. Maaten and G. Hinton, « Visualizing data using t-SNE », *Journal of Machine Learning Research*, vol. 9, *86*, pp. 2579–2605, 2008, ISSN: 1533-7928. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[88] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, « A deep learning mammography-based model for improved breast cancer risk prediction », *Radiology*, vol. 292, *1*, pp. 60–66, 2019, ISSN: 1527-1315. DOI: 10.1148/radiol.2019182716.

[89] L. Zhang, M. Wang, M. Liu, and D. Zhang, « A survey on deep learning for neuroimaging-based brain disorder analysis », *Frontiers in Neuroscience*, vol. 14, p. 779, 2020, ISSN: 1662-4548. DOI: 10.3389/fnins.2020.00779.

[90] L. Saba, S. S. Sanagala, S. K. Gupta, V. K. Koppula, A. M. Johri, A. M. Sharma, R. Kolluri, D. L. Bhatt, A. Nicolaides, and J. S. Suri, « Ultrasound-based internal carotid artery plaque characterization using deep learning paradigm on a supercomputer: a cardiovascular disease/stroke risk assessment system », *The International Journal of Cardiovascular Imaging*, vol. 37, *5*, pp. 1511–1528, 2021, ISSN: 1875-8312. DOI: 10.1007/s10554-020-02124-9.

[91] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, B. Baydar, D. Lachinov, S. Han, J. Pauli, F. Isensee, M. Perkonigg, R. Sathish, R. Rajan, D. Sheet, G. Dovletov, O. Speck, A. Nürnberger, K. H. Maier-Hein, G. Bozdağı Akar, G. Ünal, O. Dicle, and M. A. Selver, « CHAOS challenge - combined (CT-MR) healthy abdominal organ segmentation », *Medical Image Analysis*, vol. 69, p. 101 950, 2021, ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101950.

[92] F. Liu, Z. Zhou, A. Samsonov, D. Blankenbaker, W. Larison, A. Kanarek, K. Lian, S. Kambhampati, and R. Kijowski, « Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection », *Radiology*, vol. 289, *1*, pp. 160–169, 2018, ISSN: 1527-1315. DOI: 10.1148/radiol.2018172986.

[93] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G. S. Corrado, L. H. Peng, D. R. Webster, D. Ai, S. J. Huang, Y. Liu, R. C. Dunn, and D. Coz, « A deep learning system for differential diagnosis of skin diseases », *Nature Medicine*, vol. 26, *6*, pp. 900–908, 2020, ISSN: 1546-170X. DOI: 10.1038/s41591-020-0842-3.

[94] G. Quellec, M. Lamard, P.-H. Conze, P. Massin, and B. Cochener, « Automatic detection of rare pathologies in fundus photographs using few-shot learning », *Medical Image Analysis*, vol. 61, p. 101 660, 2020, ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101660.

[95] A. Ben Hamida, M. Devanne, J. Weber, C. Truntzer, V. Derangère, F. Ghiringhelli, G. Forestier, and C. Wemmert, « Deep learning for colon cancer histopathological images analysis », *Computers in Biology and Medicine*, vol. 136, p. 104 730, 2021, ISSN: 1879-0534. DOI: 10.1016/j.compbiomed.2021.104730.

[96] G. Al Hinai, S. Jammoul, Z. Vajihi, and J. Afilalo, « Deep learning analysis of resting electrocardiograms for the detection of myocardial dysfunction, hypertrophy, and ischaemia: a systematic review », *European Heart Journal. Digital Health*, vol. 2, *3*, pp. 416–423, 2021, ISSN: 2634-3916. DOI: 10.1093/ehjdh/ztab048.

[97] J. C. Quiroz, L. Laranjo, A. B. Kocaballi, S. Berkovsky, D. Rezazadegan, and E. Coiera, « Challenges of developing a digital scribe to reduce clinical documentation burden », *npj Digital Medicine*, vol. 2, *1*, pp. 1–6, 2019, ISSN: 2398-6352. DOI: 10.1038/s41746-019-0190-1.

[98] N. Kanwal and G. Rizzo, « Attention-based clinical note summarization », *in Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '22, New York, NY, USA: Association for Computing Machinery, 2022, pp. 813–820, ISBN: 978-1-4503-8713-2. DOI: 10.1145/3477314.3507256.

[99] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, « Highly accurate protein structure prediction with AlphaFold », *Nature*, vol. 596, *7873*, pp. 583–589, 2021, ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2.

[100] S. Koitka, M. S. Kim, M. Qu, A. Fischer, C. M. Friedrich, and F. Nensa, « Mimicking the radiologists' workflow: estimating pediatric hand bone age with stacked deep neural networks », *Medical Image Analysis*, vol. 64, p. 101 743, 2020, ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101743.

[101] P. Savadjiev, Y. Rathi, S. Bouix, A. R. Smith, R. T. Schultz, R. Verma, and C.-F. Westin, « Fusion of white and gray matter geometry: a framework for investigating brain development », *Medical Image Analysis*, Special Issue on the 2013 Conference on Medical Image Computing and Computer Assisted Intervention, vol. 18, *8*, pp. 1349–1360, 2014, ISSN: 1361-8415. DOI: 10.1016/j.media.2014.06.013.

[102] M. Hoheisel, « Review of medical imaging with emphasis on x-ray detectors », *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Proceedings of the 7th International Workshop on Radiation Imaging Detectors, vol. 563, *1*, pp. 215–224, 2006, ISSN: 0168-9002. DOI: 10.1016/j.nima.2006.01.123.

[103] J. H. Scatliff and P. J. Morris, « From roentgen to magnetic resonance imaging: the history of medical imaging », *North Carolina Medical Journal*, vol. 75, *2*, pp. 111–113, 2014, ISSN: 0029-2559. DOI: 10.18043/ncm.75.2.111.

[104] A. Caroppo, A. Leone, and P. Siciliano, « Deep transfer learning approaches for bleeding detection in endoscopy images », *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, vol. 88, p. 101 852, 2021, ISSN: 1879-0771. DOI: 10.1016/j.compmedimag.2020.101852.

[105] R. Behling, « X-ray sources: 125 years of developments of this intriguing technology », *Physica medica: PM: an international journal devoted to the applications of physics to medicine and biology: official journal of the Italian Association of Biomedical Physics (AIFB)*, vol. 79, pp. 162–187, 2020, ISSN: 1724-191X. DOI: 10.1016/j.ejmp.2020.07.021.

[106] L. W. Goldman, « Principles of CT and CT technology », *Journal of Nuclear Medicine Technology*, vol. 35, *3*, pp. 115–128, 2007, ISSN: 0091-4916. DOI: 10.2967/jnmt.107.042978.

[107] J. E. Aldrich, « Basic physics of ultrasound imaging », *Critical Care Medicine*, vol. 35, *5*, S131–137, 2007, ISSN: 0090-3493. DOI: 10.1097/01.CCM.0000260624.99430.22.

[108] M. D. Coltrera, « Ultrasound physics in a nutshell », *Otolaryngologic Clinics of North America*, vol. 43, *6*, pp. 1149–1159, 2010, ISSN: 1557-8259. DOI: 10.1016/j.otc.2010.08.004.

[109] D. W. Townsend, « Physical principles and technology of clinical PET imaging », *Annals of the Academy of Medicine, Singapore*, vol. 33, *2*, pp. 133–145, 2004, ISSN: 0304-4602.

[110] S. Basu, S. Hess, P.-E. Nielsen Braad, B. B. Olsen, S. Inglev, and P. F. Høilund-Carlsen, « The basic principles of FDG-PET/CT imaging », *PET clinics*, vol. 9, *4*, pp. 355–370, v, 2014, ISSN: 1879-9809. DOI: 10.1016/j.cpet.2014.07.006.

[111] B. Bybel, R. C. Brunken, F. P. DiFilippo, D. R. Neumann, G. Wu, and M. D. Cerqueira, « SPECT/CT imaging: clinical utility of an emerging technology », *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, vol. 28, *4*, pp. 1097–1113, 2008, ISSN: 1527-1323. DOI: 10.1148/rg.284075203.

[112] H. F. Wehrl, A. W. Sauter, M. R. Divine, and B. J. Pichler, « Combined PET/MR: a technology becomes mature », *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, vol. 56, *2*, pp. 165–168, 2015, ISSN: 1535-5667. DOI: 10.2967/jnumed.114.150318.

[113] P. Börnert and D. G. Norris, « A half-century of innovation in technology-preparing MRI for the 21st century », *The British Journal of Radiology*, vol. 93, *1111*, p. 20 200 113, 2020, ISSN: 1748-880X. DOI: 10.1259/bjr.20200113.

[114] T. Ai, J. N. Morelli, X. Hu, D. Hao, F. L. Goerner, B. Ager, and V. M. Runge, « A historical overview of magnetic resonance imaging, focusing on technological innovations », *Investigative Radiology*, vol. 47, *12*, pp. 725–741, 2012, ISSN: 1536-0210. DOI: 10.1097/RLI.0b013e318272d29f.

[115] U. Kanniyappan, B. Wang, C. Yang, P. Ghassemi, M. Litorja, N. Suresh, Q. Wang, Y. Chen, and T. J. Pfefer, « Performance test methods for near-infrared fluorescence imaging », *Medical Physics*, vol. 47, *8*, pp. 3389–3401, 2020, ISSN: 2473-4209. DOI: 10.1002/mp.14189.

[116] N. Talebloo, M. Gudi, N. Robertson, and P. Wang, « Magnetic particle imaging: current applications in biomedical research », *Journal of magnetic resonance imaging: JMRI*, vol. 51, *6*, pp. 1659–1668, 2020, ISSN: 1522-2586. DOI: 10.1002/jmri.26875.

[117] Y.-J. Park, D. Choi, J. Y. Choi, and S. H. Hyun, « Performance evaluation of a deep learning system for differential diagnosis of lung cancer with conventional CT and FDG PET/CT using transfer learning and metadata », *Clinical Nuclear*

Medicine, vol. 46, *8*, pp. 635–640, 2021, ISSN: 1536-0229. DOI: 10.1097/RLU.0000000000003661.

[118] D. Mu, J. Bai, W. Chen, H. Yu, J. Liang, K. Yin, H. Li, Z. Qing, K. He, H.-Y. Yang, J. Zhang, Y. Yin, H. W. McLellan, U. J. Schoepf, and B. Zhang, « Calcium scoring at coronary CT angiography using deep learning », *Radiology*, vol. 302, *2*, pp. 309–316, 2022, ISSN: 1527-1315. DOI: 10.1148/radiol.2021211483.

[119] S. Lee and R. M. Summers, « Clinical artificial intelligence applications in radiology: chest and abdomen », *Radiologic Clinics of North America*, vol. 59, *6*, pp. 987–1002, 2021, ISSN: 1557-8275. DOI: 10.1016/j.rcl.2021.07.001.

[120] P. Hamelmann, R. Vullings, A. F. Kolen, J. W. M. Bergmans, J. O. E. H. van Laar, P. Tortoli, and M. Mischi, « Doppler ultrasound technology for fetal heart rate monitoring: a review », *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 67, *2*, pp. 226–238, 2020, ISSN: 1525-8955. DOI: 10.1109/TUFFC.2019.2943626.

[121] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, « Deep learning techniques for medical image segmentation: achievements and challenges », *Journal of Digital Imaging*, vol. 32, *4*, pp. 582–596, 2019, ISSN: 1618-727X. DOI: 10.1007/s10278-019-00227-x.

[122] « The multimodal brain tumor image segmentation benchmark (BRATS) », *IEEE Transactions on Medical Imaging*, vol. 34, *10*, pp. 1993–2024, 2015, ISSN: 1558-254X. DOI: 10.1109/TMI.2014.2377694.

[123] G. Ras, N. Xie, M. v. Gerven, and D. Doran, « Explainable deep learning: a field guide for the uninitiated », *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–396, 2022, ISSN: 1076-9757. DOI: 10.1613/jair.1.13200.

[124] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu, *A survey of uncertainty in deep neural networks*, 2022. DOI: 10.48550/arXiv.2107.03342. arXiv: 2107.03342[cs,stat].

[125] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, *Geometric deep learning: grids, groups, graphs, geodesics, and gauges*, 2021. DOI: 10.48550/arXiv.2104.13478. arXiv: 2104.13478[cs,stat].

[126] K. T. Gao, V. Pedoia, K. A. Young, F. Kogan, M. F. Koff, G. E. Gold, H. G. Potter, and S. Majumdar, « Multiparametric MRI characterization of knee articular cartilage and subchondral bone shape in collegiate basketball players », *Journal of Orthopaedic Research: Official Publication of the Orthopaedic Research Society*, vol. 39, *7*, pp. 1512–1522, 2021, ISSN: 1554-527X. DOI: 10.1002/jor.24851.

[127] M. J. Paldino, E. Yang, J. Y. Jones, N. Mahmood, A. Sher, W. Zhang, S. Hayatghaibi, R. Krishnamurthy, and V. Seghers, « Comparison of the diagnostic accuracy of PET/MRI to PET/CT-acquired FDG brain exams for seizure focus detection: a prospective study », *Pediatric Radiology*, vol. 47, *11*, pp. 1500–1507, 2017, ISSN: 1432-1998. DOI: 10.1007/s00247-017-3888-8.

[128] O. M. Navarro, « Magnetic resonance imaging of pediatric soft-tissue vascular anomalies », *Pediatric Radiology*, vol. 46, *6*, pp. 891–901, 2016, ISSN: 1432-1998. DOI: 10.1007/s00247-016-3567-1.

[129] L. He, H. Li, J. Wang, M. Chen, E. Gozdas, J. R. Dillman, and N. A. Parikh, « A multi-task, multi-stage deep transfer learning model for early prediction of neurodevelopment in very preterm infants », *Scientific Reports*, vol. 10, *1*, p. 15 072, 2020, ISSN: 2045-2322. DOI: 10.1038/s41598-020-71914-x.

[130] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, F. C. Kitamura, H. H. Thodberg, L. Chen, G. Shih, K. Andriole, M. D. Kohli, B. J. Erickson, and A. E. Flanders, « The RSNA pediatric bone age machine learning challenge », *Radiology*, vol. 290, *2*, pp. 498–503, 2019, ISSN: 1527-1315. DOI: 10.1148/radiol.2018180736.

[131] Z. Wu, Z. Lin, L. Li, H. Pan, G. Chen, Y. Fu, and Q. Qiu, « Deep learning for classification of pediatric otitis media », *The Laryngoscope*, vol. 131, *7*, E2344–E2351, 2021, ISSN: 1531-4995. DOI: 10.1002/lary.29302.

[132] S. Liu, K.-H. Thung, W. Lin, P.-T. Yap, and D. Shen, « Real-time quality assessment of pediatric MRI via semi-supervised deep nonlocal residual neural networks », *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 2020, ISSN: 1941-0042. DOI: 10.1109/TIP.2020.2992079.

[133] D. G. Kendall, « Shape manifolds, procrustean metrics, and complex projective spaces », *Bulletin of the London Mathematical Society*, vol. 16, *2*, pp. 81–121, 1984, ISSN: 0024-6093. DOI: 10.1112/blms/16.2.81.

[134] H. Gray and W. H. Lewis, *Anatomy of the human body*, in collab. with Harold B. Lee Library. Philadelphia : Lea & Febiger, 1918, 1404 pp.

[135] F. Flandry and G. Hommel, « Normal anatomy and biomechanics of the knee », *Sports Medicine and Arthroscopy Review*, vol. 19, *2*, pp. 82–92, 2011, ISSN: 1538-1951. DOI: 10.1097/JSA.0b013e318210c0aa.

[136] F. E. Zajac, « Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control », *Critical Reviews in Biomedical Engineering*, vol. 17, *4*, pp. 359–411, 1989, ISSN: 0278-940X.

[137] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, « OpenSim: open-source software to create and analyze dynamic simulations of movement », *IEEE Transactions on Biomedical Engineering*, vol. 54, *11*, pp. 1940–1950, 2007, ISSN: 1558-2531. DOI: 10.1109/TBME.2007.901024.

[138] J.-R. Fouefack, « Towards a framework for multi class statistical modelling of shape, intensity, and kinematics in medical images », These de doctorat, Ecole nationale supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, 2021.

[139] J.-J. Jacq, T. Cresson, V. Burdin, and C. Roux, « Performing accurate joint kinematics from 3-d in vivo image sequences through consensus-driven simultaneous registration », *IEEE Transactions on Biomedical Engineering*, vol. 55, *5*, pp. 1620–1633, 2008, ISSN: 1558-2531. DOI: 10.1109/TBME.2008.918580.

[140] D. Breton, V. Burdin, J. Leboucher, and O. Remy-Neris, « Study of the joint configuration of the knee using a morpho-functional analysis », *IRBM*, Biomedical image segmentation using variational and statistical approaches, vol. 35, *1*, pp. 53–57, 2014, ISSN: 1959-0318. DOI: 10.1016/j.irbm.2013.12.006.

[141] A. **Boutillon**, A. Salhi, V. Burdin, and B. Borotikar, « Anatomically parameterized statistical shape model: explaining morphometry through statistical learning », *IEEE Transactions on Biomedical Engineering*, vol. 69, *9*, pp. 2733–2744, 2022, ISSN: 1558-2531. DOI: 10.1109/TBME.2022.3152833.

[142] A. Salhi, V. Burdin, A. **Boutillon**, S. Brochard, T. Mutsvangwa, and B. Borotikar, « Statistical shape modeling approach to predict missing scapular bone », *Annals of Biomedical Engineering*, vol. 48, *1*, pp. 367–379, 2020, ISSN: 1573-9686. DOI: 10.1007/s10439-019-02354-6.

[143] A. Salhi, « Towards a combined statistical shape and musculoskeletal modeling framework for pediatric shoulder joint », These de doctorat, Ecole nationale supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, 2019.

[144] A. Salhi, V. Burdin, T. Mutsvangwa, S. Sivarasu, S. Brochard, and B. Borotikar, « Subject-specific shoulder muscle attachment region prediction using statistical shape models: a validity study », *in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 1640–1643. DOI: 10.1109/EMBC.2017.8037154.

[145] S. Allaire, V. Burdin, J.-j. Jacq, G. Moineau, E. Stindel, and C. Roux, « Robust quadric fitting and mensuration comparison in a mapping space applied to 3D morphological characterization of articular surfaces », *in 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2007, pp. 972–975. DOI: 10.1109/ISBI.2007.357016.

[146] J.-R. Foueface, B. Borotikar, M. Lüthi, T. S. Douglas, V. Burdin, and T. E. M. Mutsvangwa, *Dynamic multi feature-class gaussian process models*, 2021. DOI: 10.48550/arXiv.2112.04495. arXiv: 2112.04495[cs,eess].

[147] M. Ünal, O. Akkuş, and R. E. Marcus, « Fundamentals of musculoskeletal biomechanics », *in Musculoskeletal Research and Basic Science*, F. Korkusuz, Ed., Cham: Springer International Publishing, 2016, pp. 15–36, ISBN: 978-3-319-20777-3. DOI: 10.1007/978-3-319-20777-3_2.

[148] J. R. Leschied and S. B. Soliman, « Pediatric musculoskeletal trauma: special considerations », *Seminars in Roentgenology*, Imaging of Musculoskeletal Trauma, vol. 56, *1*, pp. 70–78, 2021, ISSN: 0037-198X. DOI: 10.1053/j.ro.2020.07.017.

[149] E. W. Hubbard and A. I. Riccio, « Pediatric orthopedic trauma: an evidence-based approach », *The Orthopedic Clinics of North America*, vol. 49, *2*, pp. 195–210, 2018, ISSN: 1558-1373. DOI: 10.1016/j.ocl.2017.11.008.

[150] A. B. Meyers, « Physeal bridges: causes, diagnosis, characterization and post-treatment imaging », *Pediatric Radiology*, vol. 49, *12*, pp. 1595–1609, 2019, ISSN: 1432-1998. DOI: 10.1007/s00247-019-04461-x.

[151] J. J. Dowling, H. D. Gonorazky, R. D. Cohn, and C. Campbell, « Treating pediatric neuromuscular disorders: the future is now », *American Journal of Medical*

Genetics. Part a, vol. 176, *4*, pp. 804–841, 2018, ISSN: 1552-4825. DOI: 10.1002/ajmg.a.38418.

[152] K. Ecklund and D. Jaramillo, « Imaging of growth disturbance in children », *Radiologic Clinics of North America*, vol. 39, *4*, pp. 823–841, 2001, ISSN: 0033-8389. DOI: 10.1016/s0033-8389(05)70313-4.

[153] P. Mary, L. Servais, and R. Vialle, « Neuromuscular diseases: diagnosis and management », *Orthopaedics & traumatology, surgery & research: OTSR*, vol. 104, *1*, S89–S95, 2018, ISSN: 1877-0568. DOI: 10.1016/j.otsr.2017.04.019.

[154] J. Ropars, F. Gravot, D. Ben Salem, F. Rousseau, S. Brochard, and C. Pons, « Muscle MRI: a biomarker of disease severity in duchenne muscular dystrophy? a systematic review », *Neurology*, vol. 94, *3*, pp. 117–133, 2020, ISSN: 1526-632X. DOI: 10.1212/WNL.0000000000008811.

[155] K. Vitrikas, H. Dalton, and D. Breish, « Cerebral palsy: an overview », *American Family Physician*, vol. 101, *4*, pp. 213–220, 2020, ISSN: 1532-0650.

[156] L. A. Koman, B. P. Smith, and J. S. Shilt, « Cerebral palsy », *Lancet (London, England)*, vol. 363, *9421*, pp. 1619–1631, 2004, ISSN: 1474-547X. DOI: 10.1016/S0140-6736(04)16207-7.

[157] C. L. Piccolo, M. Galluzzo, S. Ianniello, M. Trinci, A. Russo, E. Rossi, M. Zeccolini, A. Laporta, G. Guglielmi, and V. Miele, « Pediatric musculoskeletal injuries: role of ultrasound and magnetic resonance imaging », *Musculoskeletal Surgery*, vol. 101, pp. 85–102, Suppl 1 2017, ISSN: 2035-5114. DOI: 10.1007/s12306-017-0452-5.

[158] K. Rosendahl and P. J. Strouse, « Sports injury of the pediatric musculoskeletal system », *La Radiologia Medica*, vol. 121, *5*, pp. 431–441, 2016, ISSN: 1826-6983. DOI: 10.1007/s11547-015-0615-0.

[159] P. A. DeHeer, « Equinus and lengthening techniques », *Clinics in Podiatric Medicine and Surgery*, vol. 34, *2*, pp. 207–227, 2017, ISSN: 1558-2302. DOI: 10.1016/j.cpm.2016.10.008.

[160] J. Charles, S. D. Scutter, and J. Buckley, « Static ankle joint equinus: toward a standard definition and diagnosis », *Journal of the American Podiatric Medical Association*, vol. 100, *3*, pp. 195–203, 2010, ISSN: 1930-8264. DOI: 10.7547/1000195.

[161] D. Murphy, M. Raza, H. Khan, D. M. Eastwood, and Y. Gelfer, « What is the optimal treatment for equinus deformity in walking-age children with clubfoot? a systematic review », *EFORT open reviews*, vol. 6, *5*, pp. 354–363, 2021, ISSN: 2058-5241. DOI: 10.1302/2058-5241.6.200110.

[162] D. I. Zafeiriou and K. Psychogiou, « Obstetrical brachial plexus palsy », *Pediatric Neurology*, vol. 38, *4*, pp. 235–242, 2008, ISSN: 0887-8994. DOI: 10.1016/j.pediatrneurol.2007.09.013.

[163] S. P. Chauhan, S. B. Blackwell, and C. V. Ananth, « Neonatal brachial plexus palsy: incidence, prevalence, and temporal trends », *Seminars in Perinatology*, vol. 38, *4*, pp. 210–218, 2014, ISSN: 1558-075X. DOI: 10.1053/j.semperi.2014.04.007.

[164] A. F. Hoeksma, A. M. Ter Steeg, P. Dijkstra, R. G. H. H. Nelissen, A. Beelen, and B. A. de Jong, « Shoulder contracture and osseous deformity in obstetrical brachial plexus injuries », *The Journal of Bone and Joint Surgery. American Volume*, vol. 85, *2*, pp. 316–322, 2003, ISSN: 0021-9355. DOI: 10.2106/00004623-200302000-00020.

[165] C. Pons, F. T. Sheehan, H. S. Im, S. Brochard, and K. E. Alter, « Shoulder muscle atrophy and its relation to strength loss in obstetrical brachial plexus palsy », *Clinical biomechanics (Bristol, Avon)*, vol. 48, pp. 80–87, 2017, ISSN: 0268-0033. DOI: 10.1016/j.clinbiomech.2017.07.010.

[166] C. Pons, B. Borotikar, M. Garetier, V. Burdin, D. Ben Salem, M. Lempereur, and S. Brochard, « Quantifying skeletal muscle volume and shape in humans using MRI: a systematic review of validity and reliability », *PloS One*, vol. 13, *11*, e0207847, 2018, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0207847.

[167] H. Khalatbari, M. T. Parisi, N. Kwatra, D. J. Harrison, and B. L. Shulkin, « Pediatric musculoskeletal imaging: the indications for and applications of PET/computed tomography », *PET Clinics*, vol. 14, *1*, pp. 145–174, 2019, ISSN: 1556-8598. DOI: 10.1016/j.cpet.2018.08.008.

[168] R. B. Sequeiros, J.-J. Sinikumpu, R. Ojala, J. Järvinen, and J. Fritz, « Pediatric musculoskeletal interventional MRI », *Topics in magnetic resonance imaging: TMRI*, vol. 27, *1*, pp. 39–44, 2018, ISSN: 1536-1004. DOI: 10.1097/RMR.0000000000000143.

[169] L. Barbuto, M. Di Serafino, N. Della Vecchia, G. Rea, F. Esposito, N. Vezzali, F. Ferro, M. G. Caprio, E. A. Vola, V. Romeo, and G. Vallone, « Pediatric musculoskeletal ultrasound: a pictorial essay », *Journal of Ultrasound*, vol. 22, *4*, pp. 491–502, 2019, ISSN: 1876-7931. DOI: 10.1007/s40477-018-0337-y.

[170] M. A. DiPietro and J. R. Leschied, « Pediatric musculoskeletal ultrasound », *Pediatric Radiology*, vol. 47, *9*, pp. 1144–1154, 2017, ISSN: 1432-1998. DOI: 10.1007/s00247-017-3919-5.

[171] W. R. Walter, L. H. Goldman, and Z. S. Rosenberg, « Pitfalls in MRI of the developing pediatric ankle », *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, vol. 41, *1*, pp. 210–223, 2021, ISSN: 1527-1323. DOI: 10.1148/rg.2021200088.

[172] V. M. Gylys-Morin, « MR imaging of pediatric musculoskeletal inflammatory and infectious disorders », *Magnetic Resonance Imaging Clinics of North America*, vol. 6, *3*, pp. 537–559, 1998, ISSN: 1064-9689.

[173] M. Garetier, B. Borotikar, K. Makki, S. Brochard, F. Rousseau, and D. Ben Salem, « Dynamic MRI for articulating joint evaluation on 1.5T and 3.0T scanners: setup, protocols, and real-time sequences », *Insights into Imaging*, vol. 11, *1*, p. 66, 2020, ISSN: 1869-4101. DOI: 10.1186/s13244-020-00868-5.

[174] B. Borotikar, M. Lempereur, M. Lelievre, V. Burdin, D. B. Salem, and S. Brochard, « Dynamic MRI to quantify musculoskeletal motion: a systematic review of concurrent validity and reliability, and perspectives for evaluation of musculoskeletal disorders », *PLOS ONE*, vol. 12, *12*, e0189587, 2017, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0189587.

[175] J. R. Leschied and K. G. Udager, « Imaging of the pediatric knee », *Seminars in Musculoskeletal Radiology*, vol. 21, *2*, pp. 137–146, 2017, ISSN: 1098-898X. DOI: 10.1055/s-0037-1599205.

[176] Y. Han and G. Wang, « Skeletal bone age prediction based on a deep residual network with spatial transformer », *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105 754, 2020, ISSN: 1872-7565. DOI: 10.1016/j.cmpb.2020.105754.

[177] A. Tsai, « Anatomical landmark localization via convolutional neural networks for limb-length discrepancy measurements », *Pediatric Radiology*, vol. 51, *8*, pp. 1431–1447, 2021, ISSN: 1432-1998. DOI: 10.1007/s00247-021-05004-z.

[178] D. Wei, Q. Wu, X. Wang, M. Tian, and B. Li, « Accurate instance segmentation in pediatric elbow radiographs », *Sensors (Basel, Switzerland)*, vol. 21, *23*, p. 7966, 2021, ISSN: 1424-8220. DOI: 10.3390/s21237966.

[179] J. Castiglione, E. Somasundaram, L. A. Gilligan, A. T. Trout, and S. Brady, « Automated segmentation of abdominal skeletal muscle on pediatric CT scans using deep learning », *Radiology. Artificial Intelligence*, vol. 3, *2*, e200130, 2021, ISSN: 2638-6100. DOI: 10.1148/ryai.2021200130.

[180] F. Rosenblatt, « The perceptron: a probabilistic model for information storage and organization in the brain », *Psychological Review*, vol. 65, *6*, pp. 386–408, 1958, ISSN: 1939-1471. DOI: 10.1037/h0042519.

[181] K. Fukushima, « Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position », *Biological Cybernetics*, vol. 36, *4*, pp. 193–202, 1980, ISSN: 1432-0770. DOI: 10.1007/BF00344251.

[182] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, « Backpropagation applied to handwritten zip code recognition », *Neural Computation*, vol. 1, *4*, pp. 541–551, 1989, ISSN: 0899-7667. DOI: 10.1162/neco.1989.1.4.541.

[183] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, « Learning representations by back-propagating errors », *Nature*, vol. 323, *6088*, pp. 533–536, 1986, ISSN: 1476-4687. DOI: 10.1038/323533a0.

[184] H. Robbins and S. Monro, « A stochastic approximation method », *The Annals of Mathematical Statistics*, vol. 22, *3*, pp. 400–407, 1951, ISSN: 0003-4851. URL: https://www.jstor.org/stable/2236626.

[185] G. Cybenko, « Approximation by superpositions of a sigmoidal function », *Mathematics of Control, Signals and Systems*, vol. 2, *4*, pp. 303–314, 1989, ISSN: 1435-568X. DOI: 10.1007/BF02551274.

[186] K. Hornik, M. Stinchcombe, and H. White, « Multilayer feedforward networks are universal approximators », *Neural Networks*, vol. 2, *5*, pp. 359–366, 1989, ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8.

[187] A. Pinkus, « Approximation theory of the MLP model in neural networks », *Acta Numerica*, vol. 8, pp. 143–195, 1999, ISSN: 0962-4929. DOI: 10.1017/S0962492900002919.

[188] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, « Multilayer feedforward networks with a nonpolynomial activation function can approximate any function », *Neural Networks*, vol. 6, *6*, pp. 861–867, 1993, ISSN: 0893-6080. DOI: 10.1016/S0893-6080(05)80131-5.

[189] X. Glorot, A. Bordes, and Y. Bengio, « Deep sparse rectifier neural networks », *in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15, JMLR Workshop and Conference Proceedings, 2011, pp. 315–323. URL: https://proceedings.mlr.press/v15/glorot11a.html.

[190] A. Krizhevsky, I. Sutskever, and G. E. Hinton, « ImageNet classification with deep convolutional neural networks », *in Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[191] M. Tan and Q. Le, « EfficientNet: rethinking model scaling for convolutional neural networks », *in Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html.

[192] K. Kamnitsas, E. Ferrante, S. Parisot, C. Ledig, A. V. Nori, A. Criminisi, D. Rueckert, and B. Glocker, « DeepMedic for brain tumor segmentation », *in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, B. Menze, O. Maier, M. Reyes, S. Winzeck, and H. Handels, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 138–149, ISBN: 978-3-319-55524-9. DOI: 10.1007/978-3-319-55524-9_14.

[193] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, « Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features », *Scientific Data*, vol. 4, p. 170 117, 2017, ISSN: 2052-4463. DOI: 10.1038/sdata.2017.117.

[194] M. D. Zeiler, *ADADELTA: an adaptive learning rate method*, 2012. DOI: 10.48550/arXiv.1212.5701. arXiv: 1212.5701[cs].

[195] J. Duchi, E. Hazan, and Y. Singer, « Adaptive subgradient methods for online learning and stochastic optimization », *Journal of Machine Learning Research*, vol. 12, *61*, pp. 2121–2159, 2011, ISSN: 1533-7928. URL: http://jmlr.org/papers/v12/duchi11a.html.

[196] D. P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, 2017. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980[cs].

[197] X. Glorot and Y. Bengio, « Understanding the difficulty of training deep feedforward neural networks », *in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html.

[198] P. Ramachandran, B. Zoph, and Q. V. Le, *Searching for activation functions*, 2017. DOI: 10.48550/arXiv.1710.05941. arXiv: 1710.05941[cs].

[199] J. Long, E. Shelhamer, and T. Darrell, « Fully convolutional networks for semantic segmentation », *in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965.

[200] V. Badrinarayanan, A. Kendall, and R. Cipolla, « SegNet: a deep convolutional encoder-decoder architecture for image segmentation », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, *12*, pp. 2481–2495, 2017, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2016.2644615.

[201] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, « DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, *4*, pp. 834–848, 2018, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2017.2699184.

[202] H. Noh, S. Hong, and B. Han, « Learning deconvolution network for semantic segmentation », *in 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1520–1528. DOI: 10.1109/ICCV.2015.178.

[203] G. Yang and D. Ramanan, « Volumetric correspondence networks for optical flow », *in Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran

Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/bbf94b34eb32268ada57a3be5062fe7d-Abstract.html.

[204] W.-N. Lie, H.-T. Chiu, and J.-C. Chiang, « Disparity map estimation from stereo image pair using deep convolutional network », *in 2020 International Computer Symposium (ICS)*, 2020, pp. 365–369. DOI: 10.1109/ICS51289.2020.00079.

[205] Y. Cao, A. Vassantachart, J. C. Ye, C. Yu, D. Ruan, K. Sheng, Y. Lao, Z. L. Shen, S. Balik, S. Bian, G. Zada, A. Shiu, E. L. Chang, and W. Yang, « Automatic detection and segmentation of multiple brain metastases on magnetic resonance image using asymmetric UNet architecture », *Physics in Medicine and Biology*, vol. 66, *1*, p. 015 003, 2021, ISSN: 1361-6560. DOI: 10.1088/1361-6560/abca53.

[206] E. Schnider, A. Huck, M. Toranelli, G. Rauter, M. Müller-Gerbl, and P. C. Cattin, « Improved distinct bone segmentation from upper-body CT using binary-prediction-enhanced multi-class inference. », *International Journal of Computer Assisted Radiology and Surgery*, 2022, ISSN: 1861-6429. DOI: 10.1007/s11548-022-02650-y.

[207] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, « nnU-net: a self-configuring method for deep learning-based biomedical image segmentation », *Nature Methods*, vol. 18, *2*, pp. 203–211, 2021, ISSN: 1548-7105. DOI: 10.1038/s41592-020-01008-z.

[208] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, « 3D U-Net: learning dense volumetric segmentation from sparse annotation », *in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 424–432, ISBN: 978-3-319-46723-8. DOI: 10.1007/978-3-319-46723-8_49.

[209] N. Zettler and A. Mastmeyer, *Comparison of 2D vs. 3D U-Net organ segmentation in abdominal 3D CT images*, 2021. DOI: 10.48550/arXiv.2107.04062. arXiv: 2107.04062[cs,eess].

[210] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, « How does batch normalization help optimization? », *in Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018. URL: https://papers.nips.cc/paper/2018/hash/905056c1ac1dad141560467e0a99e1cf-Abstract.html.

[211] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, « Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations », *Deep learning in medical image analysis and multimodal learning for clinical decision support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, held in conjunction with MICCAI 2017 Quebec City*, vol. 2017, pp. 240–248, 2017. DOI: 10.1007/978-3-319-67558-9_28.

[212] S. Jadon, « A survey of loss functions for semantic segmentation », *in 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2020, pp. 1–7. DOI: 10.1109/CIBCB48159.2020.9277638.

[213] D. Karimi and S. E. Salcudean, « Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks », *IEEE Transactions on Medical Imaging*, vol. 39, *2*, pp. 499–513, 2020, ISSN: 1558-254X. DOI: 10.1109/TMI.2019.2930068.

[214] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, « Pytorch: an imperative style, high-performance deep learning library », *in Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

[215] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, « TensorFlow: a system for large-scale machine learning », *in Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, ser. OSDI'16, USA: USENIX Association, 2016, pp. 265–283, ISBN: 978-1-931971-33-1. URL: https://www.tensorflow.org/.

[216] F. Chollet, *Deep Learning with Python*, 1st. USA: Manning Publications Co., 2017, ISBN: 978-1-61729-443-3. URL: https://keras.io/.

[217] P. Salembier, A. Oliveras, and L. Garrido, « Antiextensive connected operators for image and sequence processing », *IEEE Transactions on Image Processing*, vol. 7, *4*, pp. 555–570, 1998, ISSN: 1941-0042. DOI: 10.1109/83.663500.

[218] A. Reinke, M. D. Tizabi, C. H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, S. Bakas, P. Bankhead, A. Benis, M. J. Cardoso, V. Cheplygina, E. Christodoulou, B. Cimini, G. S. Collins, K. Farahani, B. van Ginneken, B. Glocker, P. Godau, F. Hamprecht, D. A. Hashimoto, D. Heckmann-Nötzel, M. M. Hoffman, M. Huisman, F. Isensee, P. Jannin, C. E. Kahn, A. Karargyris, A. Karthikesalingam, B. Kainz, E. Kavur, H. Kenngott, J. Kleesiek, T. Kooi, M. Kozubek, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, D. Moher, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, M. A. Noyan, J. Petersen, G. Polat, N. Rajpoot, M. Reyes, N. Rieke, M. Riegler, H. Rivaz, J. Saez-Rodriguez, C. S. Gutierrez, J. Schroeter, A. Saha, S. Shetty, M. van Smeden, B. Stieltjes, R. M. Summers, A. A. Taha, S. A. Tsaftaris, B. Van Calster, G. Varoquaux, M. Wiesenfarth, Z. R. Yaniv, A. Kopp-Schneider, P. Jäger, and L. Maier-Hein, *Common limitations of image processing metrics: a picture story*, 2022. DOI: 10.48550/arXiv.2104.05642. arXiv: 2104.05642[cs,eess].

[219] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, C. Feldmann, A. F. Frangi, P. M. Full, B. van Ginneken, A. Hanbury, K. Honauer, M. Kozubek, B. A. Landman, K. März, O. Maier, K. Maier-Hein, B. H. Menze, H. Müller, P. F. Neher, W. Niessen, N. Rajpoot, G. C. Sharp, K. Sirinukunwattana, S. Speidel, C. Stock, D. Stoyanov, A. A. Taha, F. van der Sommen, C.-W. Wang, M.-A. Weber, G. Zheng, P. Jannin, and A. Kopp-Schneider, « Why rankings of biomedical image analysis competitions should be interpreted with care », *Nature Communications*, vol. 9, *1*, p. 5217, 2018, ISSN: 2041-1723. DOI: 10.1038/s41467-018-07619-7.

[220] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, « Dropout: a simple way to prevent neural networks from overfitting », *Journal of Machine Learning Research*, vol. 15, *56*, pp. 1929–1958, 2014. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[221] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems*. Dordrecht: Springer Netherlands, 1995, ISBN: 978-94-015-8480-7. DOI: 10.1007/978-94-015-8480-7.

[222] G. E. Hinton, *Learning distributed representations of concepts*, ser. Parallel distributed processing: Implications for psychology and neurobiology. New York, NY,

US: Clarendon Press/Oxford University Press, 1989, 46 pp., ISBN: 978-0-19-852178-5.

[223] A. Lewkowycz and G. Gur-Ari, « On the training dynamics of deep networks with l2 regularization », *in Proceedings of the 34th International Conference on Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., ser. NIPS'20, Curran Associates Inc., 2020, pp. 4790–4799, ISBN: 978-1-71382-954-6. URL: https://papers.nips.cc/paper/2020/hash/32fcc8cfe1fa4c77b5c58dafd36d1a98-Abstract.html.

[224] L. Folle, T. Meinderink, D. Simon, A.-M. Liphardt, G. Krönke, G. Schett, A. Kleyer, and A. Maier, « Deep learning methods allow fully automated segmentation of metacarpal bones to quantify volumetric bone mineral density », *Scientific Reports*, vol. 11, *1*, p. 9697, 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-89111-9.

[225] M. Antico, F. Sasazawa, M. Dunnhofer, S. M. Camps, A. T. Jaiprakash, A. K. Pandey, R. Crawford, G. Carneiro, and D. Fontanarosa, « Deep learning-based femoral cartilage automatic segmentation in ultrasound imaging for guidance in robotic knee arthroscopy », *Ultrasound in Medicine & Biology*, vol. 46, *2*, pp. 422–435, 2020, ISSN: 1879-291X. DOI: 10.1016/j.ultrasmedbio.2019.10.015.

[226] C. A. Neves, E. D. Tran, I. M. Kessler, and N. H. Blevins, « Fully automated preoperative segmentation of temporal bone structures from clinical CT scans », *Scientific Reports*, vol. 11, *1*, p. 116, 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-020-80619-0.

[227] J. Wang, Y. Lv, J. Wang, F. Ma, Y. Du, X. Fan, M. Wang, and J. Ke, « Fully automated segmentation in temporal bone CT with neural network: a preliminary assessment study », *BMC medical imaging*, vol. 21, *1*, p. 166, 2021, ISSN: 1471-2342. DOI: 10.1186/s12880-021-00698-x.

[228] E. Schnider, A. Horváth, G. Rauter, A. Zam, M. Müller-Gerbl, and P. C. Cattin, « 3D segmentation networks for excessive numbers of classes: distinct bone segmentation in upper bodies », *in Machine Learning in Medical Imaging*, M. Liu, P. Yan, C. Lian, and X. Cao, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 40–49, ISBN: 978-3-030-59861-7. DOI: 10.1007/978-3-030-59861-7_5.

[229] Q. Lin, M. Luo, R. Gao, T. Li, Z. Man, Y. Cao, and H. Wang, « Deep learning based automatic segmentation of metastasis hotspots in thorax bone SPECT images », *PloS One*, vol. 15, *12*, e0243253, 2020, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0243253.

[230] N. Moreau, C. Rousseau, C. Fourcade, G. Santini, L. Ferrer, M. Lacombe, C. Guillerminet, M. Campone, M. Colombie, M. Rubeaux, and N. Normand, « Deep learning approaches for bone and bone lesion segmentation on 18fdg PET/CT imaging in the context of metastatic breast cancer », *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2020, pp. 1532–1535, 2020, ISSN: 2694-0604. DOI: 10.1109/EMBC44109.2020.9175904.

[231] C. E. von Schacky, N. J. Wilhelm, V. S. Schäfer, Y. Leonhardt, F. G. Gassert, S. C. Foreman, F. T. Gassert, M. Jung, P. M. Jungmann, M. F. Russe, C. Mogler, C. Knebel, R. von Eisenhart-Rothe, M. R. Makowski, K. Woertler, R. Burgkart, and A. S. Gersing, « Multitask deep learning for segmentation and classification of primary bone tumors on radiographs », *Radiology*, vol. 301, *2*, pp. 398–406, 2021, ISSN: 1527-1315. DOI: 10.1148/radiol.2021204531.

[232] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, « 3D deeply supervised network for automated segmentation of volumetric medical images », *Medical Image Analysis*, Special Issue on the 2016 Conference on Medical Image Computing and Computer Assisted Intervention (Analog to MICCAI 2015), vol. 41, pp. 40–54, 2017, ISSN: 1361-8415. DOI: 10.1016/j.media.2017.05.001.

[233] A. **Boutillon**, B. Borotikar, V. Burdin, and P.-H. Conze, « Multi-structure bone segmentation in pediatric MR images with combined regularization from shape priors and adversarial network », *Artificial Intelligence in Medicine*, vol. 132, 2022, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2022.102364.

[234] A. **Boutillon**, B. Borotikar, V. Burdin, and P.-H. Conze, « Combining shape priors with conditional adversarial networks for improved scapula segmentation in MR images », *in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1164–1167. DOI: 10.1109/ISBI45749.2020.9098360.

[235] C. Biffi, J. J. Cerrolaza, G. Tarroni, W. Bai, A. de Marvao, O. Oktay, C. Ledig, L. Le Folgoc, K. Kamnitsas, G. Doumou, J. Duan, S. K. Prasad, S. A. Cook, D. P. O'Regan, and D. Rueckert, « Explainable anatomical shape analysis through deep

hierarchical generative models », *IEEE Transactions on Medical Imaging*, vol. 39, *6*, pp. 2088–2099, 2020, ISSN: 1558-254X. DOI: 10.1109/TMI.2020.2964499.

[236] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, « Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation », *Medical Image Analysis*, vol. 36, pp. 61–78, 2017, ISSN: 1361-8415. DOI: 10.1016/j.media.2016.10.004.

[237] Q. Zhang, Y. N. Wu, and S.-C. Zhu, « Interpretable convolutional neural networks », *in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8827–8836. DOI: 10.1109/CVPR.2018.00920.

[238] Z. Zhang, C. Wu, S. Coleman, and D. Kerr, « DENSE-Inception U-Net for medical image segmentation », *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105 395, 2020, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2020.105395.

[239] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, « Medical transformer: gated axial-attention for medical image segmentation », *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 36–46, ISBN: 978-3-030-87193-2. DOI: 10.1007/978-3-030-87193-2_4.

[240] J. Cheng, S. Tian, L. Yu, C. Gao, X. Kang, X. Ma, W. Wu, S. Liu, and H. Lu, « ResGANet: residual group attention network for medical image classification and segmentation », *Medical Image Analysis*, vol. 76, p. 102 313, 2022, ISSN: 1361-8423. DOI: 10.1016/j.media.2021.102313.

[241] A. A. Kalinin, V. I. Iglovikov, A. Rakhlin, and A. A. Shvets, « Medical image segmentation using deep neural networks with pre-trained encoders », *in Deep Learning Applications*, ser. Advances in Intelligent Systems and Computing, M. A. Wani, M. Kantardzic, and M. Sayed-Mouchaweh, Eds., Singapore: Springer, 2020, pp. 39–52, ISBN: 978-981-15-1816-4. DOI: 10.1007/978-981-15-1816-4_3.

[242] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, « ImageNet large scale visual recognition challenge », *International Journal of Computer Vision*, vol. 115, *3*, pp. 211–252, 2015, ISSN: 1573-1405. DOI: 10.1007/s11263-015-0816-y.

[243] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, « Image-to-image translation with conditional adversarial networks », *in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976. DOI: 10.1109/CVPR.2017.632.

[244] F. Wilcoxon, « Individual comparisons by ranking methods », *Biometrics Bulletin*, vol. 1, *6*, pp. 80–83, 1945, ISSN: 0099-4987. DOI: 10.2307/3001968.

[245] R. B. D'Agostino, « An omnibus test of normality for moderate and large size samples », *Biometrika*, vol. 58, *2*, pp. 341–348, 1971, ISSN: 0006-3444. DOI: 10.2307/2334522.

[246] R. D'Agostino and E. S. Pearson, « Tests for departure from normality. empirical results for the distributions of b_2 and $\sqrt{b_1}$ », *Biometrika*, vol. 60, *3*, pp. 613–622, 1973, ISSN: 0006-3444. DOI: 10.2307/2335012.

[247] W. H. Kruskal and W. A. Wallis, « Use of ranks in one-criterion variance analysis », *Journal of the American Statistical Association*, vol. 47, *260*, pp. 583–621, 1952, ISSN: 0162-1459. DOI: 10.1080/01621459.1952.10483441.

[248] S. Holm, « A simple sequentially rejective multiple test procedure », *Scandinavian Journal of Statistics*, vol. 6, *2*, pp. 65–70, 1979, ISSN: 0303-6898. URL: https://www.jstor.org/stable/4615733.

[249] M. Noori, A. Bahri, and K. Mohammadi, « Attention-guided version of 2D UNet for automatic brain tumor segmentation », *in 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2019, pp. 269–275. DOI: 10.1109/ICCKE48569.2019.8964956.

[250] T. Heimann, B. Morrison, M. Styner, M. Niethammer, and S. Warfield, « Segmentation of knee images: a grand challenge », *in MICCAI Workshop on Medical Image Analysis for the Clinic*, 2010, pp. 207–214.

[251] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, « Generalization and equilibrium in generative adversarial nets (GANs) », *in Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., vol. 70, PMLR, 2017, pp. 224–232. URL: https://proceedings.mlr.press/v70/arora17a.html.

[252] M. D. Kohli, R. M. Summers, and J. R. Geis, « Medical image data and datasets in the era of machine learning-whitepaper from the 2016 c-MIMI meeting dataset session », *Journal of Digital Imaging*, vol. 30, *4*, pp. 392–399, 2017, ISSN: 1618-727X. DOI: 10.1007/s10278-017-9976-3.

[253] A. **Boutillon**, P.-H. Conze, C. Pons, V. Burdin, and B. Borotikar, « Generalizable multi-task, multi-domain deep segmentation of sparse pediatric imaging datasets via multi-scale contrastive regularization and multi-joint anatomical priors », *Medical Image Analysis*, 2022, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102556.

[254] A. **Boutillon**, P.-H. Conze, C. Pons, V. Burdin, and B. Borotikar, « Multi-task, multi-domain deep segmentation with shared representations and contrastive regularization for sparse pediatric datasets », *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 239–249, ISBN: 978-3-030-87193-2. DOI: 10.1007/978-3-030-87193-2_23.

[255] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, « An image is worth 16x16 words: Transformers for image recognition at scale », *in International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=YicbFdNTTy.

[256] X. Hu, D. Zeng, X. Xu, and Y. Shi, « Semi-supervised contrastive learning for label-efficient medical image segmentation », *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 481–490, ISBN: 978-3-030-87196-3. DOI: 10.1007/978-3-030-87196-3_45.

[257] H. Dou, D. Karimi, C. K. Rollins, C. M. Ortinau, L. Vasung, C. Velasco-Annis, A. Ouaalam, X. Yang, D. Ni, and A. Gholipour, « A deep attentive convolutional neural network for automatic cortical plate segmentation in fetal MRI », *IEEE Transactions on Medical Imaging*, vol. 40, *4*, pp. 1123–1133, 2021, ISSN: 1558-254X. DOI: 10.1109/TMI.2020.3046579.

[258] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, *MobileNets: efficient convolutional neural networks for mobile vision applications*, 2017. DOI: 10.48550/arXiv.1704.04861. arXiv: 1704.04861[cs].

[259] F. Chollet, « Xception: deep learning with depthwise separable convolutions », *in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807. DOI: 10.1109/CVPR.2017.195.

[260] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, « Squeeze-and-excitation networks », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, *8*, pp. 2011–2023, 2020, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2019.2913372.

[261] N. Smirnov, « Table for estimating the goodness of fit of empirical distributions », *The Annals of Mathematical Statistics*, vol. 19, *2*, pp. 279–281, 1948, ISSN: 0003-4851. DOI: 10.1214/aoms/1177730256.

[262] A. N. Kolmogorov, « Sulla determinazione empirica di una legge di distribuzione », *Giornale dell'Istituto Italiano degli Attuari*, *4*, pp. 83–91, 1933.

[263] S. Moccia, E. De Momi, S. El Hadji, and L. S. Mattos, « Blood vessel segmentation algorithms — review of methods, datasets and evaluation metrics », *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 71–91, 2018, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2018.02.001.

[264] G. Quellec, H. Al Hajj, M. Lamard, P.-H. Conze, P. Massin, and B. Cochener, « ExplAIn: explanatory artificial intelligence for diabetic retinopathy diagnosis », *Medical Image Analysis*, vol. 72, p. 102 118, 2021, ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102118.

[265] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, « Explainable AI: a review of machine learning interpretability methods », *Entropy*, vol. 23, *1*, p. 18, 2021, ISSN: 1099-4300. DOI: 10.3390/e23010018.

[266] C. Shen, P. Wang, H. R. Roth, D. Yang, D. Xu, M. Oda, W. Wang, C.-S. Fuh, P.-T. Chen, K.-L. Liu, W.-C. Liao, and K. Mori, « Multi-task federated learning for heterogeneous pancreas segmentation », *in Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning*, C. Oyarzun Laura, M. J. Cardoso, M. Rosen-Zvi, G. Kaissis, M. G. Linguraru, R. Shekhar, S. Wesarg, M.

Erdt, K. Drechsler, Y. Chen, S. Albarqouni, S. Bakas, B. Landman, N. Rieke, H. Roth, X. Li, D. Xu, M. Gabrani, E. Konukoglu, M. Guindy, D. Rueckert, A. Ziller, D. Usynin, and J. Passerat-Palmbach, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 101–110, ISBN: 978-3-030-90874-4. DOI: 10.1007/978-3-030-90874-4_10.

[267] F. Gao, M. Hu, M.-E. Zhong, S. Feng, X. Tian, X. Meng, M.-l. Ni-jia-ti, Z. Huang, M. Lv, T. Song, X. Zhang, X. Zou, and X. Wu, « Segmentation only uses sparse annotations: unified weakly and semi-supervised learning in medical images », *Medical Image Analysis*, vol. 80, p. 102 515, 2022, ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102515.

[268] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, « UNETR: Transformers for 3D medical image segmentation », *in 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1748–1758. DOI: 10.1109/WACV51458.2022.00181.

[269] A. Fernandez-Quilez, « Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability », *AI and Ethics*, 2022, ISSN: 2730-5961. DOI: 10.1007/s43681-022-00161-9.

[270] R. Selvan, N. Bhagwat, L. F. Wolff Anthony, B. Kanding, and E. B. Dam, « Carbon footprint of selecting and training deep learning models for medical image analysis », *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., ser. Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2022, pp. 506–516, ISBN: 978-3-031-16443-9. DOI: 10.1007/978-3-031-16443-9_49.

[271] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, « Transformers in medical image analysis: a review », *Intelligent Medicine*, 2022, ISSN: 2667-1026. DOI: 10.1016/j.imed.2022.07.002.

NOMENCLATURE

This section provides a list of abbreviations and mathematical notations employed throughout this thesis.

Abbreviations

2D, 3D Two-, Three-Dimension

CT Computed Tomography

PET Positron Emission Tomography

SPECT Single-Photon Emission Computed Tomography

ALARA As Low As Reasonably Achievable

MR, MRI Magnetic Resonance (Imaging)

TR, TE Repetition Time and Echo Time

FOV Field Of View

OBPP Obstetrical Brachial Plexus Palsy

CNN Convolutional Neural Network

SVM Support Vector Machine

SSM Statistical Shape Model

ReLU Rectified Linear Unit

SiLU Sigmoid Linear Unit

CE Cross-Entropy

BCE Binary Cross-Entropy

BN, DSBN (Domain-Specific) Batch Normalization

DSL Domain-Specific Layer

AE, MJAE (Multi-Joint) Auto-Encoder

MJSP Multi-Joint Shape Priors

SSC, MSC Single-, Multi-Scale Contrastive

ASSD Average Symmetric Surface Distance

MSSD Maximum Symmetric Surface Distance

RAVD Relative Absolute Volume Difference

CPU Central Processing Unit

GPU Graphics Processing Unit

Notations

Ω	Image grid
K	Number of domains
$\mathcal{I}_k$	k^{th} image domain [1]
$\mathcal{C}_k$	k^{th} label space
C_k	Number of structures of interest in $\mathcal{C}_k$
$\mathcal{D}_k$	k^{th} imaging dataset
n_k	Number of image in $\mathcal{D}_k$
x_i^k	i^{th} image of $\mathcal{D}_k$
y_i^k	Ground truth label associated with x_i^k
$\hat{y}_i^k$	Predicted label associated with x_i^k
S	Segmentation network
Θ	Shared parameters of S [2]
Γ	Domain-specific parameters of S
Θ_l	Convolution filter of the l^{th} layer
b_l	Bias of the l^{th} layer
$v_{i,l}^k$	Feature map of the l^{th} layer with input x_i^k

μ_l^k Domain-specific mini-batch mean of the l^{th} layer

σ_l^k Domain-specific mini-batch standard deviation of the l^{th} layer

β_l^k Domain-specific learnable shift of the l^{th} layer

γ_l^k Domain-specific learnable scale of the l^{th} layer

Λ_k Domain-specific batch normalization parameters of S

ρ Activation function

$u_{i,l}^k$ Activation map of the l^{th} layer with input x_i^k

u_i^k Activation map of the penultimate layer with input x_i^k

W_k Domain-specific 1×1 convolution of the final layer

b_k Domain-specific bias of the final layer

Ξ_k Domain-specific parameters of the final layer

F, G Shape encoder and decoder

Θ_F, Θ_G Shared parameters of F and G

Γ_F, Γ_G Domain-specific parameters of F and G

D Discriminator

Θ_D Parameters of D

$\mathcal{L}$ Loss function

$\mathcal{L}_{\text{CE}}$ Cross-entropy loss function

$\mathcal{L}_{\text{BCE}}$ Binary cross-entropy loss function

$\mathcal{L}_{\text{Dice}}$ Dice coefficient loss function

$\mathcal{L}_{\text{AE}}$, $\mathcal{L}_{\text{MJAE}}$ (Multi-joint) auto-encoder loss function

$\mathcal{L}_{\text{Shape}}$, $\mathcal{L}_{\text{MJSP}}$ (Multi-joint) shape priors regularization

$\mathcal{L}_{\text{D}}$ Discriminator loss function

$\mathcal{L}_{\text{Adv}}$ Adversarial regularization

$\mathcal{S}$ Set of layers indices corresponding to the different spatial scale of S

$\mathcal{P}_i^k$	Set of indexes of all images from the same domain as x_i^k
$z_{i,s}^k$	Embedding of x_i^k at scale s
$\mathcal{L}_{\text{MSC}}, \mathcal{L}_{\text{SSC}}$	Multi-scale and single scale contrastive regularization
τ	Temperature hyper-parameter of $\mathcal{L}_{\text{MSC}}$ and $\mathcal{L}_{\text{SSC}}$
$\lambda_{1,2,3}$	Regularization weighting hyper-parameters
α	Learning rate hyper-parameter
p_{data}	True probability distribution of the data
p_{model}	Probability distribution parameterized by the segmentation model
GT, P	Ground truth and predicted 3D segmentation masks
S_{GT}, S_P	Surface voxels of GT and P
∇	Gradient operator
$*$	Convolution product
$\cdot$	Scalar product
$\|.\|_2$	Euclidean or L_2 norm
$\|.\|$	Cardinality of a set

1. In Chapters 3, 4, and 5, the index k is omitted because only one domain is considered.
2. In Chapters 3, 4, and 5, shared and domain-specific parameters are not dissociated because only one domain is considered.

PUBLICATIONS

This section presents the outputs of research projects conducted during the thesis for which the author is the primary author or a collaborator.

Journal articles

- A. **Boutillon**, P.-H. Conze, C. Pons, V. Burdin, and B. Borotikar, « Generalizable multi-task, multi-domain deep segmentation of sparse pediatric imaging datasets via multi-scale contrastive regularization and multi-joint anatomical priors », *Medical Image Analysis*, 2022, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102556.
- A. **Boutillon**, A. Salhi, V. Burdin, and B. Borotikar, « Anatomically parameterized statistical shape model: explaining morphometry through statistical learning », *IEEE Transactions on Biomedical Engineering*, vol. 69, *9*, pp. 2733–2744, 2022, ISSN: 1558-2531. DOI: 10.1109/TBME.2022.3152833.
- A. **Boutillon**, B. Borotikar, V. Burdin, and P.-H. Conze, « Multi-structure bone segmentation in pediatric MR images with combined regularization from shape priors and adversarial network », *Artificial Intelligence in Medicine*, vol. 132, 2022, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2022.102364.
- A. Salhi, V. Burdin, A. **Boutillon**, S. Brochard, T. Mutsvangwa, and B. Borotikar, « Statistical shape modeling approach to predict missing scapular bone », *Annals of Biomedical Engineering*, vol. 48, *1*, pp. 367–379, 2020, ISSN: 1573-9686. DOI: 10.1007/s10439-019-02354-6.

Conference proceedings

- A. **Boutillon**, B. Borotikar, C. Pons, V. Burdin, and P.-H. Conze, « Multi-structure deep segmentation with shape priors and latent adversarial regulariza-

tion », *in 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 999–1002. DOI: 10.1109/ISBI48211.2021.9434104.

- A. **Boutillon**, P.-H. Conze, C. Pons, V. Burdin, and B. Borotikar, « Multi-task, multi-domain deep segmentation with shared representations and contrastive regularization for sparse pediatric datasets », *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 239–249, ISBN: 978-3-030-87193-2. DOI: 10.1007/978-3-030-87193-2_23.
- A. **Boutillon**, B. Borotikar, V. Burdin, and P.-H. Conze, « Combining shape priors with conditional adversarial networks for improved scapula segmentation in MR images », *in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1164–1167. DOI: 10.1109/ISBI45749.2020.9098360.

Conference abstracts

- A. Salhi, A. **Boutillon**, P.-H. Conze, G. Dardenne, V. Burdin, and B. Borotikar, « Segmentation automatique multi-structures: une approche par transfert d'apprentissage sur IRM de genou », presented at the Traitement et Analyse de l'Information Méthodes et Applications - TAIMA 2022, 2022.
- A. **Boutillon**, A. Salhi, R. Baily, M. Naffrechoux, S. Brochard, and B. Borotikar, « Statistical shape models (SSM) similarity comparison with vertex-wise Bhattacharya metric of the pediatric talus », presented at the XXVIII Congress of the International Society of Biomechanics (ISB), 2021.

Presentations

- A. **Boutillon**, « Multi-task, multi-domain deep segmentation with shared representations and contrastive regularization for sparse pediatric datasets », presented at the GDR ISIS workshop - Deep learning with weak or few labels in medical image analysis, Paris, France, 2022.

- A. **Boutillon**, « Anatomically parameterized statistical shape model: explaining morphometry through statistical learning », presented at the 2022 IEEE EMBS-SPS International Summer School on Biomedical Imaging, St-Jacut, France, 2022.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Modèles d'apprentissage profond régularisés pour la segmentation multi-anatomie en imagerie pédiatrique

Mot clés : Segmentation multi-domaine, Apprentissage multi-tâche, *A priori* de formes, Régularisation contrastive, Réseaux antagonistes, Système musculo-squelettique

Résumé : En imagerie médicale, la segmentation basée sur l'apprentissage profond permet de générer automatiquement des modèles anatomiques qui sont cruciaux pour l'évaluation morphologique. Cependant, la rareté des ressources d'imagerie pédiatrique peut entraîner une diminution de la précision et des performances de généralisation des réseaux de segmentation. Pour atténuer ces problèmes, notre première approche consiste en un nouveau schéma d'optimisation exploitant des *a priori* de formes visant à imposer des prédictions globalement cohérentes et un réseau antagoniste qui encourage des délimitations plus précises. Dans notre deuxième stratégie, nous concevons un nouveau réseau multi-tâche et multi-domaine optimisé sur des ensembles de données d'imagerie multi-anatomie. Pour améliorer la généralisation, nous démêlons les représentations des domaines en utilisant une régularisation contrastive et nous étendons les *a priori* de formes à l'apprentissage multi-anatomie. Nos contributions sont évaluées pour la segmentation osseuse de trois articulations (cheville, épaule, genou). Les méthodes proposées ont obtenu des résultats supérieurs ou égaux à ceux des modèles de l'état de l'art. Ces résultats ouvrent de nouvelles perspectives pour une utilisation collaborative des ressources d'imagerie pédiatrique et une meilleure gestion des troubles musculo-squelettiques.

Title: Regularized deep learning models for multi-anatomy segmentation in pediatric imaging

Keywords: Multi-domain segmentation, Multi-task learning, Shape priors, Contrastive regularization, Adversarial networks, Musculsokeletal system

Abstract: In medical imaging, segmentation using deep learning enables an automatic generation of anatomical models that are crucial for morphological evaluation. However, the scarcity of pediatric imaging resources may result in reduced accuracy and generalization performance of segmentation networks. To mitigate these issues, our first approach consists in a novel optimization scheme leveraging shape priors to enforce globally consistent predictions and an adversarial network to encourage precise delineations. In our second strategy, we design a novel multi-task, multi-domain network optimized over multi-anatomy imaging datasets. To improve generalizability, we disentangle the domains representations using a contrastive regularization and extend the shape priors to multi-anatomy learning. Our contributions are evaluated for the bone segmentation of three anatomical joints (ankle, knee, shoulder). The proposed methods performed either better or at par with state-of-the-art models. These results bring new perspectives towards a collaborative utilization of pediatric imaging resources and better management of musculoskeletal disorders.