Learning-based image synthesis for dynamic MRI

Claire Scavinner-Dorval

COLLEGES SCIENCES
BRETAGNE POUR L'INGENIEUR
LOIRE ET LE NUMERIQUE

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° 648
Sciences pour l'Ingénieur et le Numérique
Spécialité : *Signal, Image, Vision*

Par
Claire SCAVINNER- -DORVAL

Learning-based Image Synthesis for Dynamic MRI

Thèse présentée et soutenue à IMT Atlantique, Brest, le 25/11/2024
Unité de recherche : Laboratoire de Traitement de l'Information Médicale - UMR 1101
Thèse N° : 2024IMTA0437

Rapporteurs avant soutenance :

Vincent NOBLET	Ingénieur de recherche, CNRS
Ahror BELAID	Professeur, Université de Bajaia

Composition du Jury :

Président :	Nicolas PASSAT	Professeur, Université de Reims Champagne-Ardenne
Examinateurs :	Christelle PONS BECMEUR	Maître de conférence, HDR, UBO
	Ahror BELAID	Professeur, Université de Bajaia
	Vincent NOBLET	Ingénieur de recherche, CNRS
Dir. de thèse :	François ROUSSEAU	Professeur, IMT Atlantique
Co-dir. de thèse :	Douraied BEN SALEM	Professeur, UBO - CHRU de Brest

Invité(s) :

Rodolphe BAILLY	Kinésithérapeuthe, PhD, Fondation Ildys

Acknowledgement

Les recherches qui ont mené à ces résultats ont bénéficié de financements de l'ANR (AI4CHILD ANR-19-CHIA-0015-01), la Région Bretagne (projet DynMRI), Philips, la Fondation de l'Avenir, Paris, France, et de la Fondation Motrice, Paris, France. Ce travail a bénéficié de l'accès aux ressources HPC de l'IDRIS dans le cadre de l'allocation 2023-AD010314332 faite par GENCI.

Je voudrais sincèrement remercier mes directeurs de thèse: François Rousseau et Douraied Ben Salem. Leur bienveillance, leurs conseils et leur écoute durant ces années m'ont permis d'avancer dans ces recherches et d'apprendre à leurs côtés.

Je souhaite remercier l'ensemble du jury pour avoir accepté de participer à l'évaluation de ces travaux. Merci à Nicolas Passat pour avoir accepté de présider ce jury et pour ses retours constructifs sur le manuscrit. Merci également aux deux rapporteurs: Vincent Noblet et Ahror Belaid, pour leurs retours sur ce travail et ce manuscrit, et à Christelle Pons Becmeur et Rodolphe Bailly pour avoir fait partie de ce jury.

Merci à toute l'équipe du BEaCHILD pour l'ambiance chaleureuse et bienveillante de cette équipe, et pour avoir partagé leur recherches avec toujours beaucoup d'enthousiasme et de pédagogie. Je tiens particulièrement à remercier les personnes avec qui j'ai eu la chance de partager un bureau: Anne, Chloé, Nathan, Benjamin, Guillaume, Yue, Emma, Zakaria, Florian, Triet et Zhengyang. Merci pour ces moments partagés ensemble autour d'une tasse de café, pour votre gentillesse, vos conseils et pour l'atmosphère amicale et chaleureuse de nos locaux. Et un immense merci à Sarah, pour son soutien et son enthousiasme infatigables, pour ses (nombreuses) relectures, tous les fous rires et le café partagés pendant ces années face à face, et pour avoir été ma créancière attitrée au RAK.

Un grand merci aux membre de ma famille et à mes amis qui m'ont accompagnée tout au long de cette thèse, pour leur confiance et leur gentillesse. Merci en particulier à Clara, Juliette, Gaël, Coraline et Gauthier. Merci également à Lili et Polo pour leur présence affectueuse et pour leur participation très active à l'écriture de ce manuscrit.

Enfin, je tiens tout particulièrement remercier ma mère, dont les encouragements, les rires et la présence m'ont soutenue tout au long de cette thèse.

Table of Contents

List of Figures

List of Tables

List of acronyms

AdaIN Adaptative Instance Normalization

AI Artificial Intelligence

ANN Artificial Neural Network

CNN Convolutional Neural Networks

CP Cerebral Palsy

CT Computed Tomography

DL Deep Learning

DNN Deep Neural Networks

DRL Disentangled Representation Learning

FID Fréchet Inception Distance

FiLM Feature-wise Linear Modulation

FR-IQA Full-Reference Image Quality Assessment

GAN Generative Adversarial Networks

GE Gradient Echo

GPU Graphic Processing Unit

I2I Image-to-Image

IQA Image Quality Assessment

KID Kernel Inception Distance

MAE Mean Absolute Error

ML Machine Learning

MLP Multi Layer Perceptron

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

NMR Nuclear Magnetic Resonance

NR-IQA No-Reference Image Quality Assessment

PET Positron Emission Tomography

RF Radiofrequency

SE Spin Echo

Résumé étendu

Chapitre 1: Contexte

Dans un premier chapitre, le contexte clinique associé à la paralysie cérébrale et à l'équin de cheville est présenté, ainsi que les motivations du projet Equinus. Après une présentation des principes fondamentaux de l'IRM et de son application à l'étude du mouvement, nous décrivons en détail le jeu de données acquis dans le cadre du projet Equinus.

Contexte Clinique La paralysie cérébrale (PC) est la déficience motrice la plus courante chez l'enfant, avec une prévalence de 1.5 à 3 pour 1000 naissances vivantes [1]. Elle résulte de lésions cérébrales irréversibles survenues pendant la période périnatale, avant la fin du développement cérébral. L'équin de la cheville est la déformation la plus courante chez les enfants atteints de paralysie cérébrale. Il se définit comme une incapacité à effectuer une dorsiflexion du pied au-dessus du niveau plantigrade, avec l'arrière-pied en position neutre et le genou en extension. L'équin spastique se caractérise par une faiblesse des muscles du pied et de la cheville et un faible contrôle musculaire, aboutissant avec la croissance à des déformations osseuses et une démarche déséquilibrée [2].

D'après la littérature, une intervention chirurgicale après huit ans est l'option la plus viable pour stabiliser le membre inférieur et permettre au patient de marcher de manière aussi indépendante que possible. Cependant, des études de suivi à long terme font état d'un taux de récidive post-opératoire pouvant atteindre 48% [3], [4], [5], [6], et les données probantes sont insuffisantes pour déterminer la meilleure pratique clinique. Bien que des études antérieures aient examiné l'impact des facteurs extrinsèques sur les taux de récidive (âge, le type de membre affecté, type de CP, ou paramètres de marche), les études portant sur l'impact des facteur intrinsèques (biomécanique interne de l'articulation de la cheville ou mécanique musculaire *in vivo*) restent encore insuffisantes. Pour pouvoir améliorer la prise en charge de l'équin, la compréhension de la pathologie et permettre la formulation de recommandations cliniques, il est essentiel d'examiner *in vivo* la mécanique articulaire et les déformations osseuses résultant de la faiblesse de la musculature de l'articulation

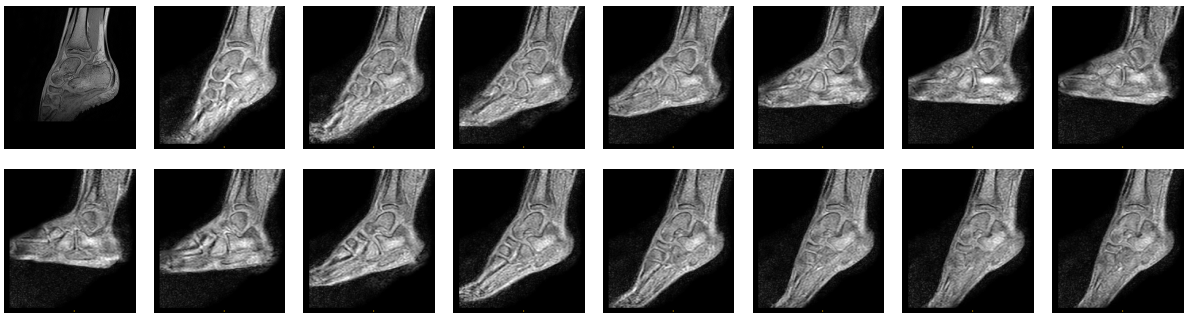
de la cheville.

Données Equinus Le projet Equinus a pour objectif de comprendre les effets in vivo de la faiblesse musculaire de l'articulation de la cheville et les déformations osseuses qui en résultent. La stratégie retenue consiste à comparer la mécanique musculaire et articulaire de la cheville entre des enfants présentant une déformation en équin et des enfants au développement typique appariés selon l'âge [7].

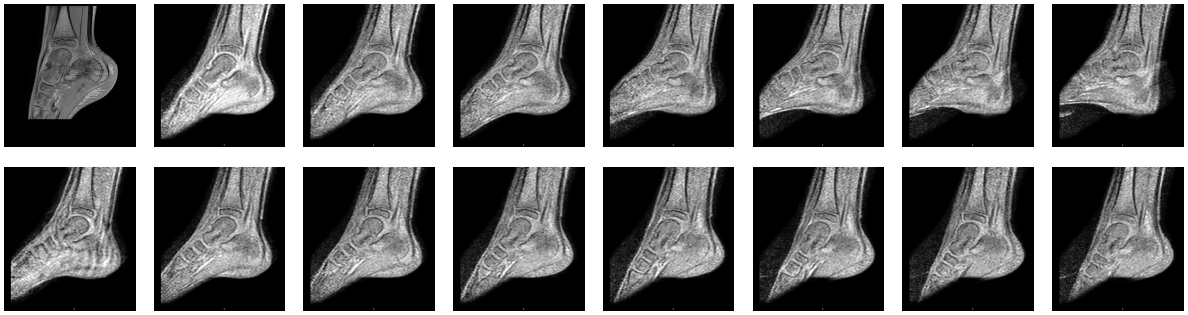
Les techniques d'imagerie modernes, en particulier l'imagerie par résonance magnétique (IRM) dynamique, offrent la possibilité d'étudier la mécanique musculaire et articulaire *in vivo*. Non-invasive et non-ionisante, cette modalité d'imagerie permet aux chercheurs d'observer les mouvements du squelette et des muscles. Pour permettre à la fois la visualisation des mécaniques articulaire au cours du mouvement et bénéficier d'une prise de vue avec une précision anatomique importante, deux types de séquences sont acquises dans le cadre du protocole: une IRM 3D statique à haute résolution et plusieurs séquences dynamiques à basse résolution de l'articulation de la cheville.

L'équin affectant la capacité de dorsiflexion, l'IRM dynamique est utilisée pour enregistrer un cycle de dorsiflexion-flexion plantaire afin d'étudier la biomécanique de l'articulation. L'étude, approuvée par le comité d'éthique régional, inclut onze enfants au développement normal et neuf enfants souffrant d'équin, âgés de 6 à 14 ans. Les données d'IRM 3D ont été recueillies au cours d'une seule visite, après que les parents aient signé des formulaires de consentement. Les données ont été acquises à l'aide d'un scanner MR 3T (Achieva dStream, Philips Medical Systems, Best, Pays-Bas). Le protocole d'acquisition comprend, pour chaque enfant, une IRM 3D statique à haute résolution de l'articulation de la cheville avec une résolution de $0.26 \times 0.26 \times 0.8$mm et trois séquences dynamiques basse résolution de l'articulation de la cheville pendant un unique cycle de dorsiflexion-flexion plantaire. L'acquisition de séquences dynamiques à haute résolution nécessite un temps d'acquisition long et des mouvements répétés à une vitesse constante, ce qui les rend difficiles à réaliser pour les patients souffrant de troubles musculo-squelettiques. Le type de séquence choisi est un compromis entre un temps d'acquisition réduit et une résolution et un contraste suffisants entre les structures anatomiques. Chaque séquence comprend 15 images temporelles 3D avec une résolution spatiale de $0.57 \times 0.57 \times 8$mm. Pour chaque IRM statique, les trois os d'intérêt de l'articulation de la cheville ont été segmentés: le talus, le tibia et le calcaneus. Ces données sont présentées sur la Figure 1 pour deux sujets (un avec équin et un au développement typique). La première image correspond à

l'image RM statique, tandis que les 15 suivantes correspondent aux images successives de la séquence IRM dynamique, montrant un cycle de dorsiflexion-flexion plantaire.



(a) Sujet avec équin.



(b) Sujet au développement typique.

Figure 1 – Exemples de données sources pour deux sujets : l'un avec un équin et l'autre avec un développement typique. Pour chacun des sujets, la première image (en haut à gauche) est une coupe issue d'une image RM statique, et les 15 suivantes sont les images successives qui composent la séquence dynamique.

Les données recueillies présentent des variations anatomiques significatives principalement dues à la présence (ou non) de pathologie et à la tranche d'âge des patients inclus dans l'étude. Ces variations anatomiques incluent des variations dans la forme des os entre les enfants typiques et les enfants équins, ainsi que des différences au niveau des cartilages de croissance en fonction de l'âge. En outre, la résolution de l'image, l'anisotropie et la présence d'artefacts contribuent à la variabilité au sein des données collectées.

Les données dynamiques du projet Equinus disposent d'une haute résolution temporelle, mais souffrent d'une faible résolution spatiale de par leur courte durée d'acquisition. A l'inverse, les images statiques bénéficient d'une meilleure résolution spatiale, mais ne permettent pas de visualiser la biomécanique de la cheville. Afin d'améliorer la compréhension de la biomécanique de la cheville en cas de l'équin de la cheville, l'objectif

est de synthétiser une séquence IRM dynamique haute résolution à partir d'une séquence IRM dynamique basse résolution.

Chapitre 2: Synthèse d'image

La méthodologie utilisée dans cette thèse est fondée sur l'application de techniques de synthèse d'images pour la génération de séquences IRM haute résolution. Ce chapitre aborde tout d'abord les problématiques liées aux données, inhérentes à l'apprentissage profond, avant d'introduire différents concepts de la synthèse d'images. Ce chapitre présente également une revue des dernières métriques d'évaluation des méthodes de synthèse d'images, avec et sans vérité terrain. En outre, les limites des modèles actuels de synthèse d'IRM dynamiques sont discutées, motivant le travail proposé dans les chapitres suivants.

Ces dernières années, des progrès significatifs ont été réalisés dans le domaine de l'apprentissage profond, et ont fait de cette technologie un outil indispensable pour le traitement des images. Leur capacité à modéliser des fonctions non linéaires complexes et leur niveau de complexité adaptable les rendent adaptés à diverses applications. L'état de l'art actuel dans le traitement des images médicales est dominé par ce nouveau paradigme. Les méthodes de synthèse d'images représentent une part importante de ces applications. Les principaux domaines d'application sont le recalage, la segmentation, la synthèse intermodalité, la reconstruction et la super-résolution [8]. Ces nouveaux développements permettent désormais de générer des images d'une modalité spécifique à partir d'images d'une autre modalité, offrant de nouvelles possibilités pour la réduction de l'exposition des patients aux modalités ionisantes, ou la réduction des temps d'acquisition en reconstruisant des images de haute qualité à partir d'images de moindre qualité acquises dans un laps de temps plus court.

Synthèse d'image Le terme « synthèse d'image » désigne le processus de génération d'une image selon un contenu souhaité. Cette synthèse peut être conditionnelle ou inconditionnelle. En synthèse d'image conditionnelle, une image est générée en fonction d'une donnée d'entrée. Cette donnée est utilisée pour décrire le contenu de l'image à générer et conditionne le résultat du processus de génération de l'image. Cette donnée peut prendre des formes diverses : texte, son, signaux cérébraux et indices visuels tels qu'une autre image, une segmentation sémantique ou des cartes de profondeur [9]. Au contraire, la

synthèse d'image inconditionnelle est caractérisée par l'absence de donnée d'entrée pour guider le processus de génération. Cette approche est fréquemment employée pour la génération de grandes quantités de données partageant un même contenu sémantique. À titre d'illustration, considérons un processus de génération de visages. En synthèse d'image conditionnelle, un visage peut être généré à partir d'un texte fournissant une description. Cette approche, bien que permettant la génération d'une image aux caractéristiques spécifiques, n'est pas adaptée à la génération de grands ensembles de données en raison de son caractère chronophage. La synthèse d'image inconditionnelle permet de générer une grande quantité de visages aux caractéristiques aléatoires, sans requérir de description. Cette caractéristique en fait une méthode de choix pour la génération de larges ensembles de données.

La synthèse d'image à image est un sous champ de la synthèse d'image conditionnelle, où la donnée d'entrée conditionnant le processus de génération est une image. Ses champs application sont variés: de la vision par ordinateur, au traitement d'image en passant par l'imagerie médicale. Les applications les plus courantes incluent la colorisation [10][11], la super-résolution [12][13], le transfert de style [14][15], la segmentation [16][17], l'estimation de pose en 3D [18], la synthèse d'image sémantique [19][20], le débruitage [21], la retouche d'images [22][23] ou la génération de dessins animés [24][25].

Synthèse d'image à image appariée / non appariée Les données d'entraînement représentent un aspect crucial pour les applications de l'apprentissage profond. Les premiers modèles de synthèse d'image à image utilisaient pour leur entraînement des images appariées entre les domaines source et cible. En traitement d'image, on parle de données appariées lorsque, pour chaque image et pour une tâche donnée, il existe une vérité terrain correspondante. Ce type de méthode exploite généralement la correspondance entre l'image d'entrée et la vérité terrain correspondante dans le domaine cible pour calculer l'erreur entre l'estimation produite par le modèle et la vérité terrain. Cette erreur va guider l'apprentissage du modèle, et est calculée en utilisant des fonctions de coût comme l'erreur quadratique moyenne (EQM) ou la norme L1 qui comparent la prédiction et la vérité terrain pixel par pixel. Cependant, l'apprentissage avec des données appariées s'avère difficile pour de nombreuses applications en raison de la difficulté ou du coût d'obtention de telles données. En effet, la création de tels ensembles de données requiert une annotation manuelle ou un accès aux vérités terrain correspondantes (masques de segmentation, acquisitions multimodales correspondantes, étiquettes de classe...), limi-

tant ainsi le nombre d'ensembles de données disponibles. Ces défis sont accentués dans le domaine de l'imagerie médicale, où la construction d'un tel ensemble de données peut être fastidieuse (par exemple, segmentation manuelle par un expert), coûteuse (nécessitant plusieurs acquisitions à partir d'un ou plusieurs systèmes d'imagerie) ou imprécise (en raison d'un recalage imparfait entre les images provenant de différentes modalités d'imagerie).

De nouvelles méthodes ont vu le jour pour pallier cette difficulté, comme l'utilisation d'ensembles de données avec peu ou pas de données appariées. L'apprentissage à partir de données non appariées autorise l'utilisation de grands ensembles de données sans nécessiter de vérité terrain [26], [27]. Ces nouvelles méthodes pour l'entraînement des modèles offrent des perspectives plus larges en termes d'applications que celles basées uniquement sur des données appariées. Plusieurs stratégies ont été conçues pour entraîner le modèle en l'absence de données appariées. Parmi les plus utilisées, la contrainte de cohérence cyclique [28] s'affranchit de l'utilisation de données appariées en proposant l'introduction d'une contrainte sur un cycle de synthèse.

Dans la pratique, les données d'entraînement et la tâche visée régissent le plus souvent le modèle d'entraînement et l'architecture choisie. Les méthodes de synthèse d'images appariées s'appuient le plus souvent sur des architectures bien connues, telles que les réseaux résiduels ou UNet, entraînés avec une fonction de coût appariée telle que L1 ou EQM. En synthèse d'images non appariées, deux approches principales dominent l'état de l'art : les réseaux antagonistes génératifs (GAN) à cohérence cyclique et l'apprentissage de représentation factorisées. Ces deux approches sont abordées dans le Chapitre 4.

Application aux données du projet Equinus L'amélioration de la qualité de l'IRM dynamique grâce à la super-résolution ou à la reconstruction est un domaine de recherche spécifique dans la synthèse d'images médicales. La contribution de l'apprentissage profond dans les méthodes proposées est relativement faible par rapport à son importance dans d'autres applications de synthèse d'images médicales. Cela pourrait s'expliquer par le manque de données accessibles au public dans cette modalité particulière. Si certaines techniques ont récemment vu le jour pour compenser le manque de données en accès libre en simulant des processus dynamiques, les méthodes développées s'appuient le plus souvent sur des données appariées. De plus, à notre connaissance, il existe un manque de méthodes qui modélisent explicitement le processus dynamique pour améliorer la résolution.

Les données collectées dans le cadre du projet Equinus sont intrinsèquement non appariées. Cependant, grâce à un processus de recalage, les images IRM statiques et dynamiques peuvent être partiellement alignées. Ce travail se concentre sur deux approches différentes pour la résolution du problème posé. La première exploite le recalage et la simulation pour générer des données partiellement ou totalement appariées. Ces images sont ensuite utilisées dans des méthodes de synthèse d'images appariées. La seconde approche repose sur des méthodes de synthèse d'image à image non appariées qui ne nécessitent donc pas de vérité terrain exacte. Le chapitre 3 porte sur l'utilisation des méthodes de synthèse appariées pour la synthèse de séquences IRM dynamiques haute-résolution, ainsi que le processus de simulation d'images RM dynamiques à partir d'images RM statiques.

Chapitre 3: Synthèse appariée d'IRM dynamique haute résolution

Le troisième chapitre examine l'utilisation de méthodes d'apprentissage profond utilisant des données appariées pour la synthèse d'images IRM dynamiques haute résolution. Nous présentons d'abord deux approches distinctes pour l'appariement des données Equinus : l'une basée sur le recalage et l'autre utilisant la simulation de données. À partir des données issues de ce processus d'appariement, nous explorons l'utilisation de plusieurs architectures pour la synthèse appariée d'images RM dynamiques haute résolution, pour lesquelles les images RM statiques sont utilisées comme vérité terrain pour l'entraînement des modèles.

Appariement des données Dans cette section, nous détaillons le processus d'appariement entre les données IRM statiques et dynamiques. L'objectif est de permettre l'utilisation de méthodes de synthèse d'image à image appariées. Deux approches pour l'appariement des données sont explorées: une approche basée sur le recalage pour aligner les images RM statiques et dynamiques les unes par rapport aux autres, et une approche de simulation de données IRM dynamiques à partir d'IRM statiques.

Le processus de recalage est divisé en deux phases: un premier recalage rigide global suivi d'une deuxième recalage rigide de précision. La première phase de recalage a pour but de déplacer l'image RM statique dans le système de coordonnées de l'image RM dynamique. À l'issue de cette première phase, un premier filtrage est intégré de façon à éliminer les images pour lesquelles la transformation estimée n'est pas réaliste. Dans

un second temps, la deuxième phase de recalage consiste a estimer un recalage plus fin, pour chacun des os de l'articulation de la cheville, entre les données IRM statiques et dynamiques. Cette approche offre un temps de calcul réduit par rapport à une approche non-rigide, et garantit la préservation des caractéristiques anatomiques en éliminant le risque de déformations au sein des tissus. Cependant, ce recalage ne produit qu'un appariement partiel des images RM statiques et dynamiques, limité à une région osseuse.

Le but de la simulation de données dynamiques est de générer des données IRM statiques et dynamiques appariées, c'est-à-dire des données parfaitement alignées entre les deux types de séquences IRM. L'objectif est d'estimer la transformation d'une image RM statique en une image RM dynamique. Deux approches sont utilisées pour la simulation de données: une approche basée sur l'apprentissage profond et une sur la simulation manuelle. Aucune des approches par apprentissage profond ne donne lieu à une synthèse réaliste de données IRM dynamiques simulées, quantitativement ou qualitativement. L'approche manuelle produit les résultats de simulation les plus réalistes parmi les méthodes explorées.

Synthèse appariée d'IRM dynamique haute résolution Cette section décrit les méthodes d'apprentissage profond utilisées pour l'estimation de séquences IRM dynamiques haute résolution en utilisant des données appariées et partiellement appariées. Ces travaux sont menés en utilisant trois jeux de données: les données réelles issu du projet Equinus, un jeu de données simulées construit à partir des données du projet Equinus et le jeu de données HCP. Les données réelles sont composées de paires d'images RM statiques et dynamiques recalées os par os. Les données simulées sont constituées de paires d'images RM statiques et d'images RM dynamiques simulées de façon manuelle. La construction de chacun de ces jeux de données est détaillée dans la section ci-dessus. Enfin, afin de valider la synthèse d'image appariée, les expérimentations sont également menées sur le jeu de données HCP, conçu pour la cartographie du connectome cérébral humain. Au sein de ce jeu de données, chaque sujet dispose de deux IRM appariées: une T1 et une T2.

Les expérimentations sont menées sur des données appariées et partiellement appariées. La fonction de coût est adapté selon le type de données utilisé pour l'entraînement. Dans le cas de données parfaitement appariées, comme dans l'ensemble de données simulées, chaque pixel de l'image prédite peut être comparé à ceux de la vérité terrain. Dans le cas des données partiellement appariées, des segmentations osseuses sont utilisées pour pondérer l'erreur quadratique moyenne classique à zéro en dehors de la région osseuse où les images sont recalées.

Bien que les expériences menées sur l'ensemble de données HCP aient donné des résultats réalistes, ni l'estimation du modèle inverse ni l'utilisation d'une estimation du modèle direct n'ont permis d'obtenir une synthèse réaliste d'IRM dynamique à haute résolution. Cela peut être attribué à la qualité du processus de recalage. En effet, la différence d'aspect importante entre les images IRM statiques et dynamiques peut rendre le processus de recalage complexe et ardu, pouvant entraîner des erreurs et compromettant l'apprentissage via des données appariées. Ces erreurs peuvent avoir un impact significatif sur l'apprentissage par paires. En outre, aucune des méthodes basées sur l'apprentissage pour simuler des données IRM dynamiques n'a donné de résultats de synthèse réalistes. L'approche manuelle, quant à elle, semble produire des donnes insuffisamment réalistes pour permettre la généralisation de l'estimation du modèle inverse à des données réelles.

Les méthodes appariées pour estimer l'IRM dynamique haute résolution à partir de l'IRM dynamique basse résolution n'ayant pas permis une synthèse réaliste de séquences IRM dynamiques haute résolution, le chapitre suivant explorera l'utilisation de méthodes de synthèse d'image à image non appariées pour la résolution du problème.

Chapitre 4: Synthèse non appariée d'IRM dynamique haute résolution

Ce chapitre propose une étude du potentiel des méthodes de synthèse d'image à image non appariées pour la synthèse de séquences d'IRM dynamiques à haute résolution. Une part importante de ces méthodes est basée sur l'apprentissage de représentations factorisées. Les représentations factorisées sont une classe de méthodes basées sur l'hypothèse qu'une distribution de données peut être décrite par un ensemble de facteurs indépendants et sémantiquement significatifs. L'objectif de ces méthodes est d'identifier ces facteurs et de les encoder dans des dimensions distinctes.

En s'appuyant sur la factorisation d'une image en facteurs de variation indépendants, cette approche offre un meilleur contrôle sur le résultat de la synthèse que les méthodes basées sur les GAN. Cependant, la factorisation des représentations latentes n'est pas une tâche triviale et peut être influencé par des biais inductifs particuliers. Dans ce chapitre, nous évaluons d'abord l'impact de plusieurs hyperparamètres et l'ajout de contraintes d'espace latent, puis nous étudions l'impact du module de fusion et de l'ajout d'une tâche auxiliaire de segmentation sur le résultat de la synthèse et la factorisation des représentations.

Ce travail apporte deux contributions principales. Premièrement, nous proposons l'ajout de contraintes sur les espaces latents afin d'améliorer la qualité de la synthèse. Deuxièmement, nous proposons d'étudier l'influence du module de fusion sur les performances de la synthèse. A notre connaissance, aucune étude n'a étudié son influence sur les représentations factorisées. Nous étudions également l'impact de l'ajout d'un module de segmentation à partir de la représentation latente d'anatomie sur les résultats de la synthèse. Dans cette seconde étude, nous proposons également une évaluation de la factorisation des représentations latentes et une analyse des espaces latents. Nos résultats démontrent que le module de fusion a un impact significatif sur les performances d'une méthode basée sur des représentations factorisées, et notamment sur le taux de factorisation de la méthode.

Méthodes La méthode utilisée s'inspire de DRIT++, un modèle de synthèse d'images non appariées basé sur l'apprentissage de représentations factorisées, proposé par [29]. L'objectif de ce travail est de parvenir à dissocier le contenu, appelé par la suite "anatomie" dans la terminologie de l'imagerie médicale, de sa représentation. Cette représentation, appelée "modalité", fait référence aux paramètres d'acquisition de l'IRM propres à chaque image et englobe les attributs spécifiques à la modalité. Généralement représentée sous la forme d'un vecteur, elle n'encode pas d'informations spatiales mais le rendu des structures anatomiques. Inversement, l'anatomie représente l'information spatiale, invariante selon la modalité, au sein de l'image. Encodée sous forme de tenseur, cette représentation préserve les corrélations spatiales des images originales, ce qui la rend adaptée aux tâches nécessitant une équivariance. La synthèse d'une modalité source à une modalité cible est réalisée en fournissant à un générateur le code latent d'anatomie de l'image source et le code latent de modalité de l'image de la modalité cible.

En nous appuyant sur cette approche, nous étudions l'impact de plusieurs paramètres sur l'apprentissage des représentations factorisées et sur la qualité de la synthèse. En particulier, nous examinons l'influence du module de fusion et l'intégration d'une tâche auxiliaire de segmentation à partir d'une représentation anatomique. Le module de fusion est un élément critique dans l'apprentissage des représentations factorisées, permettant la fusion de différentes représentations au sein du modèle. Notre analyse s'appuie sur une évaluation de la factorisation des représentations pour toutes ces méthodes.

	$\mathcal{L}_{seg}$	Entanglement	FID ↓	KID ↓	DSC ↑
DRL	✗	Conv	164.13	0.12	/
		AdaIN	195.5	0.17	/
		FiLM	367.5	0.44	/
DRL	✓	Conv	162.54	0.12	0.959 ± 0.003
		AdaIN	**134.8**	**0.076**	0.959 ± 0.004
		FiLM	201.4	0.15	0.963 ± 0.004
CycleGAN			154.81	0.10	/

Table 1 – Évaluation de la qualité de la synthèse d'une image RM dynamique haute résolution, pour différents modules de fusion, avec et sans tâche de segmentation, et comparée aux résultats obtenus par le CycleGAN. Pour l'évaluation, deux métriques pour la synthèse d'image non appariée sont utilisées, KID et FID, et complétées par un DSC calculé sur les segmentations générées. Les meilleurs performances sont indiquées en gras.

Résultats La Figure 2 présente une partie des résultats obtenus pour la synthèse d'image IRM dynamiques haute résolution. Ces résultats ont été obtenus dans le cadre de l'étude sur le module d'intégration et l'apport de la segmentation comme tache auxiliaire.

Les expériences démontrent que l'ajout d'une tâche de segmentation auxiliaire améliore considérablement la qualité de synthèse des images dynamiques haute résolution, renforce la robustesse du modèle face aux transformations spatiales et permet une meilleure factorisation entre les représentations latentes. Si les modèles basés sur les modules de fusion Conv et AdaIN produisent tous deux des images visuellement réalistes, le modèle utilisant le module AdaIN semble moins sujet aux artefacts et produit des contours plus réguliers. Si les modèles utilisant le module AdaIN présentent une amélioration systématique de la corrélation entre la représentation latente de l'anatomie et l'image source, les modèles basé sur Conv démontrent systématiquement une meilleure factorisation entre les deux représentations latentes. Le modèle CycleGAN synthétise une texture visuellement réaliste pour les structures osseuses, mais semble être plus sensible aux structures anatomiques irrégulières et aux artefacts. De plus, il est montré que sa robustesse au regard des transformations spatiales affines est inférieure à celle des méthodes basées sur les représentations factorisées. Les approches proposées dans ce chapitres permettent la génération de séquences dynamiques 3D+t haute résolution. Les résultats optimaux sont obtenus lorsque le module de fusion AdaIN est associé à une tâche de segmentation auxiliaire.

Cette étude démontre la pertinence de l'utilisation des représentations factorisées pour la synthèse de I2I non appariée appliquée à l'IRM dynamique. En appliquant la méthode proposée à une séquence dynamique, nous réussissons à générer une séquence dynamique

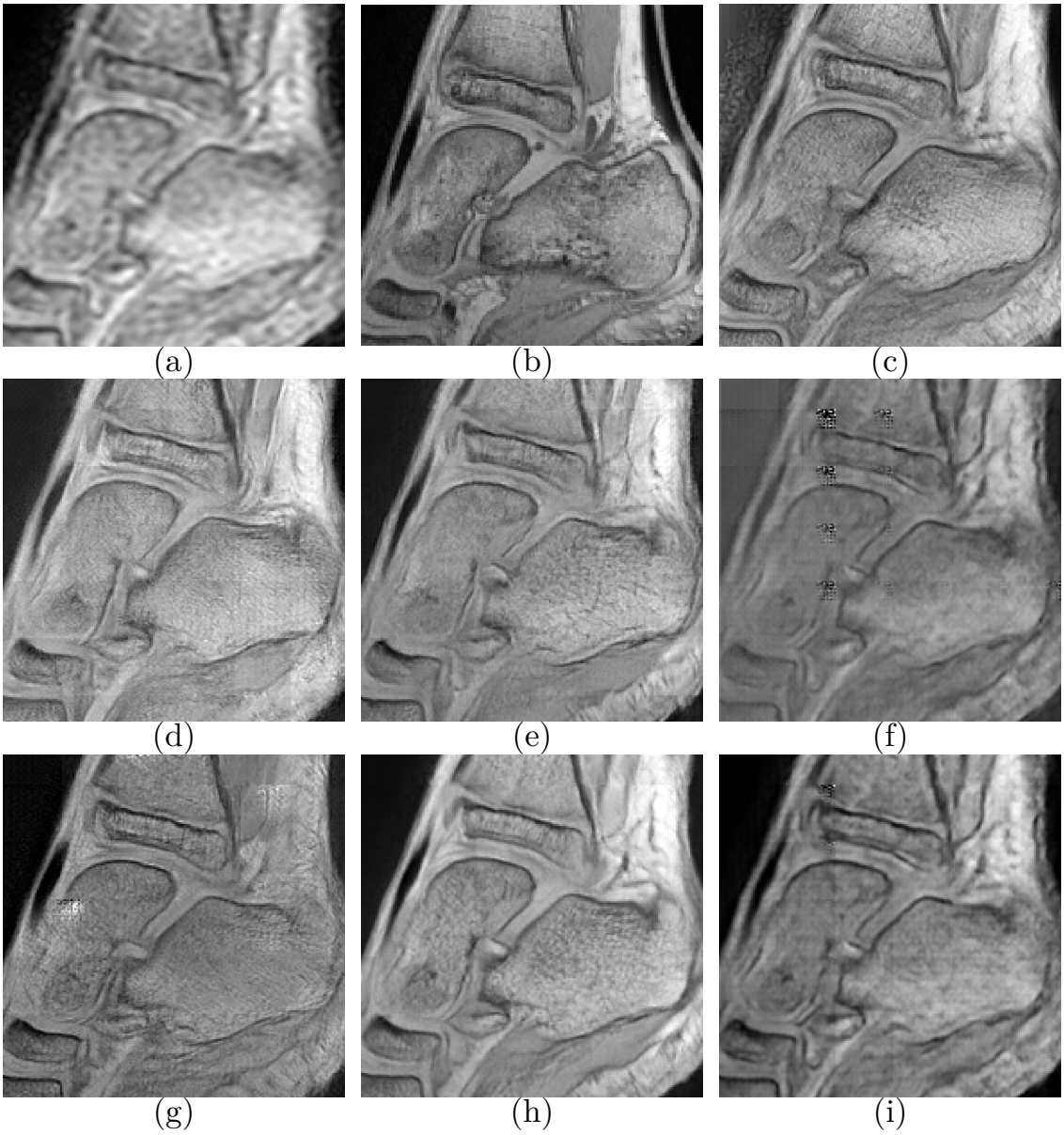


Figure 2 – Résultats de la synthèse de séquences dynamiques à haute résolution pour différents modules de fusion, avec et sans la tâche de segmentation auxiliaire. (a) Image dynamique. (b) Image statique. (c) Estimation obtenue avec CycleGAN. (d) Estimation obtenue avec mDRIT++ - Conv. (e) Estimation obtenue avec mDRIT++ - AdaIN. (f) Estimation obtenue avec mDRIT++ - FiLM. (g) Estimation obtenue avec mDRIT++ - Conv avec la tâche de segmentation. (h) Estimation obtenue avec mDRIT++ - AdaIN avec la tâche de segmentation. (i) Estimation obtenue avec mDRIT++ - FiLM avec la tâche de segmentation.

3D+t haute résolution. De plus, cette méthode offre des performances comparables, voire supérieures, à celles du CycleGAN, qui constitue une part importante de l'état de l'art des méthodes en synthèse d'image à image non appariée. Elle offre également une stabilité supérieure vis-à-vis des transformations affines.

Conclusion

L'imagerie RM dynamique est une modalité d'imagerie médicale utilisée pour observer la dynamique physiologique *in vivo*. Cette modalité d'imagerie, largement utilisée pour étudier le muscle cardiaque ou la biomécanique des articulations, est non ionisante et non invasive et permet une excellente visualisation des structures anatomiques. Cependant, l'acquisition de séquences d'IRM dynamique à haute résolution est un processus qui prend du temps et qui nécessite des mouvements répétés à une vitesse constante. Afin de réduire la pression exercée sur les patients souffrant de troubles musculo-squelettiques, le temps d'acquisition des séquences peut être réduit au détriment de la résolution spatiale. Cette thèse examine le potentiel des méthodes d'apprentissage profond pour la synthèse de séquences IRM dynamiques à haute résolution.

L'un des défis les plus importants de ce travail réside dans l'appariement des données. En effet, les données acquises dans le cadre du projet Equinus sont sujettes à des déformations non rigides entre les différentes images d'un même sujet, résultant en des images IRM statiques et dynamiques non alignées. Ces déformations sont une conséquence de la nature dynamique de l'étude et des mouvements demandés aux sujets. Ces données ont guidé le choix des approches étudiées pour la synthèse de séquences IRM dynamiques à haute résolution.

Une première approche de la synthèse de séquences dynamiques à haute résolution a été menée pour étudier le potentiel des méthodes de synthèse d'image à image utilisant des données appariées. Deux approches distinctes ont été traitées pour l'appariement des données : la première est basée sur un processus de recalage et la seconde sur la simulation des données. En utilisant ces données, nous explorons plusieurs architectures pour la synthèse de séquences IRM dynamiques à haute résolution. Après avoir validé les approches proposées sur un ensemble de données d'imagerie médicale apparié, nous démontrons l'incapacité de ces mêmes méthodes à synthétiser des images dynamiques haute résolution visuellement réalistes.

La deuxième approche étudiée s'est concentrée sur les méthodes de synthèse d'image

à image utilisant des données non appariées pour la synthèse d'IRM dynamique haute résolution. En particulier, les travaux se sont concentrés sur les méthodes basées sur des représentations factorisées. Les résultats ont démontré que cette approche permet une synthèse réaliste d'une image RM dynamique haute résolution. La méthode a été appliquée à une séquence complète d'IRM dynamique, ce qui a permis de synthétiser avec succès une séquence IRM dynamique 3D+t haute résolution.

En nous appuyant sur cette approche, nous étudions l'impact de divers paramètres sur l'apprentissage des représentations factorisées et sur la qualité de la synthèse. En particulier, nous examinons l'influence du module de fusion et de l'intégration d'une tâche auxiliaire de segmentation à partir d'une représentation anatomique. Le module de fusion est un élément critique dans l'apprentissage des représentations factorisées, permettant la fusion de différentes représentations au sein du modèle. Notre analyse s'appuie sur une évaluation du taux de factorisation pour toutes ces méthodes. Nos résultats montrent que ces deux paramètres ont un impact significatif à la fois sur la qualité de la synthèse et sur le taux de factorisation de la méthode. De plus, nous démontrons que l'introduction de contraintes sur les espaces latents peut influencer le processus de synthèse et la qualité des images générées.

La méthode employée dans cette étude présente des performances comparables, voire supérieures, à celles de CycleGAN, ainsi qu'une stabilité accrue au regard de transformations affines. En outre, la méthode permet de préserver le contenu anatomique entre l'image dynamique source et l'image synthétisée, ce qui représente un élément crucial dans l'imagerie médicale.

References

[1] H. K. Graham et al., « Musculoskeletal Pathology in Cerebral Palsy: A Classification System and Reliability Study », en, *Children*, vol. 8, *3*, p. 252, Mar. 2021, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2227-9067. DOI: 10.3390/children8030252. Accessed: Sep. 12, 2022. [Online]. Available: https://www.mdpi.com/2227-9067/8/3/252.

[2] A. Horsch, M. C. M. Klotz, H. Platzer, S. Seide, N. Zeaiter, and M. Ghandour, « Is the Prevalence of Equinus Foot in Cerebral Palsy Overestimated? Results from a Meta-Analysis of 4814 Feet », en, *Journal of Clinical Medicine*, vol. 10, *18*, p. 4128, Jan. 2021, Number: 18 Publisher: Multidisciplinary Digital Publishing Institute,

ISSN: 2077-0383. DOI: 10.3390/jcm10184128. Accessed: Sep. 12, 2022. [Online]. Available: https://www.mdpi.com/2077-0383/10/18/4128.

[3] M. Krupiński, A. Borowski, and M. Synder, « Long Term Follow-up of Subcutaneous Achilles Tendon Lengthening in the Treatment of Spastic Equinus Foot in Patients with Cerebral Palsy », eng, *Ortopedia, traumatologia, rehabilitacja*, vol. 17, *2*, pp. 155–161, Mar. 2015, ISSN: 2084-4336. DOI: 10.5604/15093492.1157092. Accessed: Aug. 16, 2023. [Online]. Available: https://doi.org/10.5604/15093492.1157092.

[4] G. B. Firth et al., « Multilevel Surgery for Equinus Gait in Children with Spastic Diplegic Cerebral Palsy: Medium-Term Follow-up with Gait Analysis », en-US, *JBJS*, vol. 95, *10*, p. 931, May 2013, ISSN: 0021-9355. DOI: 10.2106/JBJS.K.01542. Accessed: Aug. 16, 2023. [Online]. Available: https://journals.lww.com/jbjsjournal/abstract/2013/05150/multilevel_surgery_for_equinus_gait_in_children.10.aspx.

[5] S. Y. Joo, D. N. Knowtharapu, K. J. Rogers, L. Holmes, and F. Miller, « Recurrence after surgery for equinus foot deformity in children with cerebral palsy: Assessment of predisposing factors for recurrence in a long-term follow-up study », en, *Journal of Children's Orthopaedics*, vol. 5, *4*, pp. 289–296, Aug. 2011, Publisher: SAGE Publications, ISSN: 1863-2521. DOI: 10.1007/s11832-011-0352-4. Accessed: Aug. 16, 2023. [Online]. Available: https://doi.org/10.1007/s11832-011-0352-4.

[6] C. Y. Chung et al., « Recurrence of Equinus Foot Deformity After Tendo-Achilles Lengthening in Patients With Cerebral Palsy », en-US, *Journal of Pediatric Orthopaedics*, vol. 35, *4*, p. 419, Jun. 2015, ISSN: 0271-6798. DOI: 10.1097/BPO.0000000000000278. Accessed: Aug. 16, 2023. [Online]. Available: https://journals.lww.com/pedorthopaedics/abstract/2015/06000/recurrence_of_equinus_foot_deformity_after.20.aspx.

[7] University Hospital, Brest, « In Vivo Dynamic Evaluation of Ankle Joint and Muscle Mechanics in Children With Spastic Equinus Deformity Due to Cerebral Palsy: Implications for Recurrent Equinus. », clinicaltrials.gov, Clinical trial registration NCT02814786, Dec. 2021, submitted: February 10, 2016. Accessed: Aug. 15, 2023. [Online]. Available: https://clinicaltrials.gov/study/NCT02814786.

[8] S. K. Zhou et al., « 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. DOI: 10.1109/JPROC.2021.3054390.

[9] F. Zhan et al., « Multimodal Image Synthesis and Editing: A Survey and Taxonomy », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2023, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2023.3305243.

[10] R. Zhang, P. Isola, and A. A. Efros, « Colorful Image Colorization », en, *in Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 649–666, ISBN: 978-3-319-46487-9. DOI: 10.1007/978-3-319-46487-9_40.

[11] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, « Infrared Image Colorization Based on a Triplet DCGAN Architecture », *in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, ISSN: 2160-7516, Jul. 2017, pp. 212–217. DOI: 10.1109/CVPRW.2017.32.

[12] C. Ledig et al., « Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network », *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690. Accessed: Sep. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.html.

[13] Y. Hu, X. Gao, J. Li, Y. Huang, and H. Wang, « Single image super-resolution with multi-scale information cross-fusion network », en, *Signal Processing*, vol. 179, p. 107 831, Feb. 2021, ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2020.107831. Accessed: Sep. 12, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168420303753.

[14] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, « TSIT: A Simple and Versatile Framework for Image-to-Image Translation », en, *in Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 206–222, ISBN: 978-3-030-58580-8. DOI: 10.1007/978-3-030-58580-8_13.

[15] L. A. Gatys, A. S. Ecker, and M. Bethge, « Image Style Transfer Using Convolutional Neural Networks », en, *in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2414–2423, ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.265. Accessed: Sep. 8, 2022. [Online]. Available: http://ieeexplore.ieee.org/document/7780634/.

[16] J. Long, E. Shelhamer, and T. Darrell, « Fully Convolutional Networks for Semantic Segmentation », 2015, pp. 3431–3440. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html.

[17] A. Chartsias et al., « Disentangled representation learning in cardiac image analysis », en, *Medical Image Analysis*, vol. 58, p. 101 535, Dec. 2019, ISSN: 1361-8415. DOI: 10.1016/j.media.2019.101535. Accessed: Sep. 12, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841519300684.

[18] S. Li, S. Gunel, M. Ostrek, P. Ramdya, P. Fua, and H. Rhodin, « Deformation-Aware Unpaired Image Translation for Pose Estimation on Laboratory Animals », 2020, pp. 13 158–13 168. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Deformation-Aware_Unpaired_Image_Translation_for_Pose_Estimation_on_Laboratory_Animals_CVPR_2020_paper.html.

[19] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, « Semantic Image Synthesis With Spatially-Adaptive Normalization », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346. Accessed: Sep. 8, 2022. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Semantic_Image_Synthesis_With_Spatially-Adaptive_Normalization_CVPR_2019_paper.html.

[20] H. Tang, D. Xu, Y. Yan, P. H. S. Torr, and N. Sebe, « Local Class-Specific and Global Image-Level Generative Adversarial Networks for Semantic-Guided Scene Generation », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7870–7879. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Tang_Local_Class-Specific_and_Global_Image-Level_Generative_Adversarial_Networks_for_Semantic-Guided_CVPR_2020_paper.html.

[21] A. Buades, B. Coll, and J.-M. Morel, « A non-local algorithm for image denoising », *in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, ISSN: 1063-6919, vol. 2, Jun. 2005, 60–65 vol. 2. DOI: 10.1109/CVPR.2005.38.

[22] X. Liu et al., « Open-Edit: Open-Domain Image Manipulation with Open-Vocabulary Instructions », en, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 89–106, ISBN: 9783030586218. DOI: 10.1007/978-3-030-58621-8_6.

[23] K. Crowson et al., « VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance », en, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., ser. Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2022, pp. 88–105, ISBN: 9783031198366. DOI: 10.1007/978-3-031-19836-6_6.

[24] Y. Shi, D. Deb, and A. K. Jain, « WarpGAN: Automatic Caricature Generation », 2019, pp. 10 762–10 771. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Shi_WarpGAN_Automatic_Caricature_Generation_CVPR_2019_paper.html.

[25] Y. Chen, Y.-K. Lai, and Y.-J. Liu, « CartoonGAN: Generative Adversarial Networks for Photo Cartoonization », 2018, pp. 9465–9474. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Chen_CartoonGAN_Generative_Adversarial_CVPR_2018_paper.html.

[26] Z. Yi, H. Zhang, P. Tan, and M. Gong, « DualGAN: Unsupervised Dual Learning for Image-To-Image Translation », 2017, pp. 2849–2857. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Yi_DualGAN_Unsupervised_Dual_ICCV_2017_paper.html.

[27] S. Ma, J. Fu, C. W. Chen, and T. Mei, « DA-GAN: Instance-Level Image Translation by Deep Attention Generative Adversarial Networks », 2018, pp. 5657–5666. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Ma_DA-GAN_Instance-Level_Image_CVPR_2018_paper.html.

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, « Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks », *in Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232. Accessed: Sep. 12, 2022. [Online]. Available: `https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html`.

[29] H.-Y. Lee et al., « DRIT++: Diverse Image-to-Image Translation via Disentangled Representations », en, *International Journal of Computer Vision*, vol. 128, *10*, pp. 2402–2417, Nov. 2020, ISSN: 1573-1405. DOI: `10.1007/s11263-019-01284-z`. Accessed: Sep. 8, 2022. [Online]. Available: `https://doi.org/10.1007/s11263-019-01284-z`.

INTRODUCTION

Motivation

Cerebral palsy (CP) is the most prevalent motor impairment among children. It results from irreversible brain injuries that occur during the prenatal period, before the end of cerebral development. Equinus is the most prevalent deformity in children with CP [1]. It is defined as the inability to dorsiflex the foot above the level of plantigrade, with the hindfoot in a neutral position and the knee in an extended position, leading to gait and bone deformities with growth. Based on the literature, surgical intervention after eight years is the most viable option to stabilize the lower extremity and allow the patient to walk as independently as possible. However, long-term follow-up studies report a recurrence rate of up to 48% after surgery [2], [3], [4], [5], and there is insufficient evidence to determine the best clinical practice. Although previous studies have investigated the impact of extrinsic factors on recurrence rates, there is still a lack of morphological analysis of the ankle joint and muscle biomechanics and tendon strains. To improve understanding of the pathology and enable more informed clinical recommendations, it is crucial to examine joint mechanics and bone deformities caused by weak ankle joint musculature *in vivo*. State-of-the-art imaging techniques, particularly dynamic Magnetic Resonance Imaging (MRI), provide innovative ways to investigate *in vivo* muscle and joint mechanics. This imaging modality is non-invasive and non-ionizing, and has proven accuracy and precision, enabling researchers to monitor skeletal and muscle motion. Nevertheless, capturing high-resolution dynamic sequences requires a long acquisition time and repetitive motion patterns at a constant speed. To reduce the strain on patients with musculoskeletal disorders, the acquisition time of the sequences can be reduced at the cost of spatial resolution. In order to improve the understanding of the ankle biomechanics in equinus pathology, the objective is to synthesize a high-resolution dynamic MRI sequence derived from a low-resolution dynamic MRI sequence.

Medical imaging is employed in standard clinical practice to confirm or pose a diagnosis, to analyze body functions such as the brain, heart, or body kinematics, for therapeutic follow-up, or for interventional radiology. In recent years, significant progress has been

made in the field of deep learning, which has made this technology an indispensable tool for image processing. The current state-of-the-art in medical image processing is dominated by this novel paradigm, which is used for diverse purposes, including diagnosis support, medical event prediction, and decision making or classification. Image synthesis methods represent a significant part of these applications. The main fields of application include registration, segmentation, inter-modality synthesis, reconstruction, and super-resolution [6]. These novel developments now permit the generation of images of a specific modality from images of another modality, offering novel possibilities for the reduction of patient exposure to ionizing modalities, or the reduction of acquisition times by reconstructing high-quality images from lower-quality images acquired in a shorter period. While considerable attention is being devoted to enhancing the quality of static magnetic resonance (MR) images, there has been a paucity of deep learning methods developed to improve the quality of 3D+t sequence images.

This thesis examines the potential of deep learning methods for the synthesis of high-resolution dynamic MRI sequences from low-resolution dynamic MRI sequences and high-resolution static MR images. One of the most significant challenges in this work lies in the pairing of data. Indeed, the data collected as part of the Equinus project is subject to non-rigid deformations between different images of the same subject, resulting in unaligned static and dynamic MR images. These deformations are a consequence of the dynamic nature of the study and the movements required of the subjects. These data directed the selection of methodologies to be investigated for the synthesis of high-resolution dynamic MRI sequences.

As a consequence, the initial scientific question that has guided this work is whether paired image-to-image (I2I) synthesis methods can be employed to synthesize high-resolution dynamic MRI sequences. To this end, we conducted an investigation into the potential of paired image synthesis methods using both simulated and registered data. However, none of the studied approaches yielded realistic synthesis of high-resolution dynamic MR images. As a consequence, we then considered whether unpaired I2I methods might offer superior performance for high-resolution dynamic MRI sequence synthesis. We thus investigated unpaired image synthesis models and focused on a class of methods called disentangled representation learning. Building upon this approach, we investigated the impact of several architecture hyperparameters on the synthesis quality. This study has led to two additional questions that have guided our work: What is the impact of the entanglement module and the incorporation of an auxiliary segmentation task on the

learning of disentangled representations and the quality of synthesis? Does the introduction of constraints in the latent space of a disentangled representation model affect the reconstruction quality?

Thesis overview

The objective of this thesis is to propose a method for synthesizing high-resolution dynamic MRI sequences from low-resolution dynamic MRI sequences and high-resolution static MR images. This method may contribute to a greater understanding of the *in vivo* biomechanics of the ankle joint in children with cerebral palsy through the use of dynamic MRI sequences with a short acquisition time. The principal contributions of this thesis are listed bellow:

- Synthesis of high-resolution 3D+t dynamic MR sequences from low-resolution dynamic MR sequences and a high-resolution static MR image.
- Evaluation of the impact of the entanglement module on the learning of disentangled representations.
- Demonstration of the impact of incorporating an auxiliary segmentation task on the synthesis result.
- Quantitative evaluation of the disentanglement, and subsequent analysis of the latent space.
- Evaluation of the impact of latent space constraints on the synthesis result.
- Evaluation of paired image-to-image translation frameworks for the synthesis of high-resolution dynamic MR images.

Thesis organization

This thesis is organized into four chapters, followed by a conclusion and an extended abstract in French.

In the initial chapter, the clinical context associated with cerebral palsy and ankle equinus is presented, along with the motivations behind the Equinus project. Following a presentation of the fundamental principles of MRI and its application to the study of movement, we then provide a detailed description of the data set acquired as part of the Equinus project.

Following a brief introduction to the fundamental concepts of artificial intelligence

(AI) and deep learning (DL), the second chapter addresses the data-related challenges associated with DL. A presentation and state-of-the-art of image synthesis using deep learning is provided, followed by a review of metrics for evaluating image synthesis methods, with and without reference. Finally, we present examples of DL methods for dynamic MRI image synthesis.

The third chapter examines the use of deep learning methods using paired data for high-resolution dynamic MRI images synthesis. We first present two distinct approaches for the pairing of Equinus data: one registration-based and the other using data simulation. We then proceed to explore several architectures for high-resolution dynamic MRI image synthesis based on these data.

Finally, the fourth chapter explores the use of unpaired image synthesis methods for the synthesis of high-resolution dynamic MRI sequences. The employed approach is based on disentangled representations, a class of methods formulated on the assumption that a data distribution can be described by a set of independent and semantically meaningful factors of variation. Learning disentangled representations aims to discover and encode these factors into separate dimensions. The learning of such representations is not trivial, and it is influenced by several inductive biases. This chapter presents an evaluation of the impact of several hyperparameters, as well as the incorporation of latent space constraints, on the synthesis quality. Furthermore, it investigates the impact of both the entanglement module and the introduction of a segmentation auxiliary task on the quality of the synthesis and the disentanglement of the representations.

Chapter 1

Context

Abstract

This chapter introduces the clinical context of spastic equinus, which motivated the research carried out during this thesis. The motivations behind the Equinus project are presented, followed by an explanation of the principle of Magnetic Resonance Imaging (MRI), the different types of sequences, and the use of MRI to study movement. The chapter concludes with a description of the Equinus dataset, including the acquisition protocol, data, sources of variability, and derivative data.

1.1 Introduction

Cerebral palsy (CP) is the most common motor impairment in children, with a prevalence ranging from 1.5 to 3 per 1000 live births [7]. CP affects 17 million people worldwide, and 125,000 in France. Its manifestations include a range of motor disorders (ranging from limping to quadriplegia), which may be accompanied by cognitive impairment. Among the deformities associated with CP, equinus is one of the most prevalent, impacting both the growth of affected limbs and gait patterns. This pathology remains incompletely understood. However, morphological analysis of the ankle joint and muscle biomechanics may prove beneficial in advancing understanding of the mechanisms involved in equinus in children with CP and in improving medical management strategies.

This chapter is divided in three sections. Section 1.2 oversees the clinial context of the cerebral palsy and the equinus foot. Section 1.3 describes the Magnetic Resonance Imaging principle and its application to motion analysis. Finally, Section 1.4 describes the data obtained in the context of Equinus Project.

1.2 Clinical Context

1.2.1 Cerebral palsy

Cerebral palsy results from irreversible brain injuries during the perinatal period, before the end of cerebral development. The most common causes include: infection or illness during pregnancy or during the first few months, complications resulting from difficult birth, prematurity (about half the cases, particularly high-risk if the child is born before the 37th week of pregnancy), severe convulsions, neonatal stoke or abuse-related traumas. These lesions lead to permanent motor, sensitive or cognitive disorders, the nature and severity of which vary from patient to patient depending on the cerebral areas affected, the moment when the injuries occur and their size [8], [9], [10]. While cerebral lesions are stable or only slightly evolving, orthopedic sequelae tend to worsen with growth. These aggravations are induced, for example, by muscular imbalances or progressive bone or joint deformations. These disorders therefore require long-term therapeutic follow-up.

Most common symptoms includes: weak limbs, uncontrolled or chaotic movements, muscular spasms or tiptoe walking. Other symptoms can occur including vision disorders, epilepsy, speaking and writing difficulties or altered intellectual capacity. The diagnosis is done according to clinical symptoms following by brain scans (magnetic resonance imaging (MRI) or transfontanellar ultrasound) or electroencephalogram (EEG), but cannot be posed until the first 3 to 6 months. Early diagnosis and prompt medical action are decisive in minimizing long-term consequences. There are four main types of cerebral palsy, depending on the area affected (see Figure 1.1). Spastic CP (the most prevalent) is caused by damage located in the motor cortex. It is characterized by stiff and tight muscles, which can impair movement and lead to bone and joint deformities. Dyskinetic CP is caused by damage to the basal ganglia, resulting in involuntary movements or spasms. Ataxic CP results from damage to the cerebellum, affecting balance and spatial perception. It is characterized by impaired balance and coordination.

1.2.2 Equinus

Around 90% of the most common deformities in CP occurs in the ankle and foot region. Equinus is the most common deformity in children with spastic CP (incidence around 75%) [1]. It is defined as the inability to dorsiflex the foot above plantigrade, with the hindfoot in neutral position and the knee in extended position, (see Figure 1.2),

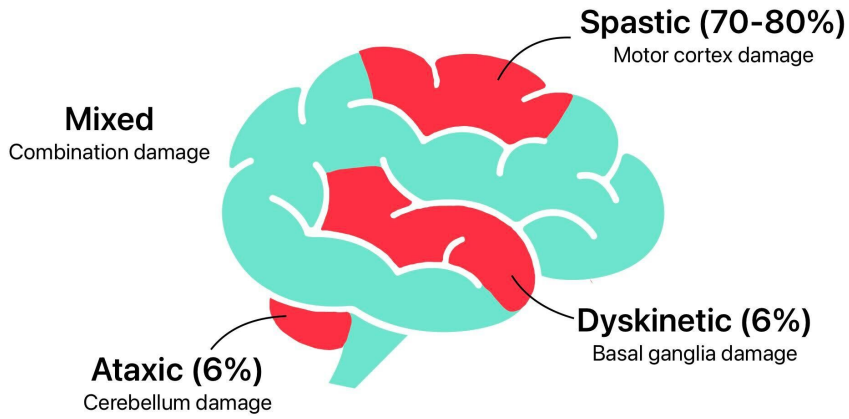


Figure 1.1 – Types of cerebral palsy. The type of CP is contingent upon the regions of the brain that have sustained damage.

preventing contact of the heel with the supporting surface. Spastic equinus is characterised by foot and ankle muscle weakness, and low muscle control, resulting in bone deformities and unbalanced gait during growth [11].

Based on prior literature, surgical intervention after eight years is the most viable option to stabilize the lower extremity and allow the patient to walk as independently as possible, while avoiding overcorrection by avoiding the high-growth phase of the child's development [12]. Nevertheless, long-term follow-up studies report a recurrence rate of up to 48% after surgery [2], [3], [4], [5], and there is insufficient evidence to determine best clinical practice. This recurrence rate reveals a misunderstanding of spastic equinus deformity in children. Although previous studies have investigated the impact of extrinsic factors, such as age and limb type, on the recurrence rate, there is still a lack of morphological analysis on the ankle joint, muscle biomechanics, and tendon strains. To improve understanding of the pathology, it is crucial to examine joint mechanics and bone deformities caused by weak ankle joint musculature *in vivo*. This may enable more informed clinical recommendations.

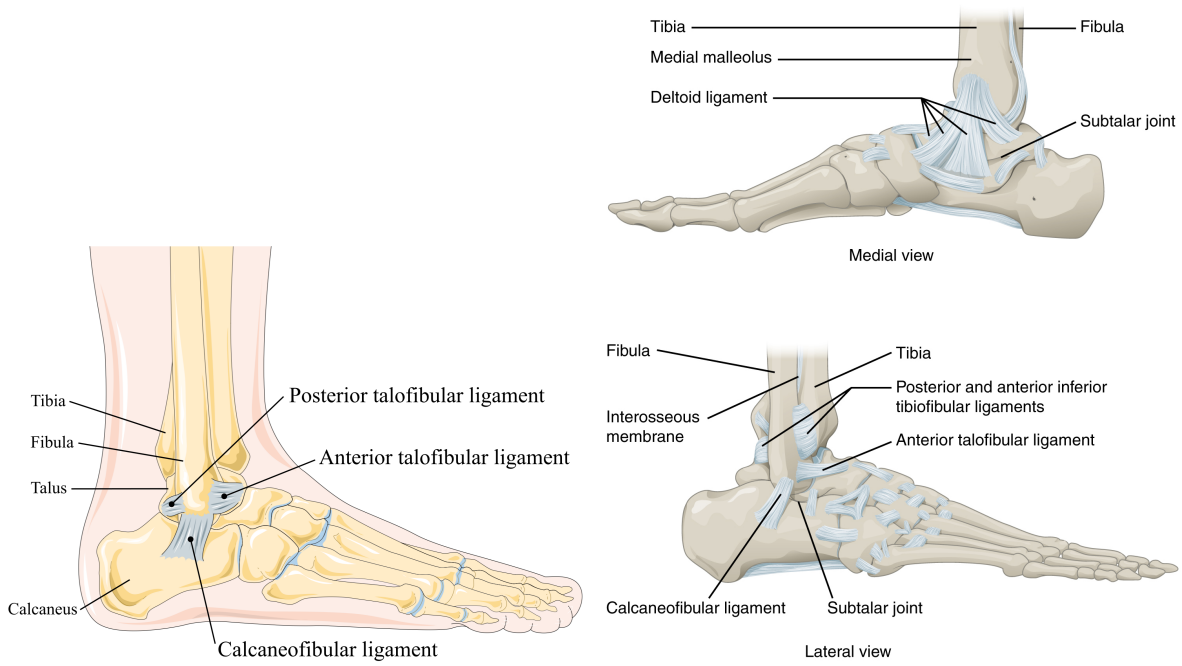
1.2.3 Ankle joint anatomy

Ankle is the part which links the foot and the leg. The ankle joint is composed of three parts: the talocrural joint, the inferior tibiofibular joint and the subtalar joint. The talocrural joint connects talus, fibula and tibia, the inferior tibiofibular joint links the lower extremities of the tibia and the fibula and the subtalar joint refers to the articulation

Figure 1.2 – Patient with an equinus deformity. The ankle cannot be dorsiflex in order to allow contact between the heel an the supporting surface. Image reproduced from Equinus project [13].

between the talus and the calcaneus. Each of these bones is shown in Figure 1.3a.

The joint bones are connected through ligaments and tendons. Ligaments connect adjacent bones and ensure passive joint stability (see Figure 1.3b), while tendons are the link between bones and muscles and contribute to movement. The ankle joint enables the foot to move in three dimensions (refer to Figure 1.4). The primary motion of the joint is dorsi/plantar flexion, which occurs in the sagittal plane and mainly involves the talocrural joint. Inversion/eversion (movement in the coronal plane) and abduction/adduction (movement in the axial plane) are secondary movements of the ankle joint. The combination of these movements results in pronation/supination [16]. All of these movements are performed by muscles, connected to the bones by tendons. The main muscles of the calf and ankle are shown in Figure 1.5. Dorsiflexion primarily involves the tibialis anterior, and secondarily the extensor digitorum longus, extensor hallucis longus, and fibularis tertius. The tibialis anterior begins on the lateral surface of the tibia and inserts into the medial cuneiform bone (between the talus and phalanges) and at the base of the first metatarsal. The primary plantarflexion muscles are the gastrocnemius and soleus (secondary: fibularis longus, fibularis brevis, plantaris, tibialis posterior, flexor digitorum longus, and flexor hallucis longus). Both insert through the Achilles tendon onto the posterior surface of the calcaneus. Eversion involves the fibularis longus and fibularis brevis (secondary: fibularis tertius) and the inversion involves the tibialis anterior, tibialis posterior (secondary: extensor hallucis longus, flexor digitorum longus, and flexor hallucis



(a) Bones of the ankle joint

(b) Ligaments of the ankle joint

Figure 1.3 – Anatomy of the ankle joint: bones and ligaments. Images reproduced from [14] and [15] ©CC-BY-3.0.

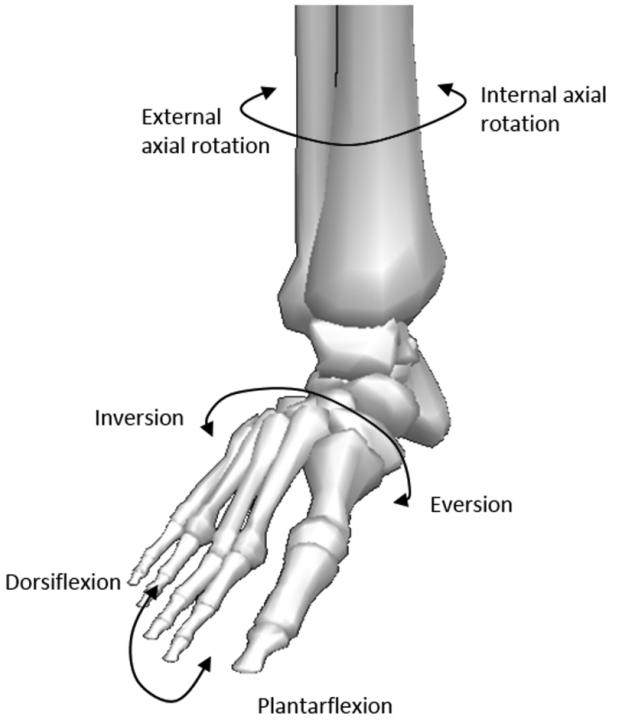


Figure 1.4 – Motions of the ankle. The ankle joint enables the foot to rotate about the three axes of space. The fundamental functional motion is dorsi/plantar flexion. Image reproduced from [16].

longus).

The spastic equinus is caused by a muscle imbalance between opposing muscle groups (agonists and antagonists). On one side, the plantarflexors muscles and on the other side, the dorsiflexors muscles (both muscle groups are described above). The three foremost mechanisms of spastic equinus include: spastic plantarflexors muscles with weaker spastic dorsiflexors; spastic plantarflexors versus normal dorsiflexors; or spastic plantarflexors versus flaccid dorsiflexors. This muscle imbalance frequently leads to bony deformities with growth and postural deformities [17].

As equinus is a dynamic pathology, understanding the influence of muscular imbalance on joint dynamics and resulting bone deformities requires *in vivo* analysis of ankle biomechanics. *In vivo* dynamic imaging provides the opportunity to observe the physiological dynamics of the human body, such as the heart and joints. *In vivo* dynamic imaging is usually performed using ultrasonography (for echocardiographic [18] and joint motion [19] analysis), Computed Tomography (CT) (respiratory, joints or organ studies [20] [21]) or MRI (joint biomechanics [22] [23] or functional cardiac analysis [24]).

Ultrasonography is non-ionizing and limited to analyzing the soft tissues surrounding the joint. On the other hand, CT scans can image bone kinematics but involve ionizing radiation. MRI provides anatomical detail of the musculoskeletal system without exposing the patient to ionizing radiation, making it a valuable tool for studying the musculoskeletal system.

1.3 Magnetic Resonance Imaging for motion analysis

1.3.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging technique providing views of specific anatomy or sequences to monitor physiological processes. It is widely used in medical imaging because unlike other common modalities such as CT (Computed Tomography) or PET (Positron Emission Tomography) scans, it does not use ionizing radiations. Applications range from pathology diagnosis or detection to therapeutic follow-up.

Nuclear Magnetic Resonance (NMR) MRI is based on the NMR principle, describing the behavior of atoms inside an intense magnetic field exposed to electromag-

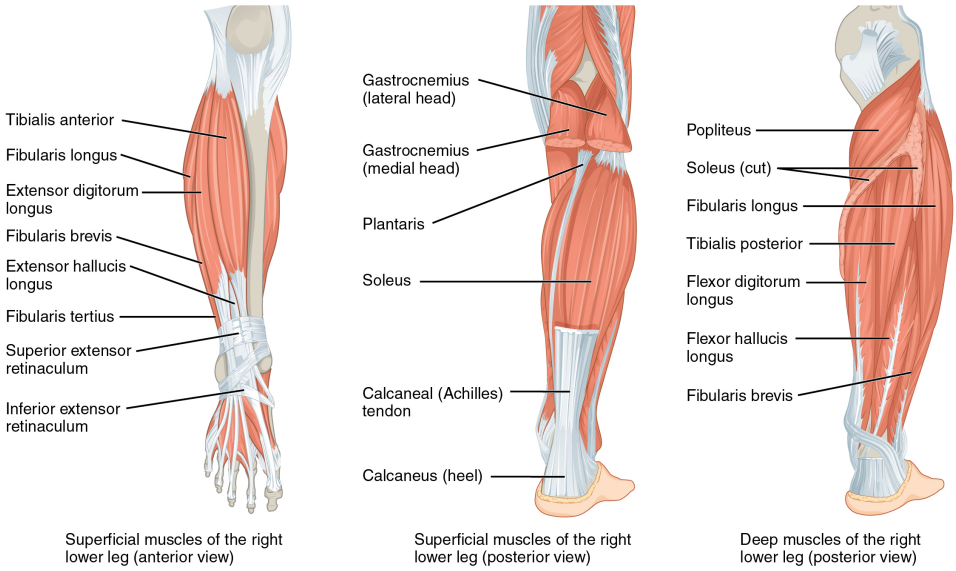


Figure 1.5 – Muscles of the calf and the ankle. The calf and ankle muscles are responsible for ankle joint motion. In treating spastic equinus deformity, surgical procedures often involve lengthening the gastrocnemius, triceps surae, and Achilles tendon to enable dorsiflexion of the foot. Image reproduced from [25] ©Mary Ann Clark/CC-BY-4.0.

netic radio-frequency. A static magnetic field, referred as $B0$, is generated through a supra-conducting coil.

As a variety of nuclei (13C, 23Na), hydrogen nuclei have magnetic properties. The magnetic moment of an atom is function of the nuclear spin. The magnetic moment is decomposed into two components: a longitudinal component and a transversal component. When nuclei are immersed in a static magnetic field, magnetic moments switch from a random orientation (null macroscopic magnetization) to an orientation that is either parallel or anti-parallel with the $B0$ direction (i.e. aligned with or against the field). The longitudinal component becomes aligned with $B0$ while the transversal one (lying on the orthogonal plane to $B0$) reflects the precession of the atom (rotation movement around $B0$ axis). The resulting macroscopic magnetization is the sum of each microscopic magnetic moment for both components. The longitudinal component is proportional to the intensity of $B0$ and the transversal one is equal to zero (since atoms are not rotating in phase).

The excitation phase corresponds to a resonance phenomenon between a pulsed electromagnetic radiofrequency (RF) field and the atoms, at the same resonance frequency (i.e., precession frequency, depending on the $B0$ intensity and atom itself). The RF field

(called $B1$) is applied perpendicular to $B0$, and the resonance leads to a transfer of energy from $B1$ to the atoms, resulting in a macroscopic magnetization lying in the transversal plane.

During the relaxation phase, the system dissipates the accumulated energy as RF radiation to return to its equilibrium state. These radiations form the NMR signal. Relaxation is divided into longitudinal relaxation and transversal relaxation. The longitudinal relaxation corresponds to a restoration of equilibrium where the spins are in a low energy state, i.e. magnetic moments parallel to $B0$. The evolution of the longitudinal magnetization during the relaxation phase follows an exponential curve, characterized by T_1. T_1 is the time required for the longitudinal magnetization to recover 36% of its final value. T_1 depends on the tissue properties. Transversal relaxation is the consequence of magnetic moment dephasing. Interaction between atoms generates field heterogeneities and modifies the precession frequency. The resulting macroscopic transversal magnetization follows an exponential decay curve defined by T_2. T_2 is the time needed for the longitudinal magnetization to return to 37% of its original value. T_2 is systematically inferior to T_1.

Magnetic Resonance Imaging The MR image is formed via the RF radiations emitted during the relaxation phase. These radiations depend on the physical properties of the tissue in which the proton is located, and allow differentiation between organs. The electromagnetic field thus generated produces an induced current into a coil and forms the measured signal.

A secondary variable magnetic field (magnetic field gradient) aims to encode voxels positions in a spatial frequency domain called k-space. The application of these gradients conditions the section of the k-space that is filled: low spatial frequency information in the center of the k-space and high spatial frequency information at the periphery.

1.3.2 MRI sequences

MRI sequences fall into two main categories: spin echo sequences and gradient echo sequences. Each category is subdivided into variations (mostly to reduce acquisition time, allowing breath-holding and thus limit movement artifacts), and some sequences are a hybrid between spin echo and gradient echo. Both are built on Hahn Echoes, discovered by Erwin Hahn in 1950 [26].

Spin Echo Spin Echo (SE) sequence is composed of two RF pulses: a first one at 90° and a second one at 180°. The first is the source of the excitation phase, moving the macroscopic magnetization on the orthogonal plane to $B0$ and generating a Free Induction Decay (FID). The second, called rephasing pulse, is emitted at $T_E/2$ (where time T_E is the time step between the 90° RF pulse and the signal measurement from transversal magnetization) to overcome the local inhomogeneities in the magnetic field that lead to local deceleration of the spins. This operation is repeated for each repetition time T_R (a repetition time corresponds to a line fill in k-space and is the time step between two 90° RF pulses). The spin echo sequence is characterized by T_R and T_E.

The different spin echo sequences are defined by a particular optimized combination of T_R and T_E. T_R determines the value of the longitudinal relaxation of the tissues (depending on T_1): a longer T_R means a more complete longitudinal magnetization recovery. Then, a short T_R enhances the influence of the relaxation time T_1. On the other hand, the transversal magnetization decrease depends on tissue's T_2 (not T_2^* due to rephasing pulse), and so: a larger T_E increases the T_2 influence on the resulting signal.

As a result, T1-weighted sequences are obtained with short T_R and T_E while T2-weighted sequence are produced by long T_R and T_E (very long acquisition times). Proton density weighted, designed to minimize the influence of T_1 and T_2 is then characterized by a long T_R (to inhibit T_1) and a short T_E (to inhibit T_2).

Fast and ultrafast spin echo sequences In fast SE sequences, the acquisition of a single signal at $T_E < T_R$ is replaced by the acquisition of a train of echoes, obtained by applying additional 180° RF pulses at each repetition time. The number of echoes for a T_R is called Echo Train Length (ETL). The phase encoding is changed between each 180° RF pulse to fill multiple lines in k-space in the same slice. As the fast SE signal is acquired at a different time from the standard SE, the image contrast is not the same between the two sequences. Multi-echo SE sequences follow the same principle, except that the phase encoding is not modified within one repetition time. Instead of filling several lines of k-space in one repetition time, several images of the same slice are acquired (with different contrasts, since the acquisition occurs at different T_E).

Ultra-fast SE sequences extend the principle of fast SE sequences by acquiring the entire k-space in a single repetition time. The process can be further accelerated by acquiring only part of the k-space. The missing parts are then reconstructed by symmetry.

Gradient Echo A gradient echo (GE) sequence, unlike a spin echo sequence, relies on a single RF pulse at an angle α of less than 90°. An angle α inferior to 90° reduces the displacement of macroscopic magnetization towards the orthogonal plane, resulting in a remanent magnetization along the longitudinal axis. Thus, α influences the transfer ratio of macroscopic magnetization on the orthogonal plane and the time T_R to recover the longitudinal magnetization (smaller α means shorter T_R and higher α means approaching a spin echo sequence). This set-up allows the use of reduced T_E and T_R, reducing the global acquisition time. The gradient echo sequence, unlike the spin echo sequence, does not apply a rephasing pulse. In spin echo sequences, this pulse corrects magnetic field inhomogeneities, and results in a transversal magnetic decay along T_2. In the absence of this pulse, transversal magnetization decays along T_2^*. A decay in T_2^* makes sequences more sensitive to artifacts linked to the magnetic susceptibilities of adjacent tissues.

In the case of very low T_R, permanent remanent transversal magnetization subsists between repetition times (if the T_R is too short for full recovery of longitudinal magnetization). The processing of this transversal magnetization defines the principal categories of gradient echo sequences.

One approach involves destroying the transversal magnetization (spoiled gradient echo sequences). This is achieved by dephasizing with an RF pulse or gradient. Dephasizing is achieved by randomly varying the phase at each T_R. The sequences are defined by the combination of T_E, T_R and α. The α angle is used to adjust the influence of T_1 on the signal. For instance, a high α boosts the relaxation of the longitudinal magnetization and intensifies the impact of T_1. In contrast to spin echo sequences, T_E weights the importance of T_2^*.

The second method preserves the remanent transversal magnetization and uses it as a part of the signal (steady-state gradient echo sequences). This magnetization is a source of new echoes, added to the Free Induction Decay. Hahn echo sequences exploit the eponymous echo resulting from two RF pulses with the same angle of incidence. This echo amplitude is directly related to T_2. Stimulated echo sequences use at least 3 RF pulses with the same angle of incidence. The amplitude of the resulting echo is linked to both T_1 and T_2.

Ultrafast gradient echo sequences Ultrafast spoiled GE sequences combine a small α with a short T_R and optimize the filling of the k-space. To compensate for the poor contrast induced by the short T_E and T_R, RF pulses are applied before the α RF pulse

series of the classical GE sequence to adjust the ponderation in T_1 or T_2. The T_1 contrast is enhanced by applying a 180° RF pulse and the T_2 contrast by using 90° and 180° RF pulses.

Hybrid sequences This particular type of sequence combines both fast SE and GE sequence techniques. The advantage in terms of imaging lies in the higher sensitivity to calcifications and hemorrhages (less visible in SE sequences), with a lower level of RF pulses than comparable SE and GE sequences. T_2 and T_2^* ponderations are obtained by adapting of the relative number of GE and SE.

1.3.3 Dynamic MRI

MRI sequences described above provide only anatomical details and cannot capture mechanical or functional information. Observation of physiological dynamics is used in particular to monitor cardiac cycles, blood flow or to study joint biomechanics. Dynamic MRI provides both anatomical detail of the musculoskeletal system and no exposure to ionizing radiation [27], making it a valuable tool in the study of the musculoskeletal system.

Dynamic MRI, developed in the 80s initially for cardiac cycles imaging, is also used in the analysis of joint kinematics [27], [28], [29]. It provides a four-dimensional (3D+t) view of a region of interest, allowing the characterization of joint biomecanics. The literature has described at least eight types of dynamic MRI sequences [30]. Two of them appear to be suitable for joints kinematics study: cine phase-contrast (PC) and real-time sequences.

Cine phase-contrast This sequence is part of the motion-triggered imaging class, which relies on identifying the phase of motion with a trigger. In cardiac MRI imaging, it is based on an electrocardiogram (ECG) in order to identify each phase of the cardiac cycle. This type of technique requires repeated motion at a constant speed to preserve quality. Cine PC combines two different techniques: cine MRI and phase contrast sequences. Phase contrast techniques, first used to quantify and describe blood flow, rely on the displacement of spins relative to fixed spins and provide measurements of velocity fields. Cine MRI techniques rely on Gradient Echo sequences where the temporal resolution depends on T_R (see section 1.3.2). In each motion cycle, one line of k-space is acquired for each defined phase of the cycle. Phases are defined and the acquisition is taken according to external triggers, such as an ECG in cardiac imaging. Cine PC

provides a velocity field with pixel resolution, making it very useful for quantifying soft tissue deformation during motion. However, this sequence suffers from significant acquisition time (acquisition in the 3 spatial directions and another acquisition without motion sensitivity), aliasing and ghosting.

Real-time sequences Such type of sequence is the result of a strong acceleration of the imaging process, using parallel imaging and a low resolution matrix. It allows to greatly reduce the acquisition time of a single image and to acquire a motion in real time without the need of repetition or a regular motion pattern. However, it suffers from low spatial and temporal resolution but can benefit from advances in parallel imaging.

1.3.4 Artifacts in MRI

A plurality of artifacts can affect the interpretation of MRI images, thereby disrupting their reading. An artefact is defined as an inappropriate signal in an image with a precise spatial location that is not present on the original object. MRI artifacts can be divided into three main categories: motion-related, hardware and signal-related.

Motion artifacts are the most common. They are caused by the movement of the patient during the scan. These movements are usually caused by breath, heartbeat, blood flow, or any other involuntary or intentional movements. They can cause blur, noise or a particular structured noise called ghosting. Ghosting corresponds to a blurred and shift version of the object. It usually occurs with periodic motion.

Hardware artifacts are those induced by the acquisition system, especially by the magnetic field. For example, B0 inhomogeneity can lead to anatomical distorsions, signal saturation or a reduction in signal-to-noise ratio (SNR). Different magnetic susceptibility (internal magnetization of a tissue) for two adjacent tissues can result in local distorsions in the magnetic field.

Signal-related artifacts refers to those caused by the signal processing, which occurs between the signal acquisition and the image formation. The most commonly observed are aliasing effects and Gibb's artifacts. These typically manifest when the dimensions of an object exceed the field-of-view boundaries. This is due to a spatial frequency error resulting from a low sampling ratio (see de Nyquist–Shannon sampling theorem). Gibbs artifacts emerge adjacent to high-contrast interfaces. They are a direct consequence of employing the Fourier transform to shift from the NMR signal to an image with a finite number of frequencies considered.

1.4 Equinus dataset

1.4.1 Equinus project

A lack of understanding of the pathology and impact of muscle imbalance on ankle biomechanics may contribute to the high postoperative recurrence rate in the treatment of fixed equinus. As explained in Section 1.2.2, surgery for correcting fixed equinus focuses on releasing or lengthening the muscles and does not involve any bone corrections. Previous studies have investigated the impact of CP type, clinical gait parameters, and demographic factors on the recurrence rate after surgery. However, no research has been conducted to analyze the effect of ankle joint biomechanics or muscle mechanics. Understanding the *in vivo* effect of ankle muscle weakness on joint mechanics and bone deformities may lead to a better comprehension of the post-surgery recurrence rate and improve medical management of equinus deformity.

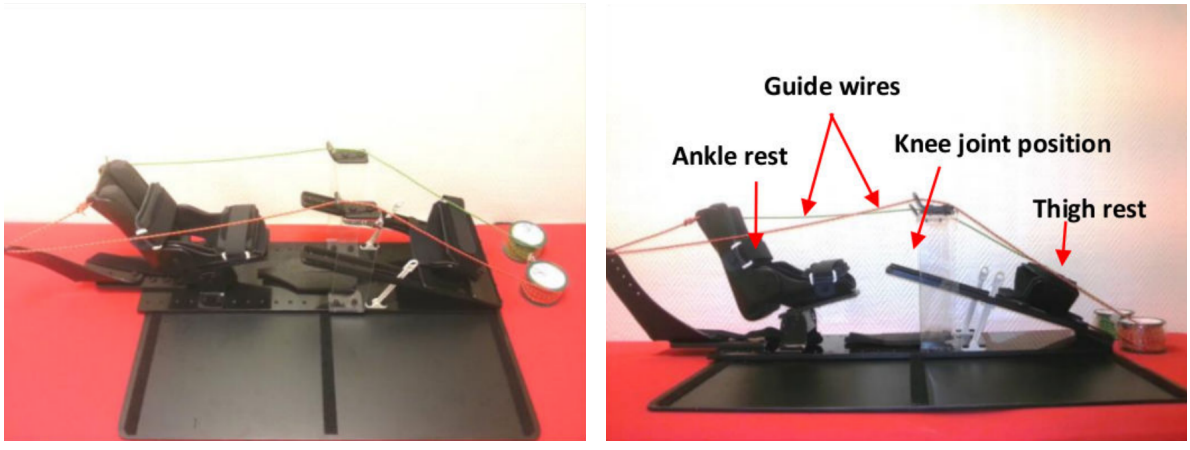
The Equinus Project aims to compare ankle joint and muscle mechanics between children with equinus deformity and typically developing age-matched children [13]. The objective is to provide insights into the relationship between equinus and bone deformities. State-of-the-art imaging techniques, particularly dynamic MRI, provide innovative ways to investigate *in vivo* muscle and joint mechanics. This modality is non-invasive and non-ionizing, and has proven accuracy and precision, enabling researchers to monitor skeletal and muscle motion. It has already been used to study joint motion *in vivo* [24][22] and thus can be successfully applied in equinus studies to evaluate ankle joint and muscle mechanics *in vivo*.

1.4.2 Acquisition protocol

The acquisition protocol includes two different MRI sequences: a high-resolution static 3D MRI scan and three low-resolution dynamic sequences of the ankle joint. Since the equinus foot affects the ability to dorsiflex, dynamic MRI is used to record a dorsiplantar flexion cycle to study the injured biomechanics of the joint. These sequences can provide a better understanding of the mechanisms involved in the pathology, contribute to the quantification and the adaptation of the therapeutic follow-up, and ease the diagnosis. All these data were acquired using a 3T MRI scanner (Achieva dStream, Philips Medical Systems, Best, Netherlands) [31].

Static MRI Static MRI images have a resolution of $0.26 \times 0.26 \times 0.5$mm (T1-weighted gradient-echo, flip angle 10°, matrix 576576, FOV 150mm×150mm, T_R/T_E 7.81/2.75ms, mean acquisition duration: 424.32 s). These images provide a detailed view of the ankle joint and the surrounding tissues.

Dynamic MRI Capturing high-resolution dynamic sequences requires a long acquisition time and repeated motion patterns at a constant speed, making them difficult to handle for patients with musculoskeletal disorders. To limit the patient discomfort, the choice of sequence type is a trade-off between reduced acquisition time and a sufficient resolution and contrast between anatomical structures. The selected one is Real-time T1 Fast Field Echo, allowing a 18s acquisition where the child performs (actively or passively) a single cycle of dorsi-plantar flexion with a rotation speed between 4°/s and 5°/s. Each sequence consists of 15 3D time frames with a spatial resolution of $0.57 \times 0.57 \times 8$mm (flip angle 15°, matrix 352352, FOV 200mm×200mm, T_R/T_E 20.61/1.8ms, acquisition duration: 18.98 s). Passive acquisition is obtained through the use of a fixture, moved by a technician, while the children are asked to relax the lower limb muscle. In the active sequences, the children perform the cycle of dorsiplantar flexion alone and voluntarily.



(a) (b)

Figure 1.6 – Orthotic fixture designed for passive dynamic MRI acquisition. Guide wires are used by the technician to control the dorsiplantar flexion cycle. Foot and lower limb are secured with straps. Image reproduced from [31].

1.4.3 Cohort

Initially, the study included data from 24 children: 13 with a typical developement and 11 with an equinus. Each patient have at least one complete static T1 MRI image, as well as at least one dynamic MRI sequence to be included in the dataset. Four subjects were initially excluded for the following reasons: no static T1 MRI, no static image (T1 or T2), missing part of the foot on an image or blurred image. After completing the initial filtering process, 20 subjects were deemed eligible for the study: 11 typical and 9 with equinus between the ages of 6 and 13. The age distribution within this population is shown in Figure 1.7.

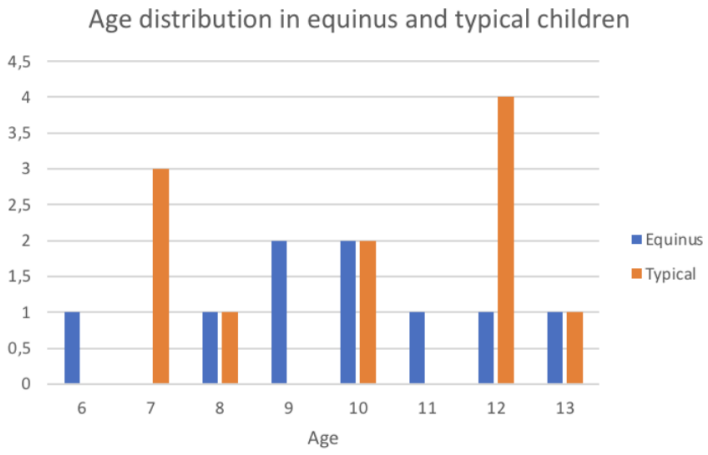
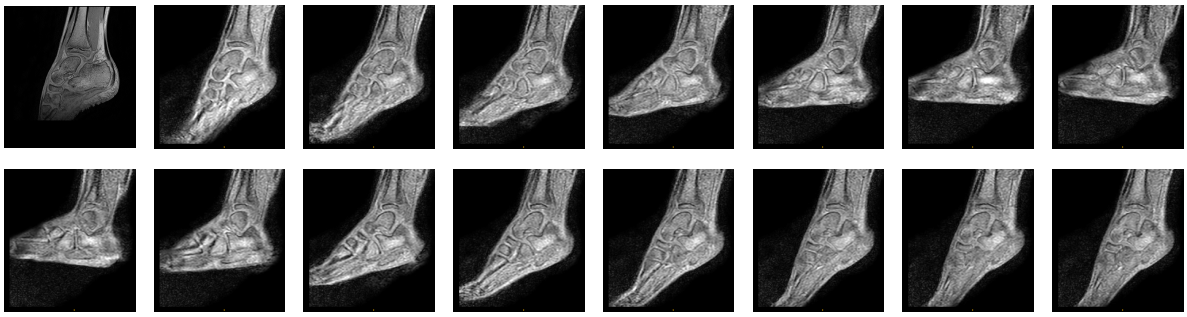


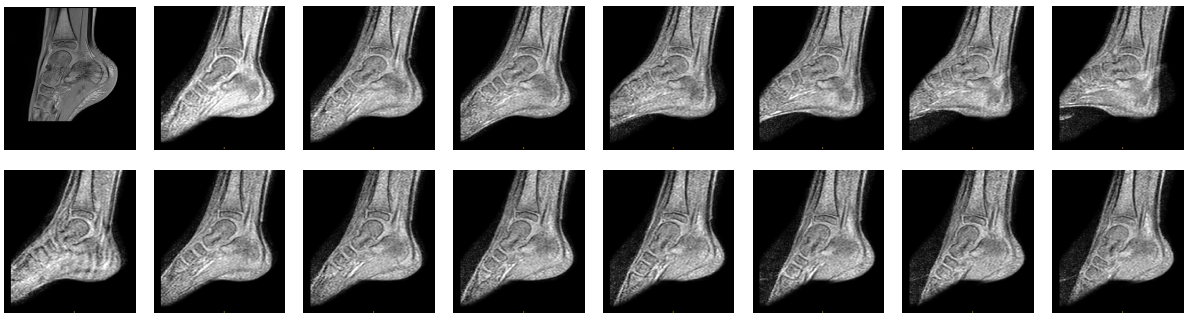
Figure 1.7 – Age distribution of equinus and typical populations.

1.4.4 Source data

The final dataset includes 20 subjects: 11 typical and 9 equinus. For each subject, there is one static image of the ankle joint with a spatial resolution of $0.26 \times 0.26 \times 0.5mm$. The number of dynamic sequences depends on the subject and ranges from 2 to 5. As mentioned above, each sequence contains 15 time frames with a temporal resolution of 1.2s. Each frame has a spatial resolution of $0.57 \times 0.57 \times 8mm$. There are 62 dynamic sequences in total: 32 from typical developing subjects and 30 from equinus subjects. Figure 1.8 shows these data for two subjects (equinus and typical). The first image corresponds to the static MR image while the next 15 correspond to the successive frames of the dynamic MR sequence, showing a cycle of dorsiplantar flexion.



(a) Subject with equinus.



(b) Subject with typical development.

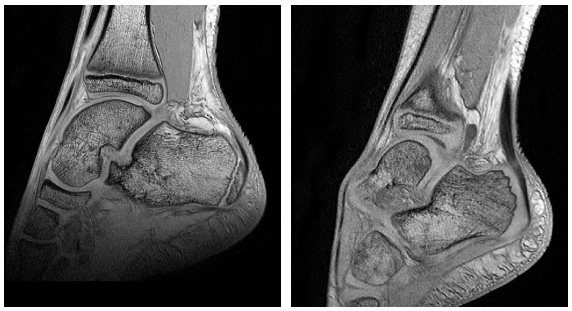
Figure 1.8 – Examples of source data for two subjects: one with equinus and one with typical development. For each, the first image (top left) is a static MRI slice, and the following 15 are extracted from the successive frames of the dynamic sequence.

1.4.5 Variability in source data

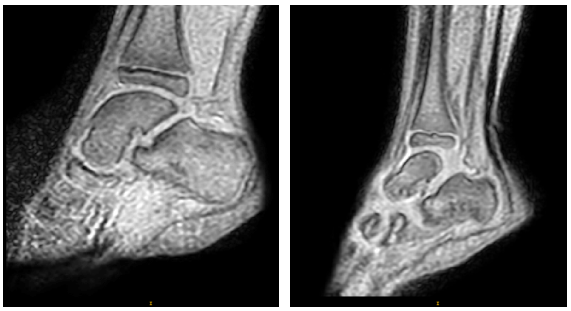
The dataset is made up of images of different subjects, from both static and dynamic MRI. Figure 1.9 shows some of the sources of variability observed within the data. High data variability offers a more complete insight of the situation under study and favors a more general analysis. Low variability makes the analysis more specific, but less robust to new data.

The cohort includes patients with equinus and typical development, from 6 to 13 years old. Children with equinus are subjected to bones deformations and foot abnormalities compared to typical children. Figures 1.9a and 1.9b illustrate examples of morphological differences between equinus and typical subjects on both static and dynamic images. In addition, despite the fact that all the children selected are within a six-year age range, this window corresponds to a period of major development in the musculoskeletal system. In particular, the growth plates (the cartilage on the child's bones) shrink and turn into bone in order to achieve a mature skeletal structure. Figures 1.9c and 1.9d show the evolution of child's bones with age on dynamic and static images.

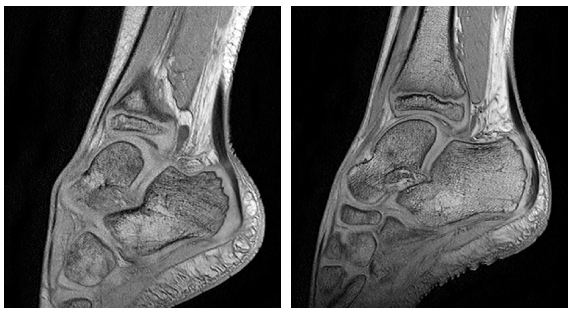
The dataset contains two types of MRI sequences, each providing images with different properties. The first difference is the resolution gap between static and dynamic images. While static images have a spatial resolution of $0.26 \times 0.26 \times 0.5mm$, dynamic images have a spatial resolution of $0.57 \times 0.57 \times 8mm$ for each frame, i.e. a factor of 2 in the sagittal plane and 16 in the orthogonal axis. In addition, since both modalities suffer from anisotropy, it is stronger in dynamic data. Dynamic and static slices are shown in Figure 1.9. In addition to their lower spatial resolution, dynamic images are more prone to noise and various artifacts. Figure 1.9e shows some examples of artifacts present in these sequences. The four images are taken from the same sequence. The first two images (corresponding to two consecutive frames) show variations in saturation between particular frames. These variations are due to inhomogeneities in the magnetic field, which has to be calibrated for a given ankle position. During movement, the position of the ankle varies and can cause this type of artefact. The third image shows both ghosting and aliasing (see section 1.3.4 on the ball of the foot and heel. Finally, the last image shows band artifacts, caused by non-uniformity of $B0$. The use of the orthotic fixture during the passive sequences results in non-rigid soft tissue deformations concentrated in the heel area (see Figure 1.9f). These deformations are not present in the active sequences or in the static images. Finally, there may be variations in the field of view between images of the same modality (see Figure 1.9g for a variation between two static images).



(a) Typical versus equinus foot on static MRI.

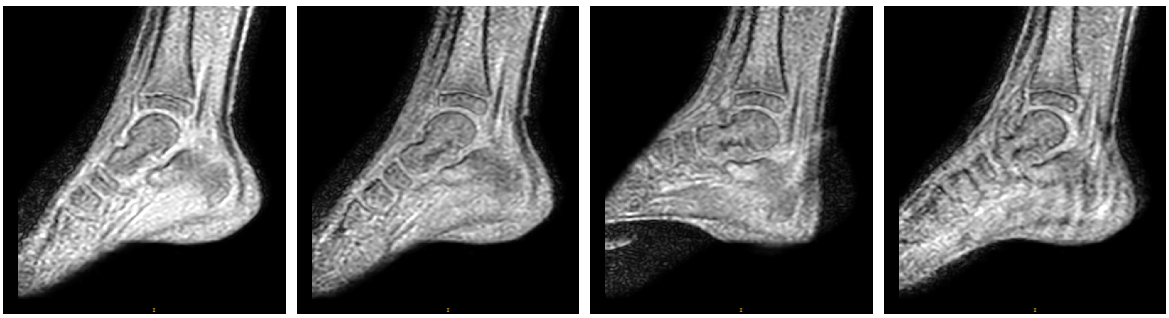


(b) Typical versus equinus foot on dynamic MRI.

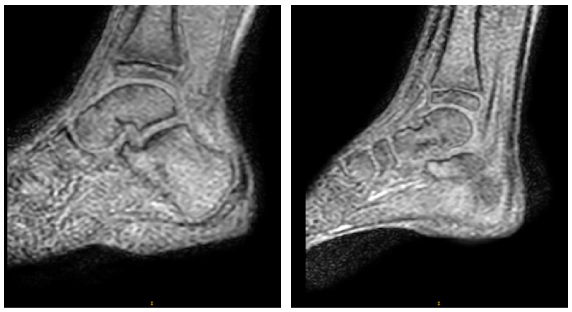


(c) 6.7 y/o equinus versus 10 on static MRI.

(d) 6.7 y/o equinus versus 10 on dynamic MRI.



(e) artifacts in dynamic sequences. From left to right: field heterogeneity between the first two images, causing brightness variation within a sequence, the third image exhibits ghosting and aliasing, and the last image shows band artifacts along the heel.



(f) Active versus passive sequence.

(g) Field of view in static image.

Figure 1.9 – Variability in the source data. In the first two rows, the typical and equinus subjects are the same in both static and dynamic images. (a)-(b) Foot deformations between children with equinus and those with typical development on static and dynamic images. (c)-(d) Effects of aging on bone and growth plate in static and dynamic images. Images come from two equinus subjects aged 6.7 and 9.95 years. (e) A series of examples of artifacts found in a dynamic sequence. (f) Non rigid deformation on the heel, due to the fixture uses for passive motion. (g) Different fields of view between two static MRI scans.

1.4.6 Segmentation

On each static MRI, three bones of interest are segmented: the calcaneus, the talus and the tibia. These bones were chosen because of their major participation in the dorsiplantar flexion movement, which is impaired in subjects with equinus. Dorsiplantar flexion is predominantly active at the talocrural joint (connecting the talus, tibia, and fibula), and the primary plantarflexion muscles are connected on the posterior surface of the calcaneus. Because the influence of the fibula in the talocrural joint is less important than the talus and tibia in this particular motion, it was not considered.

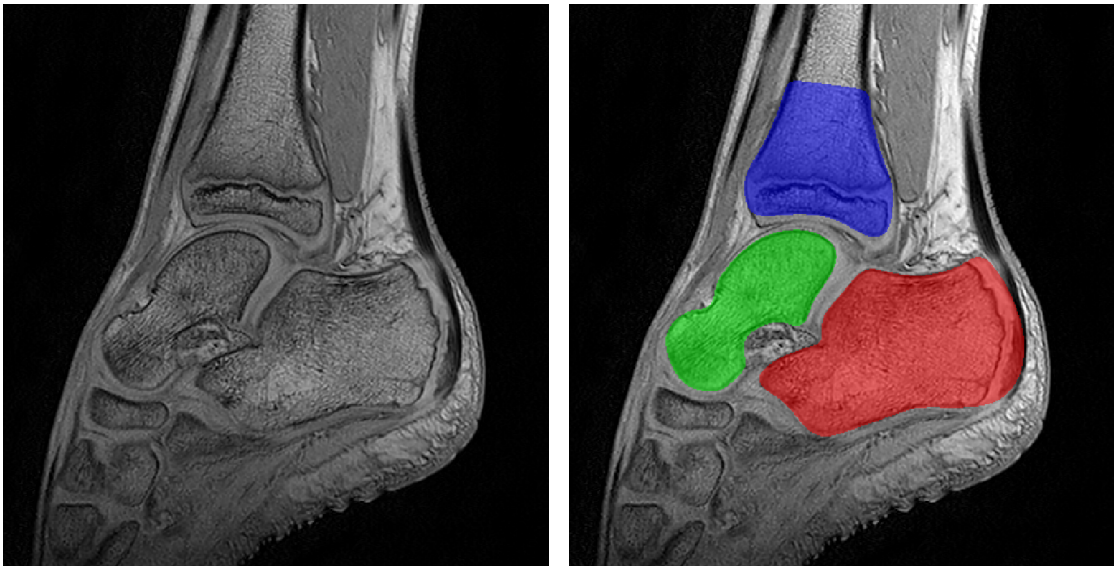
The segmentation is performed using a semi-automatic approach. Given the age variability among subjects, which results in growth plate variations between subjects, the bone segmentation is defined to include growth plates. The segmentation does not include articular cartilage. In the initial stage, three subjects of different ages (7, 10, and 12 years old) are manually segmented by human experts using ITK-Snap [32]. The remaining subjects are then segmented using a registration-based label propagation method, followed by manual correction if necessary. In the final stage, a patch-based UNet trained on the expert-based segmentations is employed to refine the generated segmentations, followed by manual correction if necessary. Figure 1.10 shows the resulting segmentation of the three bones for an equinus subject.

1.4.7 Previous works

To evaluate the *in vivo* biomechanics of the ankle joint and associated muscles, several studies have been conducted on these data. These studies focus on two main areas: analyzing the biomechanical parameters of the ankle and enhancing the resolution of dynamic sequences.

In 2022, Cheng et al. demonstrated the difference in talus and calcaneus volumes between typically developing children and those affected by ankle equinus. They also illustrated the bone deformities between the two groups. Furthermore, they highlighted the absence of a common deformation pattern for all subjects in the equinus group [33]. In 2019, Makki et al. proposed a method for quantifying biomechanical parameters, including ankle joint space width during movement [34] and *in vivo* kinematics such as the temporal non-rigid deformation of the joint [35].

The second part of the method focuses on improving the resolution of the dynamic sequences acquired in the Equinus project, which suffer from low temporal and spatial



(a) Original static image

(b) Segmentation of three bones of the ankle joint

Figure 1.10 – Segmentation of the talus (green), calcaneus (red) and tibia (blue).

resolution. In order to ease the study of the ankle biomechanics by clinicians, several studies have been made to improve both temporal and spatial resolution. Makki et al. proposed a method to reconstruct missing frames in dynamic sequences using a spatio-temporal Log-euclidean polyrigid registration framework [36]. This approach exploits the individual rigid registration of bones to generate a nonlinear joint deformation field between successive time frames. In 2018, Makki et al. proposed a method for synthesizing high-resolution dynamic MRI by fusing locally rigid registration fields [37].

1.5 Conclusions

This chapter provides the clinical context of ankle equinus and cerebral palsy in children. The Equinus project aims to study joint biomechanics in patients with ankle equinus. Compared to ultrasonography and CT scans, dynamic MRI provides accurate anatomical visualization of both *in vivo* joint and muscular biomechanics without any ionizing radiation.

The Equinus dataset comprises 20 subjects suitable for inclusion in this work (13 with typical development and 11 with equinus) for whom a static T1-weighted gradient-echo

MRI scan with a resolution of $0.26 \times 0.26 \times 0.5$mm, and between two and five real-time dynamic T1 fast field echo sequences of 15 time frames with a resolution of $0.57 \times 0.57 \times 8$mm were acquired. For each static image, the three bones involved in the dorsiplantar flexion movement are segmented using a semi-automatic approach. The dataset exhibits significant anatomical variation due to differences in pathology and patient age. This includes variations in bone shape between typical and equinus children, as well as differences in the importance of growth plates depending on age. Additionally, image resolution, anisotropy, and the presence of artifacts contribute to variability within the data. These data have already been used as the basis for studies by Cheng et al. and Makki et al. to investigate ankle biomechanics in children with CP and enhance the spatial resolution of the dynamic sequence through a registration-based method.

The Equinus project's dynamic data suffer from low spatial resolution owing to the short acquisition time. Conversely, the static images benefit from a higher spatial resolution, but do not display the ankle's biomechanics. To ease diagnostics and therapeutical follow-up, and improve pathology comprehension, we aim to synthesize a high-resolution dynamic MRI sequence from a low-resolution dynamic MRI sequence and a high-resolution static MR image. Methods for image synthesis using deep learning offer the best performance in the literature for multi-modal synthesis tasks, with highly realistic results. In this application, we focus on a subdomain of image synthesis called 'image-to-image translation'. The next chapter introduces the domain of image synthesis with a focus on image-to-image translation, which represents the technical application field of this thesis.

Chapter 2

Deep learning for image synthesis

Abstract

The methodology used in this thesis is based on the application of deep image synthesis techniques for the generation of high-resolution MRI sequences. This chapter first addresses the data-related issues inherent to deep learning, before introducing the concepts of image synthesis. This research focuses on two major frameworks for image-to-image translation with unpaired data: Generative Adversarial Networks (GANs) and disentangled representations. This chapter also presents a review of the latest metrics for evaluating image synthesis methods, both with and without reference. Additionally, the limitations of dynamic MRI synthesis models are discussed, which have motivated the work proposed in the following chapters.

2.1 Introduction

Over the past decade, Artificial Intelligence (AI) has gained significant popularity and has been increasingly integrated into a wide range of systems used in daily life. Specifically, AI has exhibited its efficiency in state-of-the-art domains, and ongoing advancements contribute to its continual progression. This section briefly introduces the main deep learning concepts and architectures for image synthesis. We first provide reminders on AI and neural networks, and then define the concept of generative modeling and the associate common architectures.

2.1.1 Artificial Neural Networks

The first artificial neuron model is proposed by Warren McCulloch and Walter Pitts in 1943. This mathematical model is a simplified version of the biological neuron and has been a foundation in the development of modern neural networks. When multiple artificial neurons are connected together, it is called an **artificial neural network (ANN)**. ANN neurons are typically arranged in layers. The neurons in each layer get input from the neurons in the previous layer. The initial layer of the network is known as the **input layer**. In contrast to the additional layers of the ANN, it takes as input the data to which the ANN is being applied, rather than processing the outputs of a previous layer. The last layer of the ANN is called the **output layer**. The **hidden layers** refer to the layers between the input and output layers. The number of layers within a neural network varies depending on the application and the desired level of complexity. As each neuron models a non-linear mathematical operator due to the non-linear activation function, a neural network is capable of modeling a non-linear function. The complexity of the function is determined by the number of neurons within the network.

The learning of an ANN consists of minimizing, through an iterative process, the difference between the desired response to a problem and the prediction of the network. In practice, this difference is represented by a **loss function**. When the desired response is known, it is referred to as **ground truth**. During training, the network's output predictions are adapted to the specific problem by iteratively adjusting the weights of each neuron. Weight adjustment is calculated based on the value of the loss function and the importance of its contribution to the prediction. Learning is an optimization problem, which objective is to minimize the loss function score by adjusting the network weights.

There are two main types of learning: **supervised learning** and **unsupervised learning**. Unsupervised learning is applied when a neural network is used to identify underlying patterns in the data on its own, without external guidance. Supervised learning is employed when a specific answer to a particular problem is desired. Teaching a neural network to classify images is an example of a supervised learning problem, since the correct labels are provided during training. In contrast, identifying risk factors for a disease using a network involves unsupervised learning, as the relevant factors are not known beforehand.

2.1.2 AI, Machine Learning & Deep Learning

Artificial Intelligence (AI) is an almost recent field of research, began for about sixty years. Since Alan Turing is the first to discuss the potential intelligence of a machine [38], the term AI is attributed to John McCarthy from the Massachusetts Institute of Technology. Since its inception, the field has periodically alternated between periods of infatuation and decline (related to the limitations imposed by material constraints in the 1960s, then to the difficulty of formalizing complex decision rules in the early 1990s), but the 2000s mark a new upswing for the discipline. The combination of access to large amounts of data and the development of graphics processing units (GPUs) capable of significantly accelerating computations has led to massive growth in this discipline [39].

The term AI encompasses all the techniques that make it possible for computers to emulate human behavior in order to solve a given problem with minimal human intervention. While early AI techniques focused on handcrafted extraction of relevant features and programming of explicit decision rules, often challenging due to the difficulty of formalizing all the steps in a decision process, **machine learning (ML)** algorithms aim to learn from task-specific training data the way to achieve the considered task.

> *"It is considered easier to explain to a child the nature of what constitutes a sports car as opposed to a normal car by showing him or her examples, rather than trying to formulate explicit rules that define a sports car."* [40]

ML algorithms are able to capture the underlying patterns and information in the data, and these characteristics make them more suitable for real-world applications. As early AI algorithms, they take as input handcrafted features extracted from the training data. However, they differ in that they automatically learn the model-building required to solve the problem. Notable algorithms include random forest, k-nearest neighbors, and ANNs. A significant number of these algorithms are regarded as "white box," indicating that they can be interpreted by humans.

Deep Neural Networks (DNNs) constitute a subcategory of ANNs that are characterized by a larger number of layers. Although there is no clear distinction between ANN and DNN, DNN are often described as ANN with more than one hidden layer. They are also characterized by more complex operations and multiple activations. Pure ML algorithms and shallow ANN are grouped under the term "shallow ML", while deeper ANN and DNN fall under the term "**deep learning**" **(DL)**. While the performance of ML algorithms is highly dependent on the selected features, DL algorithms overcome this issue by taking raw data as input. This construction makes it possible for DL algorithms not

only to learn and adapt the model building, but also to automatically select the relevant feature for a given problem.

2.1.3 Generative Modeling

Deep learning algorithms can be divided into two prominent categories: *generative models* and *discriminative models*. Discriminative models aim to model the posterior probability describing the conditional probability distribution of labels given input data, i.e. the decision boundaries that separate different classes describing the input data. Main applications include image classification and segmentation where the classes are respectively assigned at the image level and at the pixel level. Generative models, on the other hand, model the input distribution as well as the output distribution from a set of input training data. Once training is complete, this allows for the generation of new data points similar to the training data by sampling from the estimated distributions [41]. Generative models are the foundation of methods for image synthesis (see Section 2.3).

The objective of a generative model is to approximate an intractable probability distribution $\mathcal{X}$ defined on $\mathbb{R}^n$, where n is typically large. The training data is considered as independent and identically distributed samples from $\mathcal{X}$. The accuracy of $\mathcal{X}$ representation increases with the number of samples. In practice, generative models aim to find the mapping function from samples of a tractable distribution $\mathcal{Z}$, defined on $\mathbb{R}^q$ (where q is typically smaller than n), to a data point from $\mathbb{R}^n$, similar to the training data. $\mathcal{Z}$ is known and is typically assumed to be a univariate Gaussian distribution in $\mathbb{R}^q$. Given $z \sim \mathcal{Z}$, the objective is to estimate the mapping function $g : \mathbb{R}^q \to \mathbb{R}^n$ (also known as the generator) such that $g(z) \sim \mathcal{X}$, and more generally $g(\mathcal{Z}) \approx \mathcal{X}$. As a result, once the generator is estimated, it becomes possible to generate samples from the complex distribution $\mathcal{X}$ by sampling from the distribution $\mathcal{Z}$ and applying the function g [42]. In practice, estimating real-world data distributions, such as those used for image generation, can be challenging due to their complex and high-dimensional nature. Given their ability to model complex functions, deep learning-based approaches have become a powerful tool in g function modelling, significantly improving the performance of generative models.

The applications of generative models in image synthesis are further discussed in Section 2.3. The subsequent section describes the main deep learning architectures found in most generative model approaches.

2.1.4 Common Deep Neural Networks architectures

Advancements in Big Data have enabled the widespread use of deep learning in various applications. DL is gaining traction in numerous fields, such as autonomous driving, industrial compliance assessments, retail dynamic pricing or personalized email generation, energy supply optimization, risk analysis and reduction in finance, or in voice assistants. In healthcare, it is used for various purposes such as medical imaging, risk reduction, diagnosis using clinical measurements, cancer classification using gene expression profiles, and gait analysis or EEG signals.

DL is a major breakthrough in AI. DL algorithms have replaced the classic, shallow ML models as reference methods, exhibiting remarkable performance in various fields, ranging from image recognition to language processing. They are particularly effective in analyzing high-dimensional data.

Since 2010, major advances have been successfully developed, such as **Generative Adversarial Networks (GANs)**, transformers and, more recently, diffusion models. This section provides an overview of the three most prevalent types of neural networks: **Multi-Layer Perceptrons (MLPs)**, **Convolutional Neural Networks (CNNs)**, and GANs. Although CNNs and MLPs are not exclusive to generative models and can be found in discriminative model architectures, GANs have been specifically designed for data generation.

Multi-layer perceptron A multilayer perceptron is a fully connected neural network, wherein each neuron in a given layer is connected to every neuron in the preceding layer (see Figure 2.1). Originally introduced by Frank Rosenblatt in 1957, it contains at least one hidden layer. A disadvantage of fully connected networks is the rapid increase in parameter numbers as the number of neurons per layer or the number of layers increases, resulting in higher computational expenses.

Convolutional neural network Convolutional neural networks (CNNs) have been engineered to process data in matrix form, making them particularly suitable for handling 2D or 3D images and 1D signals. In practice, they are commonly utilized in imaging applications as they effectively decrease the dimensions of images while preserving information. They can be applied to diverse applications, such as segmentation, classification or detection.

CNNs use convolutional filters, also known as convolution kernels, which are matrices

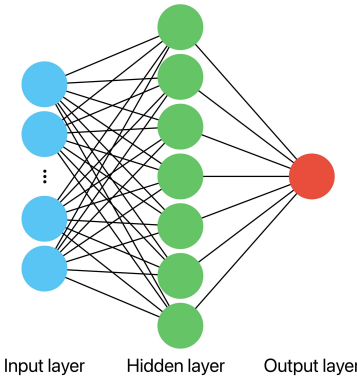


Figure 2.1 – Structure of a multi-layer perceptron.

of variable size (e.g. 3×3) that iterate over the image to extract local features. The convolution operation involves iterating the convolution kernel over the entire image, and is implemented by taking the scalar product between the kernel and the corresponding image portion. This process is repeated as the filter moves through the image (see Figure 2.2). The outcome of performing a convolution of an image and a convolution kernel is known as a feature map. A convolution typically employs multiple convolution kernels. Consequently, performing convolution using n kernels on an image leads to n feature maps. While a convolution operation can be applied to an image, it is also applicable to feature maps. CNNs often comprise a series of consecutive convolutional operations. The convolution operation capitalizes on both the signal structure, which often exhibits adjacent values that are correlated to create a unique pattern, and the position invariance of said patterns [43].

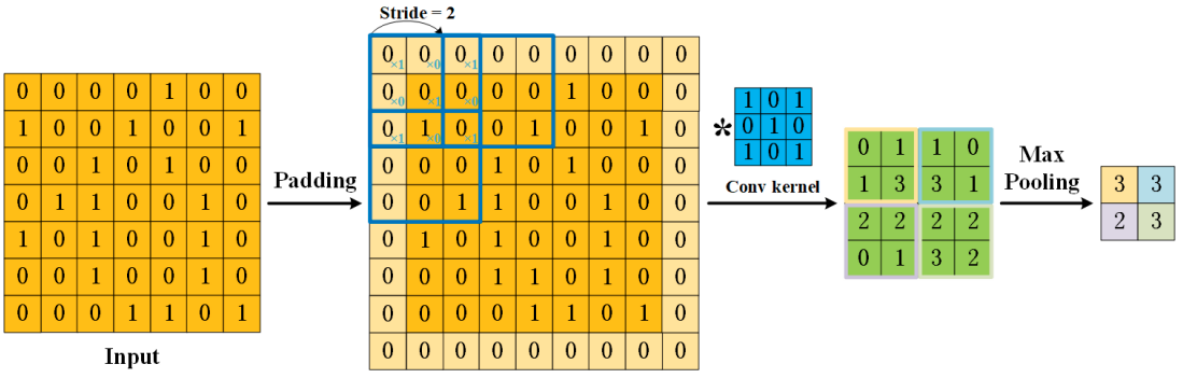


Figure 2.2 – Process of a CNN in 2D. Reproduced from [44]. ©2021 IEEE.

The structure of a CNN typically consists of a sequence of convolutional operations, followed by an activation function, mostly ReLU, and possibly a pooling operation. The purpose of the pooling operation is to reduce the dimensionality by preserving only the most relevant information within a given receptive field. Max pooling (preserving the highest value) or average pooling (averaging values over the receptive field) are the most common operators.

In practice, learning a CNN consists in adjusting the values of the convolution filters. The stacked convolutional layers allow the CNN to learn features in a hierarchical way. For instance, in a classification task, the initial convolution layers extract simple image features like edges. With each subsequent convolutional layer, the extracted features become more complex, such as shapes or objects.

Generative Adversarial Networks GANs belong to the category of deep learning algorithms and are a part of the generative models field. It is assumed that the data set D can be described by a specific distribution. Generative networks aim to reproduce the distribution of dataset D in order to generate new, realistic images that are comparable to other images in D.

GANs were first introduced in 2014 by Goodfellow et al. [45]. They became the state-of-the-art methods for unpaired image synthesis and provided the basis for a significant number of such methods [46], [47], [48]. Since their beginning, they have been widely developed and adapted to multiple applications. Initially used for image generation drawing from noise, they have been extended to conditional data synthesis. GAN relies on two separate networks which learn in parallel and competitively. A generator draws images from its inputs (noise vector or images), and a discriminator classifies images into "real" ones (original images from the target domain) and "fake" (synthetized images from the generator). The generator's objective is to fool the discriminator by generating such realist images that the discriminator cannot identify which are real from which are fake. The discriminator's goal is to classify the real and fake images.

Given z the input of the generator, and x the images from the training set and those synthesized by the generator G. The discriminator D is trained to maximize the probability of assigning the correct label to both real images (with the distribution p_{data}: label = 1) and synthesized images (with the distribution p_g, label = 0). The generator G is trained to minimize the $\log(1 - D(G(z)))$ term, i.e., to reach a state in which the discriminator classifies its outputs as real images. The corresponding objective function is

expressed by the following equation:

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.1}$$

Conditional GANs differ from traditional GANs in the sense that they enable the manipulation of synthesized data through a condition, which takes the form of supplementary input data, such as a class label. Denoting the condition as y, equation 2.1 is modified as follows:

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)}[\log D(x|y)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z|y)))] \tag{2.2}$$

Conditional GANs are used in image editing and generation (to manipulate specific image attributes) [49][50], time series generation [51], or image captioning [52].

2.2 Data for learning

Data is critical to the deep learning process. Unlike machine learning algorithms, which require a manual feature selection stage, deep learning algorithms learn directly from the raw data and identify the relevant features for solving a problem. The distribution, quantity, representativeness, and quality of the training data significantly impact the final model generated during the training process and its ability to generalize.

Following data collection, the data processing pipeline typically begins with a cleaning and preprocessing phase. The cleaning phase consists of the selection of data that will be used to train the model. As an example, in the case of MRI, some images may be subject to high noise levels, the presence of artifacts, or a field of view defect in comparison to others. In the previous chapter, Figure 1.9 illustrates some of these anomalies. An unbalanced data set can also cause the model to learn bias (see Section 2.2.4). The preprocessing phase includes the processing done on the data prior to training, including flipping, cropping, and normalization (see Section 2.2.1). In practice, data processing may depend on the specific application. Once the data has been preprocessed, it can be fed to the model for training. If the amount of training data is too small, a data augmentation phase can be performed during training (see Section 2.2.2).

2.2.1 Preprocessing

Preprocessing is a crucial component of deep learning pipelines, designed to clean and standardize the data. Data standardization process is highly dependent on several factors, including the data itself, the neural network being utilized, and the specific applications involved. For instance, MLP requires all input data to be of the same size, while a CNN takes inputs of different sizes. However, in practice, standardization always includes a data normalization step. Data normalization is the process of adjusting data to the same scale. This is a significant aspect of deep learning that enhances convergence, minimizes the influence of outliers and ignores the magnitude impact in data, and promotes generalization, among other benefits. In image processing, the most typical normalization techniques are feature scaling and z-score normalization. Singh et al. list the various data standardization techniques in more exhaustive detail [53].

Feature scaling normalizes pixel values for each image to fit within a given range, typically between 0 and 1 or 0 and 255. If the original image is denoted as x and the goal is to normalize it between two values a and b, the resulting normalized image x' can be calculated using:

$$x' = a + \frac{(x - x_{min})(b - a)}{x_{max} - x_{min}} \tag{2.3}$$

where x_{min} and x_{max} represent the minimum and maximum values of the original image x, respectively.

Z-score normalization constrains the mean and standard deviation of a population of values (such as the pixels of an image) to 0 and 1. This is achieved by subtracting the mean from each value and dividing the result by the standard deviation of the image or dataset.

The Z-score normalization method usually shows the best results [53] and is more appropriate than feature scaling in medical image classification [54]. Moreover, Z-score normalization is best suited to images with wide dynamics, such as medical images, as it acts on the shape of the distribution rather than constraining image dynamics like feature scaling.

2.2.2 Data augmentation

The size of the dataset used for training has a significant impact on the performance of deep learning models. If the dataset is too small, the model may overfit and result in inaccurate generalization. Overfitting occurs when the model excessively specializes in the

training data, resulting in an error while generalizing to new data. Unfortunately, many applications lack large datasets for model training. Data augmentation is a technique used to expand the size of a dataset by enriching or balancing existing data. This can be achieved through methods such as geometric transformations (e.g. flipping, rotation, translation, cropping, and non-rigid deformation), noise addition, filtering, or data generation using generative models (see section 2.1.4) [55]. The selection of data augmentation techniques should be made with careful consideration to ensure consistency with the intended analysis and to enhance the training process. The application of a large number of transformations does not necessarily guarantee improved performance if the selection is not informed by a robust foundation.

In the context of medical imaging, large datasets are often scarce due to privacy concerns and the high costs associated with imaging time. Consequently, data augmentation may be a suitable approach to address this limitation. This type of data requires dedicated data augmentation engineering compared to natural imaging. Generally speaking, data augmentation relies on simulating variations in the imaging device properties to increase the amount of available data. In natural imaging, modifications may include zooming, color changes, blurring, or changes in the field of view. However, due to the specificity of the imaging devices, data types, and particular artifacts, medical imaging data augmentation should be considered separately. Typical artifacts include ghosting effects in MRI and streak artifacts in CT. Elastic deformations allow for simulating variations in anatomical structures, effectively increasing the number of anatomically diverse subjects.

2.2.3 Paired/unpaired data

In image processing, the term *paired* refers to datasets in which each data point has an exact corresponding ground truth for a given task [46], [56]. Creating such datasets involves manual annotation or accessing to the corresponding ground truths (segmentation masks, corresponding multi-modal acquisitions, class labels...), resulting in a limited number of available datasets. These challenges are compounded in medical imaging, where constructing such a dataset can be tedious (e.g., manual segmentation by an expert), expensive (requiring multiple acquisitions from one or more imaging systems), or inaccurate (due to imperfect registration between images from different imaging modalities).

2.2.4 Data biases

Training data determines the functionality of deep learning algorithms. Biased data can result in propagation and perpetuation of such biases by these algorithms [57] [58]. Recently, studies have demonstrated how biases in data affect ML systems [59]. In the United States, the COMPAS software is used in the judicial system to decide whether to release or remand individuals. This software has exhibited racial bias against African Americans, as documented in [60]. Such biases may result in under-representation of ethnicities in generated faces and failures of facial recognition software, as well as gender bias during recruitment processes. Data biases manifest in various ways and affect numerous applications [59] [61].

In the field of medical imaging, the presence of biased data may contribute to the incidence of diagnostic errors in clinical practice. As demonstrated in [62], the models trained on three public chest X-ray datasets have demonstrated a systematic underdiagnosis bias against several underrepresented subpopulations, including female patients, Black and Hispanic patients, and patients of lower socioeconomic status. The most prevalent biases in medical imaging are threefold: a training data distribution that does not reflect the data distribution in real clinical practice (in terms of gender or ethnicity prevalence), variations in the acquisition protocols, or those introduced by expert labeling. Such biases can be corrected by balancing the dataset, employing data augmentation, or post-processing of the model outcome [63], [64].

2.3 Image synthesis

Image synthesis refers to the process of generating an image that contains the desired content. Image synthesis can be either conditional or unconditional. In conditional image synthesis, an image is generated in accordance with the input data. This input serves as a description of the desired image content and conditions the generation of the final image. This input can assume a variety of forms, including text, audio, brain signals, and visual cues such as another image, semantic segmentation, or depth maps [65]. Some of these cases are shown in Figure 2.3. In unconditional image synthesis, no specific guidance is provided as input to constrain particular aspects of the image. This approach is frequently employed for the generation of large amounts of diverse new data sharing the same semantic content. As an illustration, let us consider the process of generating faces. In conditional image synthesis, a face can be created from text by providing a description,

such as "a woman with blue eyes, dark hair, and freckles." However, this approach can be time-consuming when generating large amounts of data. In contrast, unconditional image synthesis allows for the generation of a random face without any specification, making it a more efficient method for large-scale data generation.

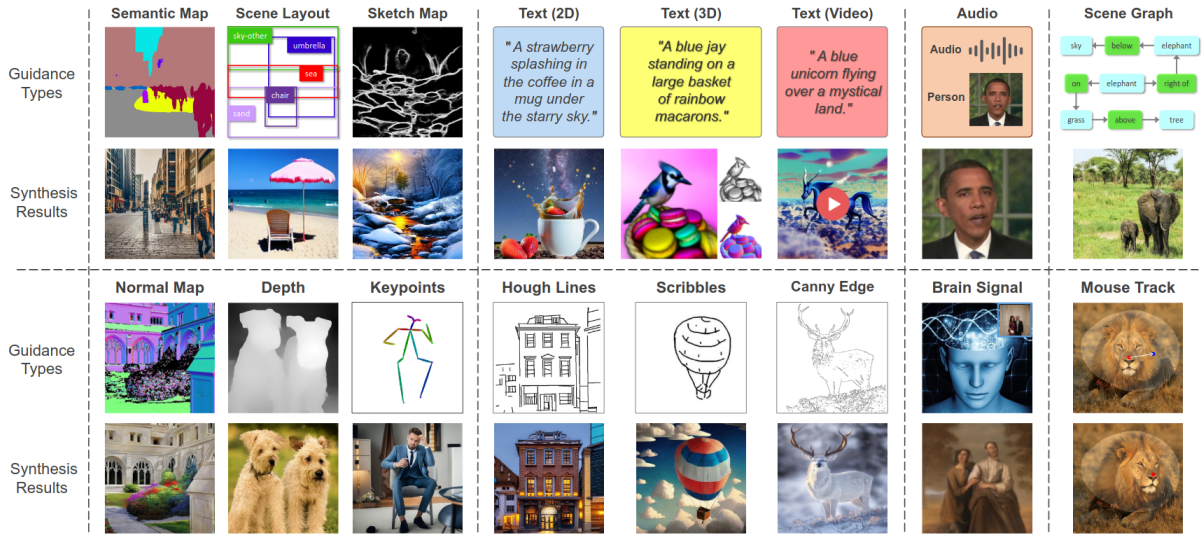


Figure 2.3 – Different categories of image synthesis according to the guidance types. The most common types of guidance are visual information, text, audio, scene graph, brain signal and mouse track. Image reproduced from [65]. ©2023 IEEE.

Today, image synthesis is involved in a wide variety of practical applications in domain crossing for healthcare to security. The most common applications include: animation, super-resolution, computer-aided design or photo editing. It is particularly active in medical imaging where each imaging procedure is not only expensive, but also time-consuming, potentially ionizing and not comfortable for the patient.

This section presents a review of the current state of the art of the literature on image synthesis using deep learning. Section 2.3.1 describes unconditional image synthesis while sections 2.3.2 and 2.3.3 details conditional image synthesis. Section 2.3.2 focus on multimodal image synthesis, i.e. image synthesis from inputs which not lie in the imaging domain while section 2.3.3 details monomodal image synthesis, also referred as image-to-image translation.

2.3.1 Unconditional image synthesis

As previously stated, unconditional image synthesis is defined as a process of image generation devoid of any contextual or descriptive elements. The trained model is expected to generate images whose distribution is similar to that of the training data. The main challenges are twofold: firstly, the newly created images are meant to be different from the ones used for training the model; secondly, they must show a sufficient amount of variability and realism.

Unconditional image synthesis has demonstrated a wide variety of applications, particularly in art generation (in order to create new artworks from a specific artist), supplementing databases (to improve performances of other algorithms by artificially augmented the data amount for training) and even in industrial design (to create new product designs that do not yet exist). Unconditional generative models, including VAEs, GANs, and diffusion models, play a central role in these approaches. Figure 2.4 shows examples of unconditional medical image synthesis on the IXI dataset [66] with three different methods: Diff-Med-Synth [67], DDPM [68] and StyleGAN [69].

2.3.2 Multimodal conditional image synthesis

Excluding the practical applications mentioned above, image synthesis is usually used in a conditional setting where the content of the synthesized image is guided through an input data. Multimodal image synthesis term gathers all the applications where an image is synthesized from any other information modality such as text, audio and sketches inputs. If multimodal image synthesis is almost trivial for the human brain, thanks to the imagination, it is far more complex for a machine. The following paragraphs briefly detailed principles, challenges and popular methods for image synthesis from the abovemention input modalities.

Text-to-image synthesis Text-to-image synthesis is inspired by the way humans mentally visualize scenes when they hear or read a story. The conversion from language to a mental image performed by the human brain inspired researchers to conceive a system able to make the link between the language domain and the visual domain [70]. Text-to-image synthesis allows the representation of sentences written by humans as an image where the semantically meaningful concepts are preserved and represented (layouts, class labels and keywords). The main challenges in this task are the interpretations that text

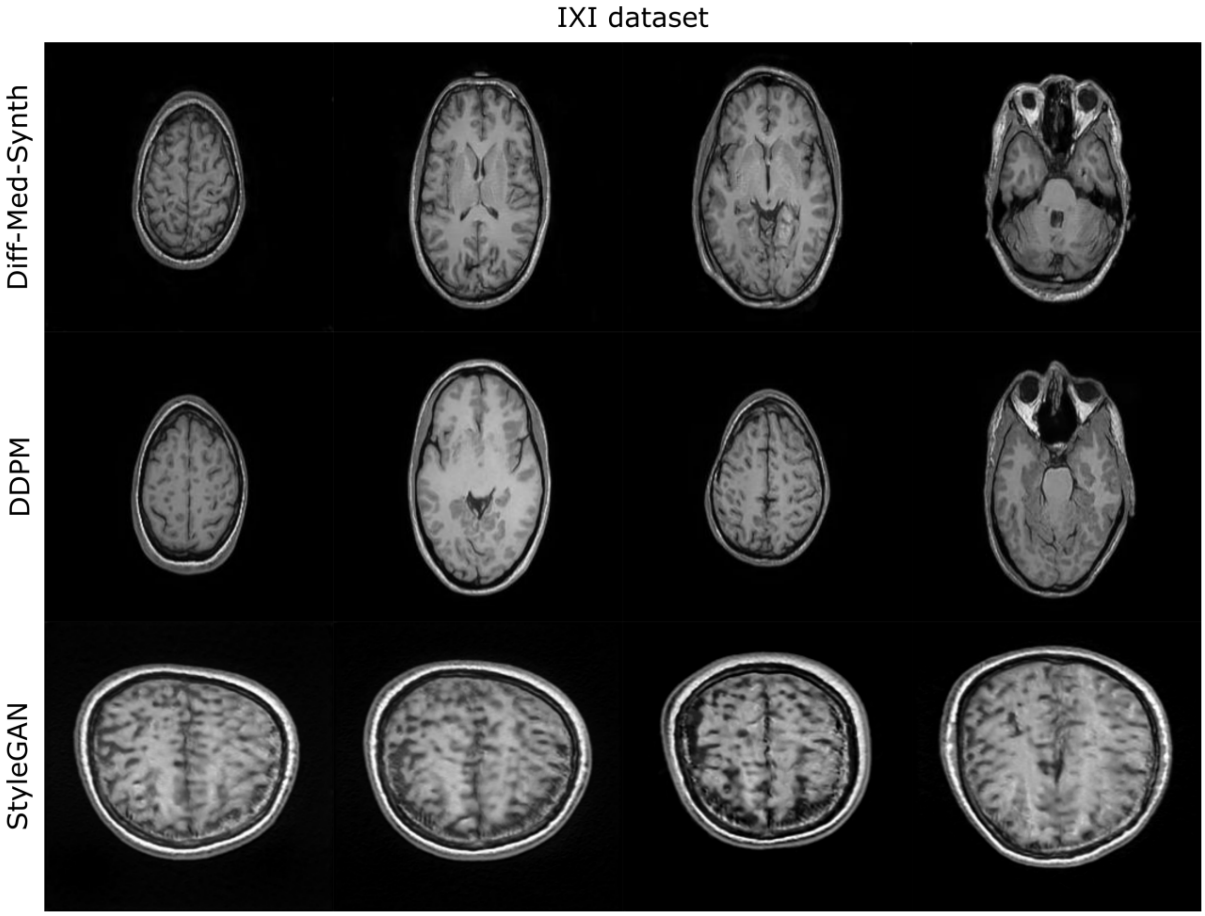


Figure 2.4 – Demonstration of three unconditional medical image synthesis algorithms trained on the IXI dataset for T_1 synthesis. All the image are randomly generated from random noise. Image reproduced from [67]. ©2023 IEEE.

is open to and the possible ambiguity, which can lead to a wide range of possible images for a unique text, making more difficult the output prediction.

The early research focuses on supervised methods to optimize the image content and the textual information, and so face performance gap when the text description was not seen in the training phase. However, since the advance of deep learning methods, and particularly GANs, the developed methods experiences a significant improvement and became more robust to text description unseen in the training phase (StackGAN [71], GigaGAN [72]) [73]. Recently, new methods have emerged such as diffusion-based models (Imagen [74], DALL-E 2 [75]), autoregressive model (Muse [76], DALL-E [77]) and Neural Radiance Fields (DreamFusion [78], Magic3D [79]).

In these families of methods, the condition is typically incorporated after an initial processing stage that is unique to each of the modalities. In this process, the condition is projected into embedding vectors that can be processed by the model. This condition-specific encoder can be either learned with the whole model or be a pretrained model. Learning a representation from a text is not trivial since the encoder should take into account the global structure of the text and the relations between each word. Encoders often rely on traditional text representations such as Word2Vec [80], and deep learning models handling sequences of data as RNN [81] or LSTM [82].

Audio-to-image synthesis If text-driven image synthesis relies on an explicit description of the scene, audio-to-image image synthesis typically relies on more abstract information, such as environmental context and voice tones. Two distinct categories of audio-to-image synthesis can be identified: speech-to-image synthesis, which aims to generate an image that is semantically consistent with the speech, and image synthesis from audio, which does not involve speech. Both types of inputs involve a temporal component, which differentiates them from the other input modality discussed in this section. This temporal component can be used for image sequence generation. However, if an audio input is used to generate an image based on the sounds (for example: musician playing the corresponding instrument [83]), speech-to-image synthesis entails generating an image or an image sequence according to a given speech. This process requires language understanding, analogous to text-to-image synthesis. The speech may be either a description of the image itself [84] or a talk with which the image should be synchronized (for example: lip synchronization [85], [86]).

The audio signal can be represented by features extracted from spectrograms, Mel-

Frequency Cepstral Coefficients or hidden layers of a pre-trained SoundNet model. For the specific case of face generation according to a given speech (also called talking face generation), Action Units are largely used to generate a consistent image according to the audio [65].

Sketch-to-image synthesis Sketch-to-image synthesis is at the cutting edge of image-to-image translation and multimodal image synthesis. Sketch corresponds to an intuitive, fast and simple way to represent an image, and does not require a huge accuracy in the drawing. It is qualified as visual content, yet it is categorized as a distinct modality because the corresponding generating image is not confined by the drawing lines, but rather by the concepts represented. Sketch-to-image synthesis is divided into two main categories: the first consists in querying large image datasets with the input sketch in order to return the closer images and fuse all information into a single image [87], while the second generates images according to the sketch [88][89].

This task is particularly challenging because of the multiplicity of representations for each object, depending of each user and its own mental representation and portraying capacity. Moreover, it may be different level of details across a single sketch (more attention paid to the main subject and less to the background).

2.3.3 Image-to-image translation

Image-to-image (I2I) synthesis is a field of conditional image synthesis where the input data lies in the imaging domain. Since the deep learning methods constitutes the state-of-the-art in the domain, this section will only focus on deep learning methods for image-to-image translation.

> *"Just as a concept may be expressed in either English or French, a scene may be rendered as an RGB image, a gradient field, an edge map, a semantic label map, etc. In analogy to automatic language translation, we define automatic image-to-image translation as the problem of translating one possible representation of a scene into another, given sufficient training data." [56]* ©2017 IEEE

Many problems in computer vision, image processing or medical imaging aim at translating images. The most common includes colorization [90][91], super-resolution [92][93], style transfer [94][95], segmentation [96][97], 3D pose estimation [98], inpainting [99], semantic image synthesis [100][101], denoising [102], image editing [103][104] or cartoon

generation [105][106]. Some of them are illustrated on Figure 2.5.

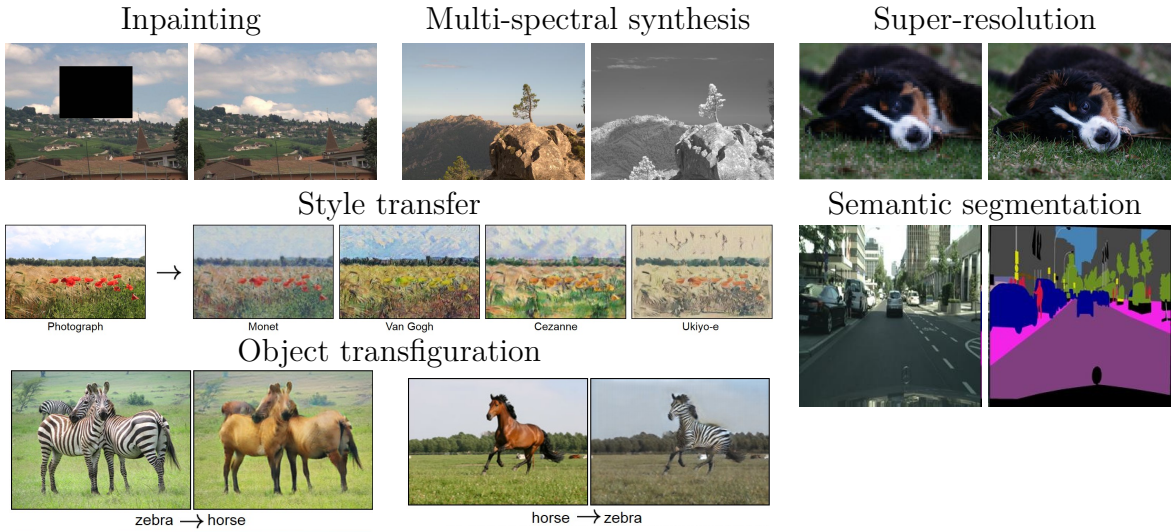


Figure 2.5 – Examples of image-to-image synthesis tasks. Inpainting is the process of filling in missing or degraded parts of an image. Multi-spectral synthesis refers to synthesizing an image in another spectral domain, here visible to near-infrared. Super-resolution encompasses methods for increasing the resolution of an image. Style transfer is the process of transferring the content of a visual source domain to a target visual domain, such as the landscape on a photo rendered in different painting styles [46]. Semantic segmentation groups techniques that apply a label to each pixel of an input image. Object transfiguration consists of translating an object to another of a similar category [46]. Images reproduced from RGB-NIR dataset [107], ImageNet dataset [108], Cityscapes dataset [109] and Zhu et al. [46] ©IEEE 2017.

As described above, image-to-image translation often refers to the process of mapping an input image from a source domain to an output image in a target domain. I2I translation implicitly assumes that an image can be divided into two categories of features: domain-specific features and domain-independent features. Usually, domain-specific features are referred as *style*, while domain-independent ones are referred as *content*. The goal is to preserve the main content of the input image while transferring the style characterizing the target domain. This task relies on the assumption that content can have multiple representations depending on the information to emphasize. In this point of view, a given landscape can be represented by a photo, a infrared map, a depth map or a semantic map. Despite a same content, there are a multiplicity of representations. This formulation assumes that the two imaging domains share some visual or conceptual properties to achieve the desired results. For example, translating human keypoints to realistic rainbows may be difficult to compute in an accurate way. Compared to the other

modalities (especially text or audio), image-to-image translation is less subject to interpretation and nuance. While a text or a talk can capture loads of different concepts and are subject to interpretation, image constitutes a representation of a reality.

More formally, given X a source imaging domain and Y a target imaging domain. Assuming $x \in X$, the objective of image-to-image translation is to generate an image from a source image x while preserving its content and transferring the style of the target domain. The objective is that the translated image x_{XY} seems to have been drawn from the target domain distribution.

Paired versus unpaired image-to-image translation Training data is a crucial aspect of deep learning applications. In the initial image-to-image translation models, training relied on paired images between the source and target domains. Methods typically exploit the correspondence between the input image and a corresponding ground truth in the target domain to compute the error between the estimation and the ground truth through pixel-to-pixel objective functions such as MSE or L1, guiding the training. Figure 2.5 provides an illustration of paired image synthesis for inpainting, multi-spectral synthesis, super-resolution and semantic segmentation. In 2017 and 2018, Pix2Pix [56] and Pix2PixHD [110] models were presented. Both learn image-to-image translation tasks from paired data. However, training with paired data proves challenging in many applications due to the high difficulty or cost of obtaining such datasets. The acquisition of massive amounts of paired data can be extremely difficult or even impossible to achieve. Using horse-to-zebra translation as an example, which is a common example in image-to-image translation, acquiring a dataset wherein a zebra image accurately corresponds to a horse image is unfeasible. Figure 2.6 illustrates the differences between paired and unpaired data for image-to-image translation.

New methods have emerged to address this problem, such as using datasets with limited or no paired information. By learning from unpaired data, it is possible to use large datasets for image-to-image translation without needing ground truth [111], [112]. This setting offers larger applications perspectives, in a broader range of scenarios compared to a paired setting (see style transfer and object transfiguration on Figure 2.5). Multiple strategies have been employed to estimate transformations in the absence of paired data. The **cycle consistency** constraint offers a way to compensate for the absence of pairing constraint. Using the notations mentioned above, $x \in X$ denotes the source image, $x_{XY} \in Y$ denotes the source image projected into the target domain, and $\tilde{x} \in X$ refers

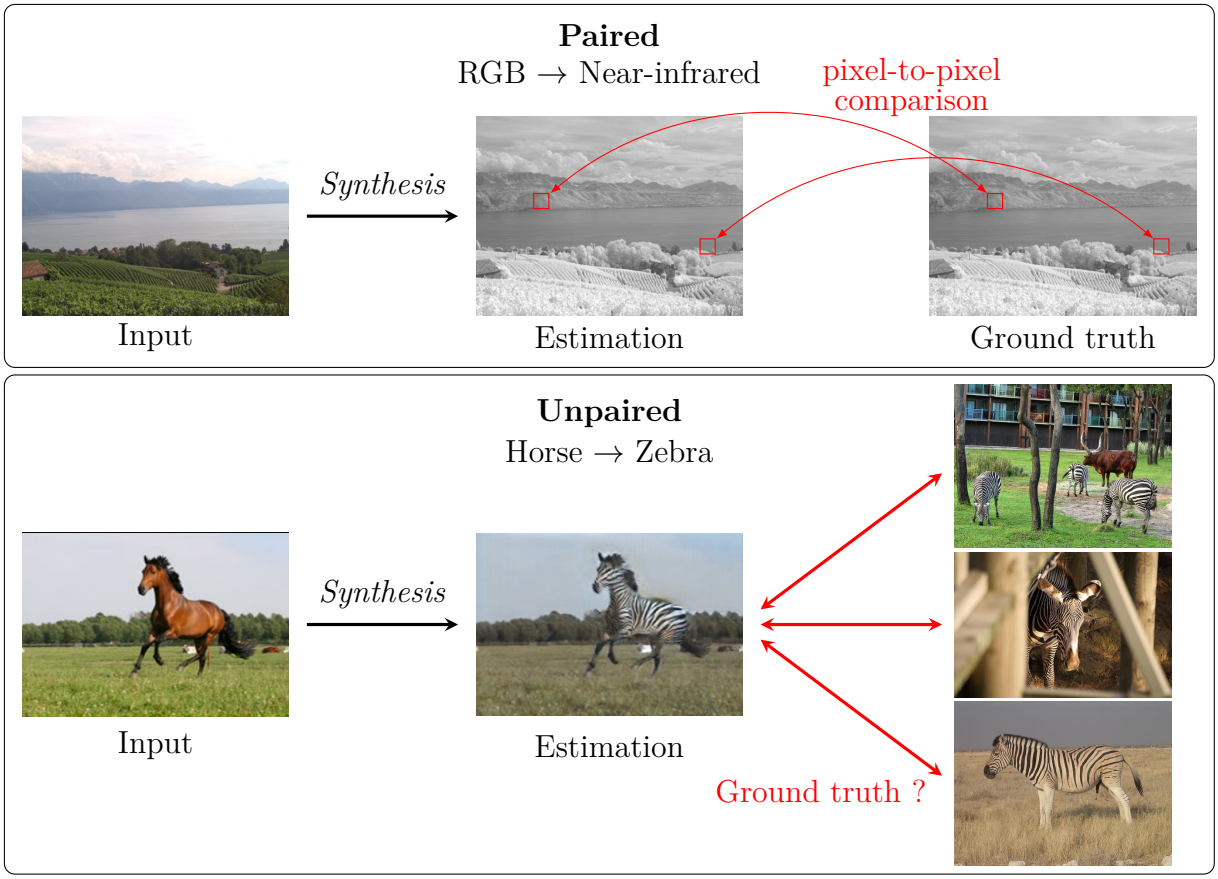


Figure 2.6 – In the paired setting, the estimated image can be compared to a provided ground truth. However, in unpaired I2I, this exact correspondence does not exist. Although some methods have been proposed to address this problem, overcoming this issue is still challenging. Images reproduced from Zhu et al. [46] ©IEEE 2017, ImageNet dataset [108] and RGB-NIR dataset [107].

to x_{XY} projected back into the source domain. The expression of the cycle consistency constraint is as follows:

$$\mathcal{L}_{cycle} = \mathcal{L}(x, \tilde{x}) \tag{2.4}$$

where $\mathcal{L}$ can be any supervised loss, usually L1 norm or Mean Squared Error. Since this technique has demonstrated its effectiveness in unpaired data settings, it also tends to enforce the preservation of source information in the generated image, which becomes a challenge when extensive transformations are required between the source and target domains. Furthermore, the assumption of a bijective relationship between two domains may be overly restrictive in some applications. Consequently, techniques have been devised to eliminate cyclic constraints and to conduct solely a unilateral translation [47], [94]. Despite their overall good performance, these models may be challenging to train due to the significant constraints introduced in their loss functions.

As a result, techniques have been developed to remove cyclic constraints and to perform unilateral translations [47], [94]. Despite their overall good performance, these models may be challenging to train because of the large constraints introduced in their loss functions.

In 2014, Kingma et al. [113] presented a method for image-to-image synthesis from a partially annotated dataset. Such datasets enable cost reduction for data pairing while providing a portion of ground truth. The piece of paired data among the unpaired ones can significantly improve the result of the synthesis. The proportion of paired data can be less than 1% and still enable successful image-to-image translation tasks [114]. All of these techniques require large data sets to accurately estimate the transformation between domains. Few-shot learning techniques refer to all methods that use a minimal amount of training data, even a single training data. In image-to-image translation, methods often leverage knowledge transfer between tasks (also referred to as transfer learning) to adapt I2I algorithms trained on large-scale datasets to small-scale datasets [115][116].

In practice, specific frameworks are favored depending on the training data. Paired image synthesis methods often rely on well-known computer vision architectures such as Residual Networks or UNet, which are trained in a classical way with a paired loss function such as L1 or MSE. In unpaired image synthesis, two main frameworks dominate the state-of-the-art: cycle-consistent Generative Adversarial Networks and Disentangled Representation Learning. Both are discussed in detail in the following paragraph.

Conditional versus unconditional image-to-image translation Image-to-image translation methods typically fall into two categories: non-conditional and conditional.

The former generates an image in the target domain based on an image in the source domain, while the latter requires an image from the target domain or a guidance in addition to the source domain. Providing the model with an image from the target domain gives more control over the synthesized result by specifying features from the target domain (see Figure 2.7) while other types of guidance can edit some particular aspects of the input image.

Unconditional methods are mainly used in cross-domain synthesis, such as CycleGAN from Zhu et al. [46] or Art2Real [117]. Their primary aim is to estimate the transformation that connects one domain to another and frequently depend on a cycle consistency constraint when the data are unpaired. Nonetheless, despite their strong abilities in image-to-image translations, the lack of conditioning in these approaches results in limited control over the synthesis outcome which can lead to realistic images with structural inconsistencies, hinders the manipulation of image features, and are not human interpretable.

On the other hand, conditional methods are mostly utilized for the purpose of altering distinct characteristics present within an image, such as pose morphing [112] or within the domain of arbitrary style transfer [95], [118], [119], [120]. Style transfer encompasses a significant portion of image-to-image translation applications and aims to represent a source image with the stylistic features of the target image. Arbitrary style transfer enables the transfer of any new style, rather than only a predetermined set of styles. Consequently, templates are designed to not only take the image to be modified as input, but also the desired style to be applied. Other methods, such as **disentangled representations**, are inherently conditional.

Disentangled representations are formulated on the assumption that the data distribution can be described by a set of independent and semantically meaningful factors of variation. Learning a disentangled representation aims at discovering these underlying factors and encoding them into separate dimensions [121], [122], [123]. A representation is considered "disentangled" if an alteration of a single factor results in a specific change only along the corresponding mode of variation within the data [124]. Disentangled representations learning frameworks are promising in terms of explainability, generalizability and controllability properties [125]. Assuming a representation to be disentangled, each factor of variation modeled by a corresponding latent code retrieves a particular semantic aspect of a set of images. Thus, the generation of a new image becomes controllable and interpretable. Moreover, assuming the learned representation to be disentangled, then

each factor of the representation is decorrelated from the other, offering better generalization capabilities and transferability than the classical algorithms. These techniques have already been applied to perform segmentation [97] or inter-domain synthesis [126], [127].

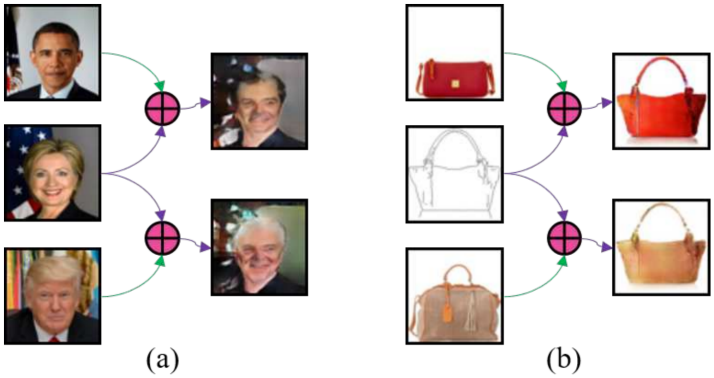


Figure 2.7 – Conditional I2I translation. (a) Conditional women-to-men translation (b) Conditional edges-to-handbags translation. The purple arrow represent the translation flow and the green arrow represent the conditional flow. It demonstrates that conditional I2I translation allow a greater control on the features of the synthesized image in the target domain. Image reproduced from [128]. ©2018 IEEE.

Two-domain I2I versus multi-domain I2I Two-domain image-to-image translation encompasses many problems and constitutes the majority of applications. However, some problems involve more than two domains and are referred as multi-domain image-to-image translation. Algorithms for image-to-image translation between two domains train a model for each translation between a pair of domains. To apply this approach to a multi-domain setting, the system's complexity increases significantly as a model is needed for each translation between all domain pairs. A key challenge in synthesizing images between multiple domains is thus to reduce the complexity of the model.

Among the proposed approaches, two primary categories emerge. The first category views translation between various domains as a combination of translations between two domains. This approach requires the training of multiple domain-specific generation units. To reduce model size and computational cost, most of these approaches rely on a single training to achieve synthesis across all domains. In addition, shared representation spaces can be constructed to reduce the number of generation units required within the model [129]. In the second category, a single generator is used for translations in all the domains. The generator is conditioned by both the source image and a target domain label [130].

2.3.4 Applications in medical imaging

Medical imaging is used in standard clinic practice to confirm or pose a diagnosis (for diseases or injuries), to analyze the body functions such as brain, heart, or body kinematics, for the therapeutical follow up or for interventional radiology. Today, deep learning methods are commonly utilized for medical image processing, serving various purposes such as diagnosis support, medical event prediction, and decision-making or classification. Image synthesis methods serve as the foundation for many of these applications. Their main fields of application are registration, segmentation, inter-modality synthesis, reconstruction and super-resolution.

Image registration is the process of aligning multiple images by analysis of their visual features [131]. In medical imaging, the objective is to align images of one or more modalities and/or one or more patients by matching anatomical structures. This technique finds applications in clinical practice (inter-patient registration for atlas creation, therapeutic follow-up) as well as in other algorithms applied to medical imaging (image segmentation or reconstruction).

Medical image segmentation is the process of extracting specific anatomical structures or structures of interest from a medical image. Manual segmentation by a specialist is a time-consuming and laborious process, whereas deep learning methods offer automatic segmentation alternatives. Segmentations are used for diagnostic purposes, therapeutic monitoring or quantitative analysis: segmentation of brain tumors [132], muscles [133] or vessels [134] or cardiac structures [97].

To obtain complementary information between modalities, many clinical practices require multimodal acquisition protocols. Inter-modality synthesis can limit patients' exposure to ionizing radiation, decrease the cost of multiple acquisitions, and shorten imaging time. For example, CT images can be synthesized from MRI images [135][136][137], or T1 MRI images can be synthesized from T2 [97].

Image reconstruction and super-resolution are motivated by the goal of obtaining high-quality images while minimizing costs and risks to the patient (use of low doses of ionizing radiation, shorter imaging times, etc.) [138] [139]. Super-resolution refers to the process of generating high-resolution images from low-resolution images [140][141]. It differs from interpolation in that it learns the relationship between low-resolution and high-resolution data from the training set, whereas interpolation fills in missing data, often resulting in blurry images. Image reconstruction refers to the field of image restoration in computer vision. Image reconstruction can be utilized for raw data within the realm of compressed

sensing in MRI [142] [143] as well as at image level [144].

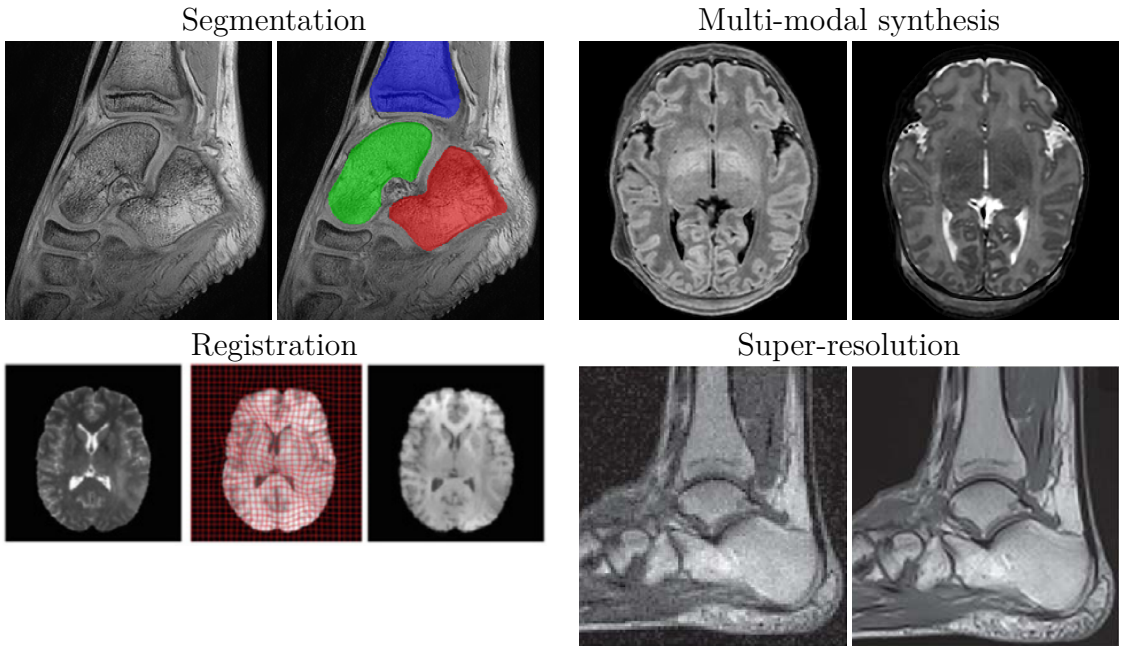


Figure 2.8 – Examples of image synthesis in medical imaging. Each application is detailed in Section 2.3.4. Segmentation consists of the identification of specific anatomical structures, illustrated using data from the Equinus dataset. Multi-modal synthesis is the process of synthesizing an image in a target modality from an image of a source modality, such as T1 and T2 MRI scans [145]. Registration stands for the process of aligning different images in a common coordinate system (Figure reproduced from [121]). Super-resolution methods aim to increase the spatial resolution of given images [146] ©IEEE 2015.

When images are synthesized, evaluating the quality of the synthesis becomes a challenge in itself. While humans excel at judging image quality, designing an image quality metric that captures the perceived quality and aligns with human judgement remains tedious. Furthermore, while well-known and extensively validated metrics exist for paired data, image quality assessment for unpaired data is much more difficult. This section presents an overview of image quality assessment metrics in both paired and unpaired data settings.

2.4 Image quality assessment

Image quality assessment (IQA) gathered the methods used to evaluate the quality of images. These measures are usually designed to express a perceptual quality, according to human perception. IQA methods fall into two categories: objective and subjective [147]. Subjective methods laid on the judgment of humans usually through user studies, while objective ones are built under computational algorithms. Objective IQA can be either divided into three categories, depending on the level of reference information available. Full reference IQA (FR-IQA), also reffered as fidelity/similarity measurement, is used when a reference (ground truth) image, perfectly matched with the test one, is available to do the comparison (examples: colorization, edges to photo, aerial to map...) [148]. Reduced reference IQA (RR-IQA) provides measures when only partial information from the reference image is available (examples: telecommunications) [149]. Finally, no-reference IQA (NR-IQA) is used when there is no sort of information about a reference image, aiming to evaluate the image quality in a blind set up (examples: faces synthesis, summer to winter synthesis, data simulation...) [150], [151].

In practice, FR-IQA methods were developed in a first place, and benefits for a wide variety of methods and often solid physical background. However, in real world applications, NR-IQA or RR-IQA are more applicable, since the ground truth is barely available in applications such as real-time, medical imaging or image synthesis. In particular, NR-IQA is a growing field of interest since the last decade. The increasing demand of image and video services and the frequent lack of ground truth leads to an increased research around this field.

The following sections will describe metrics for each type of IQA and metrics in the specific case of medical imaging.

2.4.1 Full-reference image quality assessment

A great part of the literature is using FR-IQA metrics. Most of them rely on a strong physical background and have proved their efficiency and relevance. They are often computationally low cost and simple for setting and application. The most common include Peak Signal-to-Noise Ratio (**PSNR**) or Mean Squared Error (**MSE**).

Some of the most common metrics are derived from norm functions such as MSE and Mean Absolute Error (**MAE**). MSE is one of the most common metrics in FR-IQA and corresponds to a squared L2 norm while MAE corresponds to a L1 norm. Both MAE

and MSE quantify the prediction error between a simulated image and its ground truth. Given I a simulated image and R its corresponding ground truth, MSE and MAE can be written as follow:

$$\text{MSE}(I, R) = \frac{1}{N}\sum_{i=1}^{N}(R_i - I_i)^2 \tag{2.5}$$

$$\text{MAE}(I, R) = \frac{1}{N}\sum_{i=1}^{N}|R_i - I_i| \tag{2.6}$$

For both, a lower score suggests better results.

PSNR is a widely adopted metric for image quality quantification. Initially developed to assess the quality of a compression algorithm, it expresses the ratio between the maximum possible value of a signal and the power of the error (computed between ground truth and simulated image). The PSNR is usually rescaled into a logarithm decibel scale to deal with wide dynamic range of signals. The formula is detailed in equation 2.7. PSNR is expressed in decibels (dB) and a higher score means a better synthesized image.

$$\text{PSNR}(I, R) = 20.\log_{10}\left(\frac{\max_I}{\text{MSE}(I, R)}\right) \tag{2.7}$$

However, both PSNR, MAE or MSE do not reflect the human perception of image quality. The Structural Similarity (**SSIM**)[152] index was introduced to handle this issue. SSIM index focuses on the structural similarities between two images. It is developed under the assumption that the luminance of an image does not affect the perception of the structures and of the image quality. The SSIM index itself is computed under a three step comparison of luminance, contrast and structure. A luminance comparison function l compares mean intensites of both images. After a luminance subtraction step from both images, the function c computes and compares the contrast from both images through standard deviation. Finally, after a luminance subtraction and a contrast normalization, a structure comparison function s is computed on the resulting images, invariant in contrast and luminance. The structure comparison is computed through correlation coefficient. The SSIM index is obtained by combining all of these functions (see equation 2.8), each one independent of the others:

$$\text{SSIM}(I, R) = [l(I, R)]^{\alpha}.[c(I, R)]^{\beta}.[s(I, R)]^{\gamma} \tag{2.8}$$

The parameters α, β, and γ are introduced to weight the relative importance of each term in the resulting measure.

2.4.2 No-reference image quality assessment

Despite the considerable promise of NR-IQA techniques, it remains challenging to quantify the quality of generated images in the absence of ground truth. Currently, there is a scarcity of established reference metrics for evaluating these techniques. Typically, these metrics aim to emulate the human visual system, which does not necessitate the use of external references for assessing image quality. The most popular includes Fréchet Inception Distance [153] (**FID**), Kernel Inception Distance [154] (**KID**), or Learned Perceptual Image Patch Similarity [155] (**LPIPS**).

A great part of these metrics are built on the activations of secondary neural networks. Since the last decade, new methods have emerged to evaluate those methods. The Inception Score [156] (**IS**) is one of the first appeared to quantify synthesized image quality. It both evaluates the diversity of the generated images and their realism and is known to be well-correlated with the human judgment. It exploits the intermediate activation of an Inception network pretrained on ImageNet [157] for classification task. The closer to 1 it is, the better quality the generated images are.

Following this idea, the FID score was designed as an improvement of the IS, and became the widely used metric for NR-IQA. In contrast to the IS, the FID score compares the features of both real and synthetic images extracted from an Inception v3 model trained on the ImageNet dataset:

$$\text{FID} = d^2\Big((\mu, C), (\mu_w, C_w)\Big) = \|\mu - \mu_w\|_2^2 + \text{Tr}\Big(C + C_w - 2(CC_w)^{1/2}\Big) \tag{2.9}$$

where d(.,.) is the Wasserstein-2 distance [158], (μ, C) the multivariate Gaussian obtained from the intermediate activation of the Inception network fed by generated data and (μ_w, C_w) the Gaussian obtained intermediate activation of the Inception network fed by real data. μ stands for the mean and C for the covariance matrix. Although FID was initially developed for natural image analysis, there is evidence that it can also be relevant in medical imaging [159]. A lower FID means a better quality of synthesis.

KID is constructed as a variation of FID. Instead of FID, it computes the squared maximum mean discrepancy between Inception v3's intermediate activations from real images and generated ones, and relaxes the Gaussian assumption by using a polynomial kernel:

$$\text{KID} = MMD(f_{real}, f_{fake})^2 \tag{2.10}$$

where f_{real} define the intermediate features from real images and f_{fake} those from gen-

erated images. The kernel for MMD computation is defined as $k(x, y) = (1/dx^T y + 1)^3$. Moreover, it does not assume an activation distribution's parametric form. KID score reflects the shared visual similarities between I and R. A lower KID means higher quality of synthesis.

Finally, the LPIPS is designed as a perceptual metric. As the previous metrics, it uses the intermediate activations of a secondary network, usually pretrained on a classification task (AlexNet, VGG or SqueezeNet). It exploits the deep activations of the secondary network, normalizes along the channel dimension, and then computes a perceptual distance. However, LPIPS evaluates diversity, as well as perceptual quality, which makes it unsuitable for the application case of this thesis.

2.4.3 Image quality assessment in medical imaging

Medical imaging takes a great part in the field of medical diagnosis. A large part of the research aims at improving image quality to reduce the acquisition time, for ionizing radiation reduction purposes or reduction of long-time acquisition. This quality improvement is done either by optimizing the acquisition part on a hardware point of view, or by designing reconstruction methods to recover a similar quality as an original image with a lower acquisition time. However, standard practice rarely includes reference images to perform a comparison, restricting the use of FR-IQA metrics. Moreover, medical imaging brings its own specificity compared to natural image quality assessment. The most noticeable are the importance of anatomical features conservation, the specific kind of artifacts of medical imaging (such as ghosting, movement artifacts, or some modality-specific artifacts described in the MRI case in section 1.3.4).

The quality of synthesized images is a critical aspect in medical image synthesis, as it may lead to misinterpretations and diagnostic errors. While image quality and diagnostic quality are not necessarily correlated, it is generally accepted that higher quality imagery is preferable.

Subjective IQA is a classic and relevant tool for medical image quality assessment. It includes diverse methods such as Double-Stimulus Continuous-Quality Scale or Difference Mean Opinion Score (DMOS) [150]. However, these evaluation methods are time-consuming, biased and possibly dependent on the viewing context of the evaluator. On the other side of objective IQA, FR-IQA and RR-IQA metrics are most of the time unusable in real-world applications because of the absence of ground truth image. In practice, these methods are used to assess the quality of artificially distorted images, while NR-IQA

metrics appear to be ideal for medical image quality assessment in daily practice.

Some new NR-IQA metrics have been introduced especially for medical imaging. According to the literature, these metrics aim at producing results that are closest to subjective IQA conclusions [160]. A great part of them are learning-based and/or specific to a given medical imaging modality. Since this work is focused on MR imaging, we shall only consider those MRI-compatible. If some of them have been developed to catch a specific type of artifacts [161], [162], some others are designed to assess the overall image quality. Those have been listed in medical image quality assessment reviews [150], [160] and provide the validation methods used to assess the quality of the listed metrics [163]. Some of these metrics are straightly adapted from existing ones designed for natural images [164], [165], while others are designed specifically for MR imaging like [166], [167], [168], [169].

In medical imaging, image quality evaluation typically relies on classical FR-IQA metrics such as MAE, MSE, PSNR, or SSIM, which are widely used in image synthesis literature. Objective IQA metrics are more commonly used than subjective IQA methods due to the time-consuming nature of the latter.

2.5 Conclusions

Deep learning models are a major component of state-of-the-art methods in image synthesis for computer vision and medical imaging. Their ability to model complex non-linear functions and their adaptable level of complexity make them suitable for various applications. The choice of a framework is highly influenced by the data characteristics such as the size and the pairing of the dataset. In order to synthesize high-resolution dynamic MRI sequences, we focus on image synthesis methods, in particular image-to-image translation methods.

Improving dynamic MRI quality through super-resolution or reconstruction is a specific area of research in medical image-to-image translation. The contribution of deep learning in the proposed methods is relatively lower compared to its significance in other medical image-to-image translation applications. This could be attributed to the lack of publicly available datasets in this particular modality. However, some deep learning methods emerge by using techniques to compensate for the lack of dynamic datasets. For instance, some methods simulate dynamic processes on publicly available datasets, such as the CHAOS dataset [170], by implementing non-rigid deformations to simulate specific physiological biomechanisms, such as breathing [171]. The methods proposed

for improving dynamic MRI quality can be divided into two categories: those that use registration-based frameworks to register a higher-quality MRI on the dynamic one [37], and those that aim to estimate the transformation from a dynamic image to a higher quality one. However, the last type of methods often relies on the use of paired datasets [171], [172], [173], [174]. Additionally, to the best of our knowledge, there is a lack of methods that explicitly model the dynamic process to enforce the resolution.

The dataset collected within the Equinus project is intrinsically unpaired. However, through a registration process, the static and dynamic MRI images can be partially aligned. This work focuses on two different approaches. The first approach leverages registration and simulation to generate partially paired and paired data for use in paired image synthesis methods. The second approach is based on unpaired image-to-image synthesis methods that do not require exact ground truth. The next chapter explores the paired frameworks for high-resolution dynamic MRI sequence and the dynamic MRI image simulation process from static MR image.

Chapter 3

Paired synthesis of high-resolution dynamic MRI

Abstract

The objective of this work is to estimate a high-resolution dynamic MRI sequence from a low-resolution dynamic MRI sequence. This chapter presents an investigation into the potential of deep learning methods for paired image synthesis. The chapter begins by describing two different data pairing processes that were investigated: one based on registration and the other on data simulation. Based on these datasets, we investigate several architectures for the paired synthesis of dynamic MRI sequences using the static MR images as ground truth to train the models.

3.1 Introduction

Paired image-to-image translation is an image synthesis technique that relies on the use of aligned image pairs from the source and target domains to address the translation between these two domains. This data pairing can be provided by the acquisition protocol, for instance, a single camera acquiring an image of a given scene with two different acquisition systems. Alternatively, it can be set with post-processing. The data used in this study make it impossible to have initially paired data.

The objective of this work is to estimate the transformation for synthesizing high-resolution dynamic MRI data from low-resolution dynamic MRI data and a high-resolution static MR. This chapter examines the potential of paired I2I translation methods for estimating this transformation. The main contributions are listed below:

— We propose a simulation process of dynamic MR data from static MR data.
— We demonstrate the inefficiency of paired image synthesis methods for high-resolution dynamic MRI synthesis.

This first section provides an overview of inverse problems, image registration and data simulation.

3.1.1 Inverse problem

The objective of solving an inverse problem is to identify the causes of observed effects. This is in contrast to a forward problem, which seeks to infer effects from causes. Forward models are generally much more investigated than inverse problems, because they explore the fundamental aspects of a system, such as defining the laws and properties of a system. Inverse problem solving is common in imaging (from astronomy to biomedical) and signal processing (seismic, electroencephalographic (EEG), geophysical signals). For instance, the localization of active neuronal sources in EEG from signals collected by scalp electrodes is an inverse problem [175]. Conversely, the prediction of the future state of a physical system from its current state is a direct problem.

Definition The mathematical formulation of the problem is expressed by the following equation:

$$y = H(x) \tag{3.1}$$

where y represents the observations, x the state to be estimated, and H the forward model. A problem can be described as either *well-posed* or *ill-posed*. The concept was introduced by Jacques Hadamard in 1923. A problem is considered well-posed if it satisfies the following three conditions:

— Existence: There exists at least one solution x to the problem $y = H(x)$.
— Uniqueness: There is exactly one solution x to the problem $y = H(x)$.
— Stability: The solution is continuously dependent on the observations y.

If at least one of the aforementioned conditions is not met, the problem is deemed ill-posed.

In practice, however, H is only a theoretical model and does not take into account noise and measurement errors due to instrumental constraints. The inverse problem can then be expressed as $y = H(x) + \epsilon$, where ϵ represents an additive noise that models the measurement bias.

Resolution Inverse problems can be broadly classified into two main categories: linear and nonlinear [176]. In the case of a linear problem, assuming a discrete configuration, the aforementioned relationship can be written as $y = Hx + \epsilon$, where H is a possibly non-square matrix. Theoretically, solving the inverse problem necessitates inverting the matrix H. If the matrix H is square and regular, then the solution exists and is unique. Otherwise, the problem is ill-posed, as there is either no solution or an infinite number of solutions. However, these systems can be solved by matrix factorization, the introduction of a regularization term (by restricting the space of possible solutions, the addition of a regularization term transforms an ill-posed problem into a solvable one), or the use of probabilistic methods.

If the problem is nonlinear, the inverse problem can be written as $y = H(x) + \epsilon$. The solution to the problem is then determined by solving an optimization problem $\text{argmin}_x U(x, y)$, where U corresponds to a data fidelity term. U quantifies the difference between the unknown state and the observations, which are assumed to be known. For instance, U can be defined as either a mean squared error or a Euclidean norm. However, in the case of an infinite number of solutions, this formulation is insufficient for solving the problem due to the frequent multiplicity of local minima. In most cases, a regularization term R is incorporated into the formulation. This term typically assumes the form of an a priori, providing a comprehensive framework for the solution.

Inverse problem in imaging Let X be the set of images and Y the set of measurements. Y depends on the application, and its structure may vary depending on the measurement device. H can be linear or nonlinear and known, partially known, or unknown. The most common applications in computer vision include inpainting, image interpolation, reconstruction, deblurring, and super-resolution. In medical imaging, it covers reconstruction of CT image from sinogram, reconstruction from partial measurements in CT and MRI, and denoising. The inverse problem in imaging can be solved by finding the transformation H^{-1} that connects an element of Y to an element of X. The solution can be elaborated from an analytical model, as in compressed sensing [177], but although these solutions have a stronger mathematical and physical background, they require a complete understanding of the underlying physics (which is not systematically possible), and are usually very specific.

In recent years, numerous studies have demonstrated the effectiveness of deep neural networks for inferring the underlying physical model from large imaging datasets [178]. In

applications where both X and Y are image spaces, such as denoising or super-resolution, these approaches often rely on paired (x, y) images. However, should the images be unpaired, post-processing techniques may be employed to achieve data pairing. The most prevalent methodologies employed in imaging are registration and data simulation.

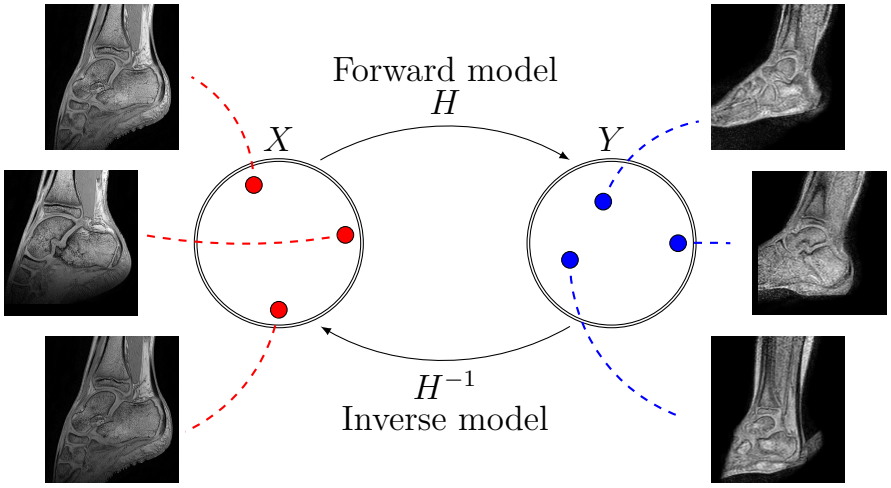


Figure 3.1 – Inverse problem in the Equinus dataset. Considering the formulation $y = H(x) + \epsilon$ where H is the forward model, x corresponds to the static MR image, i.e. the real state and y corresponds to the dynamic MR images, i.e. the observations.

3.1.2 Registration

The objective of registration is to determine the geometric transformation that will align the structures between the images to be registered. The primary applications of registration in medical imaging include multi-modal fusion (intra-patient registration) and atlas construction (inter-patient registration). The registration process is constrained by at least three parameters that condition the final result

— Similarity criterion: it may be regarded as the cost function in the optimization problem. The objective is to identify the global maximum of this function, which represents the optimal alignment between images. The similarity criterion is selected according to the data, i.e., the correlation ratio or mutual information, as examples.

— Optimization strategy: it defines the method used to identify the optimal transformation, given a specified cost function.

— Transformation model: the transformation model delineates the range of permissible transformations, i.e., the degrees of freedom (dof). In three-dimensional space, a rigid transformation is defined by six dof. Three translations and three rotations were considered. A non-rigid transformation may have up to 12 degrees of freedom (dof), including three for shears and three for scale factors.

The output of the registration process is a deformation field that represents the estimated geometrical transformation between the images.

However, the accuracy of the registration process can be influenced by multiple factors, some of which may be related to the data or parameters chosen for the registration process. Among the primary sources of error in medical imaging applications are: image noise, repeated patterns, large transformations between the two images, and contrast or modality differences. The similarity criterion of the registration process is of critical importance, particularly in the case of different imaging contrasts or modalities. Furthermore, a large number of dof may result in slower and more difficult convergence.

In some cases, the outcome of the registration process does not result in correctly paired images. The objective of data simulation is to identify and reproduce the data generation process in order to artificially generate paired data.

Registration validation The quality of the registration process can be evaluated in either a qualitative or quantitative manner. Qualitative assessment is inherently subjective and time-consuming, rendering it unsuitable for the evaluation of large amounts of data. Quantitative methods can be applied automatically and offer an objective evaluation of registration. The following paragraphs describe both quantitative and qualitative validation in medical imaging.

Qualitative validation is a subjective evaluation of the registration result conducted by a human observer, who is referred to as an expert. It is a common practice in medical imaging to assess the clinical quality of images using this method. In most cases, the evaluation is based on a visual assessment of the expert, who determines whether the registration result is accurate and rejects those that are not. The accuracy of the registration is typically evaluated using images and structures overlays, which highlight any misregistrations.

Although this process may include a clinical expertise, it is subject to a high degree of subjectivity and is highly time-consuming, making quantitative approaches preferable for validating registration results.

Quantitative evaluation can rely on ground truth or a gold standard. A ground truth provides the exact result of registration, while a gold standard offers an approximation of the correct result.

In practice, there are seldom applications where ground truth is available for registration validation. Synthetic data obtained by applying artificial transformations or simulated motions to images constitute the majority of these applications. The known perfectly registered target image allows for the computation of exact registration errors at each image position. The most common metrics include SSIM, MSE, and PSNR (see Section 2.4). However, these approaches are often a simplified version of the registration problem and suffer from a lack of realism.

In real-world applications, since a ground truth is (most of the time) unavailable for validating registration results, validation methods rely on a gold standard. A significant proportion of these applications involves multimodal registration (i.e., registration between images from different modalities, such as MRI and CT) or monomodal multicontrast registration (i.e., registration between images from the same modality but with different contrasts, such as T1 and T2 MRI). Consequently, in contrast to ground truth approaches, the target image does not correspond to the desired result of the registration process; rather, it serves as a guide with which the anatomical features of the source image should be matched. Consequently, the validation of the registration process should be conducted using metrics that can accommodate different contrasts or modalities.

Some metrics have been developed to compute the similarity between two images in a multimodal or multicontrast setup. The most common metrics include the sum-of-square distance (SSD), the mean square distance (MSD), and the (normalized) cross-correlation (NCC). These metrics are effective in measuring the accuracy of image registration when both images exhibit a similar intensity distribution, such as in the case of magnetic resonance imaging (MRI) and computed tomography (CT). In the case of disparate image intensity distributions, the mutual information (MI) metric, its variants, and learning-based metrics are more suitable for use. Nevertheless, these metrics consider the overall similarity of an image, yet fail to accurately capture registration error in specific structures.

The alignment between the anatomical structures of registered and target images offers an accurate measure of the registration error on these particular structures. This method is the most commonly used gold standard approach. The structures of interest are typically delineated by segmentation, and the alignment can be quantified by the degree of overlap

between the registered and target structures. This is commonly expressed by the Dice Similarity Coefficient (DSC) and the Tanimoto coefficient. Given two finite sets, A and B, the DSC and the Tanimoto coefficient can be expressed by the following equations [179]:

$$\text{DSC} = \frac{2\,Card(A \cap B)}{Card(A) + Card(B)} \tag{3.2}$$

$$\text{Tanimoto} = \frac{Card(A \cap B)}{Card(A \cup B)} \tag{3.3}$$

where the $Card$ operator denotes the cardinality of a set. In the computation of segmentations overlap, the sets of pixels or voxels of an anatomical structure on the registered and the target images are denoted by A and B, respectively. This method requires corresponding structures to be segmented on both the source and target images. A particular focus should be placed on the quality of the segmentation, as any errors in the segmentation can propagate to errors in the registration. This method is particularly accurate for the quantification of registration quality with respect to small and local structures, as opposed to large ones.

A more precise alternative employs corresponding anatomical keypoints in place of structure segmentations. This method circumvents the "coarse" estimation of the overlap and the propagation of potential segmentation errors by computing the distance between registered and target corresponding anatomical keypoints. Despite the more precise computation of the registration error, this method suffers from significant drawbacks. The method is highly dependent on the accuracy of the keypoints. While segmentations allow for some small errors regarding the size of the structure, keypoints must be positioned with high precision to compute accurate registration errors. Additionally, the method is time-consuming, depending on the registration type. It requires a greater number of keypoints to assess the quality of a non-rigid registration process than a rigid one.

3.1.3 Data simulation

The field of data simulation represents a hot area of research in the domain of medical imaging. Its initial purpose is to serve as a substitute for imaging modalities when they are unavailable or difficult to obtain due to constraints on time, personnel, or the high costs associated with multiple imaging modalities. Furthermore, they can reduce patient exposure to ionizing radiation, which, at high doses, can cause cancer, lesions, and organ and tissue dysfunction (synthesis of CT scan images from MRI images). Finally, they

provide an alternative when images from different modalities are not precisely registered. The primary applications of data simulation encompass inter-modality synthesis (between CT, MR, or PET) and intra-modality synthesis (differences in resolution or protocols).

In accordance with the inverse problem formulation introduced in Section 3.1.1: $y \in Y$ represents the dynamic MR images, $x \in X$ represents the real state symbolized by the static MR images, and H represents the degradation model between x and y. It is also assumed that the data are acquired in such a way that a dynamic MR image y has no corresponding static MR image x and vice versa. Finally, it is assumed that the data are affected by additive noise, represented by the variable ϵ. The primary objective is to solve the inverse problem, which entails obtaining a static MR image, x, from a dynamic MR image, y. In practice, the forward model, H, is unknown, and there is no pair of images (x, y) that satisfies the equation $y = H(x) + \epsilon$. While the inverse problem is often ill-posed (i.e., multiple y values can result in the same x value), the direct problem is assumed to be well-posed (i.e., the transformation of y by H^{-1} will always yield the same x value).

The purpose of the simulation process is to estimate the forward model for generating observations y' from the different x values. Consequently, the inverse problem is to be solved by estimating the inverse model H^{-1} that satisfies the equation $H^{-1}(y') = x$. The use of data simulation for inverse problem solving is found in a wide variety of applications. This simulation is most often based on manual model design or learning. Regression networks are already employed for the purpose of learning forward models [180].

In the application case, paired data can be obtained via two distinct methods: physical implementation (new acquisition to generate paired data) or estimation of the forward model [178]. By excluding the acquisition of new data, data simulation offers an opportunity to generate paired data for inverse model learning. The quality of inverse model estimation is contingent upon the accurate design of the forward model.

3.2 Data pairing

This section details the pairing process between the static and dynamic MR data from the Equinus project. The objective is to create a set of paired data for high-resolution dynamic MRI synthesis learning. Three approaches are explored: a registration-based process, a learning-based approach to simulate dynamic MR data from static MR data, and a manual simulation process of dynamic MR data from static MR data.

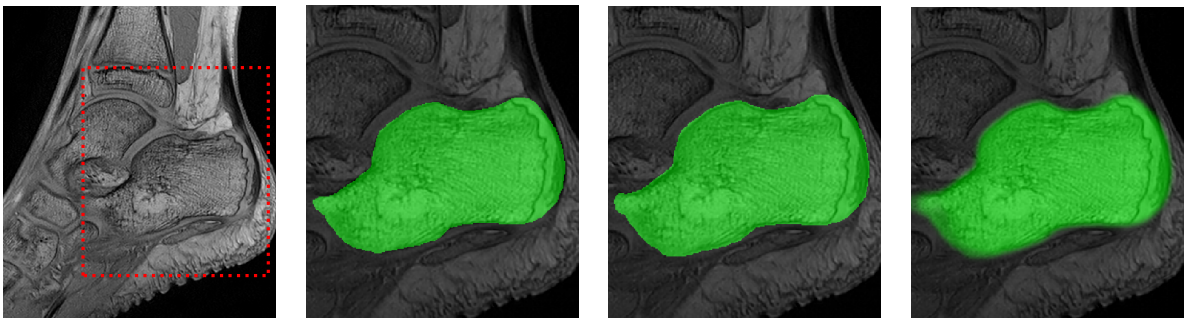
3.2.1 Registration

Global registration The aim is to register each static MR image on each dynamic volume. The initial step is a global rigid registration with six degrees of freedom. This process aims to move the static MR image into the dynamic coordinate system and provide a preliminary estimation of the transformation. The registration is performed using FLIRT, a tool for linear affine registration that is intermodality compatible. The optimization process is described in detail in the source papers [181] [182]. The correlation ratio is employed as a similarity criterion.

Following this initial step, a preliminary filtering stage is introduced to exclude images for which the estimated transformation is unrealistic. To this end, the cross-correlation is computed between each pair of images: the reference dynamic volume and the registered static volume. A threshold, which indicates whether a registration is accepted or not, is defined by exploiting the relationship between the DSC and the cross-correlation. This relationship is demonstrated by using the few bone segmentations available for the dynamic MR images, allowing the DSC computation between the static and dynamic segmentations. The threshold is first arbitrarily defined to 0.85 for the DSC, and then adapted to the cross-correlation, and set to 0.64. After this initial filtering stage, 55 of the 573 registered images were rejected, which represents approximately 9.6% of the registered images.

Registration bone to bone Following the initial registration step, each static MR image is roughly registered to each dynamic volume of the same subject. During the dorsiplantar flexion movement performed on the dynamic sequences, non-rigid deformations are introduced between the bones and in the surrounding tissues. The different positions taken by the foot during movement make it impossible to achieve data pairing by rigidly registering the static MR image on each dynamic data. The use of non-rigid registration to achieve a complete alignment between static and dynamic data requires long computation times and is prone to significant errors. These errors are favored by the noise between the two images, the complexity of the computation, and the fineness of the anatomical structures to be registered. To avoid the anatomical errors introduced by non-rigid registration, the registration problem is simplified by considering a rigid registration per bone, and therefore only partial pairing of the data.

The subsequent stage is thus to estimate a finest bone-to-bone rigid registration. This approach provides a more accurate registration in the bone region and allows for the use of



(a) (b) (c) (d)

Figure 3.2 – Bones segmentation processing for registration. (a) Original static MR image. (b) Segmentation of the calcaneus. (c) Segmentation of the calcaneus after a morphological dilation. (d) Segmentation of the calcaneus after a morphological dilation and a blurring by a Gaussian filter.

rigid transformation in the three-dimensional space, since the bones are not deformable. This rigid configuration offers a reduced computational cost and execution time in comparison to a non-rigid one. Moreover, a rigid transformation ensures that the anatomical features remain unaltered, thereby eliminating any suspicion of unwanted distortions.

The previously registered static MR image is taken as input for this second registration step and successively registered on the three bones of interest, resulting in three additional registrations. The cost function (similarity criterion) is weighted using the static segmentation of the three bones of interest to force the registration to be consistent in the bone region. Given this setting, the segmentations are morphologically dilated by a spherical structuring element with a radius depending on the bone. The radius is empirically determined under the criterion of registration quality. It is set to 1 for the calcaneus and to 2 for the talus and the tibia. The dilated segmentations are then blurred by a Gaussian filter with a standard deviation of 2. This manipulation adapts the weighting map (i.e., the segmentation) to include the edges in the voxels used for the cost function computation (dilation) and to weight the cost function decreasingly for pixels outside of the bone.

3.2.2 Learning-based dynamic MRI simulation

The goal of dynamic data simulation is to generate paired static and dynamic MRI data, i.e. data that are perfectly aligned with each other. The learning-based approach

employs artificial neural networks to learn the transformation for estimating a dynamic MR image from a static MR image. These approaches demonstrate excellent performance in the learning of non-linear functions [183]. However, they frequently encounter limitations in terms of physical explicability and reproducibility, as well as inductive bias.

The objective is to estimate the transformation from a static MR image to a dynamic one. In the context of our application, the data available for training the model consists of unpaired static and dynamic MR images. In other words, there is no dynamic MR image that corresponds to the application of H on any static MR image. Under this limitation, two distinct methodological approaches are investigated. The first method exploits the registered Equinus data, forming a partially paired dataset, allowing the use of standard methods for paired I2I translation. The second method is designed to handle unpaired data and approximate the transformation from static MR images to dynamic MR images by approximating the data distributions of the target set.

Partially paired data We first address the problem of simulating dynamic MRI data using registered images to construct a paired dataset.

However, these data are only partially registered since the registration process has been chosen to be rigid and bone specific. Consequently, the resulting data comprises a set of dynamic MR images and the corresponding static MR images registered on each ankle joint bone. As a result, we dispose of pairs (x, y) verifying the Equation 3.1 only in the registered bone section of the image. To address this partial pairing constraint, bone segmentations are used to weight the classical MSE to zero outside the bone region where the images are registered (see Equation 3.4).

$$\text{MSE} = \frac{1}{N}\sum_{i}^{N}(y_i * m_i - \hat{y}_i * m_i)^2 \tag{3.4}$$

The hypothesis is that the transformation learned between static and dynamic MR images solely on the paired bones can be extrapolated to the entire image at the test time, thereby enabling the generation of complete translated images while learning only on parts.

Unpaired data We then propose to address the problem of simulating dynamic MRI data by overcoming the need to use partially paired data. To this end, we propose to use a GAN architecture [45]. The principle of this model is described in detail in Section 2.1.4.

In this model, the generator is fed with static MR images. Subsequently, the outputs of the generator are compared to the original dynamic MR images through the discriminator, which is trained to distinguish between the original and simulated dynamic MR images.

3.2.3 Handcrafted dynamic MRI simulation

In addition to learning-based methods for dynamic MR image simulation, we propose the manual design of a dynamic MR image simulation model using mathematical and imaging functions. The manual design of a data simulation model typically necessitates an understanding of the underlying physical processes. In the absence of precise knowledge of the data construction process, the selection of the various model components is made in order to approximate the real data as closely as possible, although it may not faithfully reproduce the underlying physical process. This approach provides a preliminary simulation model and offers greater transparency over the simulation process than learning approaches.

The manual design of a data simulation model employs mathematical functions to mimic the transition from a static to a dynamic MR image. The designed simulation process of a dynamic MR image from a static one is illustrated in Figure 3.3.

The simulation model is divided into two main stages: a data augmentation phase, followed by the degradation model itself.

The purpose of the data augmentation phase is to virtually increase the amount of data available for each subject. As a reminder, the Equinus dataset includes a single static MR image for each subject. We propose to increase this number by applying random transformations to the source image. The transformations are chosen to reflect what is observable in the data. To mimic variations in foot position and size, we propose the application of random affine transformations. The permitted transformations include rotations (in the sagittal plane with a range of $\pm 30°$) and scaling effects (according to a ratio of ± 0.1). To simulate anatomical differences between patients, we propose to introduce random elastic deformation.

Secondly, the degradation model used to simulate dynamic MR images from a static MR image is applied to augmented images. This model is comprised of a succession of filters associated with additive Gaussian noise. With this degradation module, the aim is to simulate the texture of dynamic MR images, as well as the loss of detail compared to static images. The degradation operator consists of the successive application of gray-scale erosion by a structuring element of size 1×1, followed by a median filter (disk

structuring element of radius $r = 2$), the addition of Gaussian noise, an undersampling with a sampling factor of 2, a Gaussian filter and an oversampling by a factor of 2 to return to the initial resolution. In order to limit the complexity of the interpolation algorithm as much as possible and thus increase the possibility of introducing artifacts, over- and under-sampling interpolations are set to "nearest neighbor". The additive noise is generated from a Gaussian distribution with a standard deviation of $\sigma = 0.4$. This noise is then processed in parallel on two different pipelines. The first method employs a series of grayscale erosion operations with structuring elements of increasing size, while the second method employs a series of grayscale dilation operations with the same structuring elements. Finally, for each pipeline, a Gaussian filter with a standard deviation for the gaussian kernel of 0.8 is applied. This noise is carefully constructed in order to replicate the texture observed on dynamic MR images. Finally, ghosting artifacts are added, as observed in the dynamic data.

3.2.4 Experimental setup

Dataset The experimentation was conducted on images from the Equinus dataset. The learning-based methods for dynamic MRI simulation using partially paired data leverage the data resulting from the registration process described in Section 3.2.1.

Implementation details The training set for learning-based methods consists of 11 subjects, five with equinus and six typically developing. The validation and test sets each include one subject, respectively typically developing and with equinus. For a fair comparison, images from both simulated and real datasets are sampled to a resolution of $0.41 \times 0.41 \times 8$mm. The resolution in the sagittal plane was set to be halfway between the static and dynamic resolutions in order to limit information loss in static MR images and interpolation artifacts in dynamic MR images. Finally, in order to limit the proportion of interpolated images in the training set and given the significant differences in resolution between static and dynamic MR images, the resolution along the orthogonal axis to the sagittal plane was kept consistent with that of the dynamic MR images.

As previously described in Section 3.2, the real dataset includes static and dynamic MR images registered bone-to-bone. For the simulated dataset, we set the number of augmented static images to 20 per subject. We set the probability of random affine transformation to 0.8 and the probability for an elastic deformation to 0.2.

Throughout this work, we used 2D patches from MRI sequences for training. The

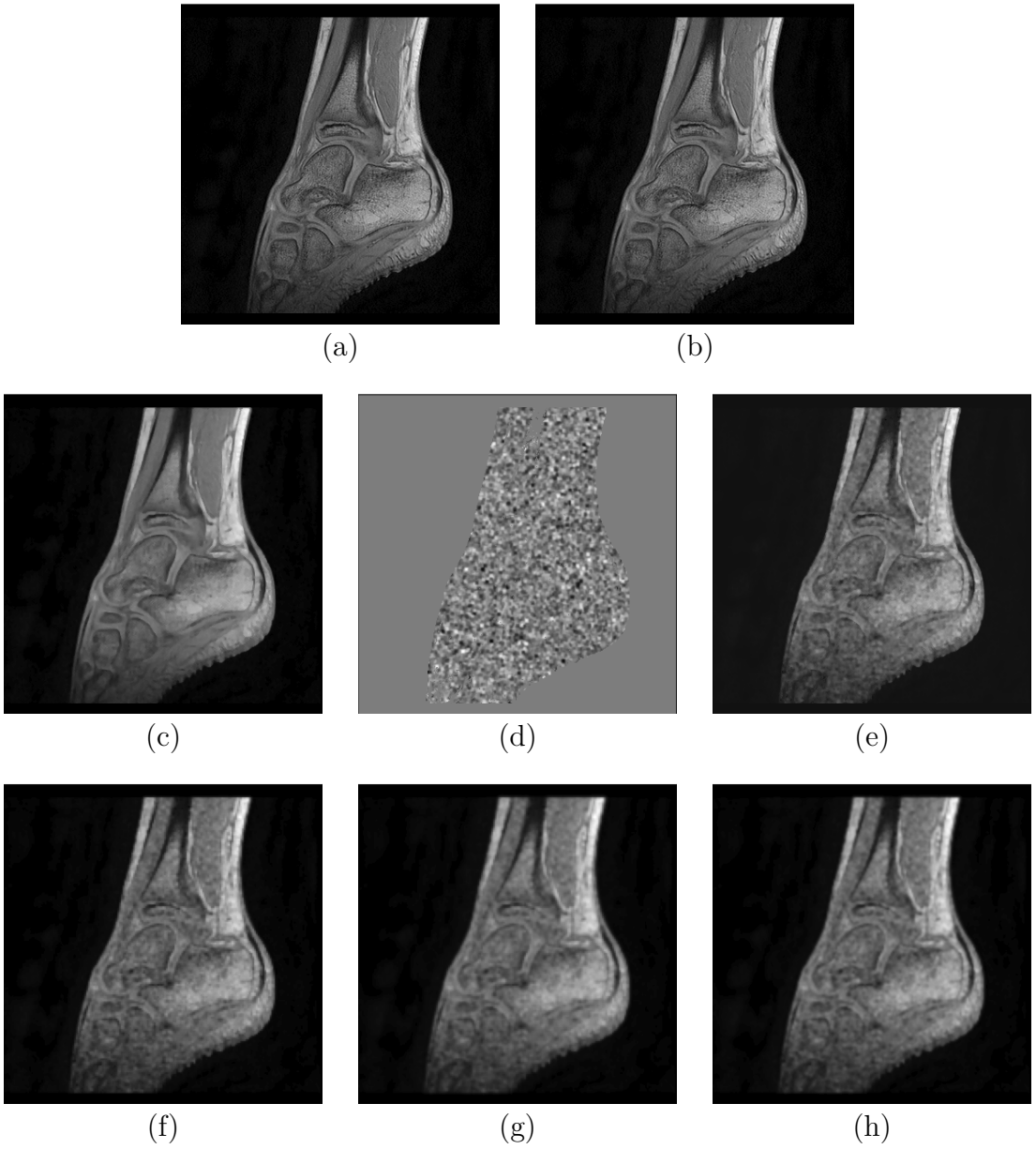


Figure 3.3 – Handcrafted dynamic MRI simulation process. Each successive step is illustrated. (a) Original static MR image (b) First grayscale erosion (c) Median filtering (d) Additive Gaussian noise (e) Addition o the Gaussian noise (f) Undersampling with a sampling factor of 2 (g) Gaussian filter (h) Oversampling by a factor of 2.

use of 3D patches increases computation times, thereby impeding the training process. Furthermore, the strong anisotropy in dynamic MRI images between the resolution in the sagittal plane and along the frontal axis limits the accuracy and robustness of volumetric image processing. For all experiments, the patch size is set to 64×64. The number of training patches is augmented to 98,000 through data augmentation using TorchIO [184], an open-source library designed for medical imaging. TorchIO enables the generation of MRI-specific artifacts, such as magnetic field inhomogeneity and motion artifacts, as well as typical computer vision augmentations. The random transformations applied include flips along the lateral axis, bias field with a maximal magnitude of polynomial coefficient equal to 0.5, Gaussian noise with a standard deviation of 0.1, and affine transformations including scaling (with an amplitude of 0.2) and rotations (along the sagittal plane, with a 40° amplitude).

All the models have been implemented using PyTorch. Two classical architectures in paired image synthesis are used for dynamic MRI simulation: UNet [185] and ResNet [186], both fully-convolutional. To assess the influence of architectural hyperparameters on synthesis quality, three network sizes are considered for each architecture. The network sizes are obtained by adjusting the number of layers in each network. The unpaired method employs a fully convolutional classifier architecture as a discriminator. In this study, the same discriminator architecture is used throughout the experiments. The generator architecture is identical to that employed in the partially-paired methods.

For training, we use Adam optimizer, with a learning rate of 10^{-5}. The batch size is set to 32, exponential decay rates to $(\beta_1, \beta_2) = (0.5, 0.999)$ and weight decay to 0.0001.

Registration validation A paired framework is dependent upon an accurate correspondence between two data. In this case study, the quality of the registration process is a crucial factor in ensuring the performance of image-to-image translation for the synthesis of high-resolution dynamic MRI.

The Equinus dataset provides tibia, talus, and calcaneus segmentations on the static MR images. In the dynamic MR images, however, only a few dynamic sequences have been given the same bone segmentations. Indeed, the manual segmentation of the bones in dynamic sequences is a time-consuming task and is prone to segmentation bias due to image noise or motion artifacts. Consequently, the quality of the registration cannot be assessed using overlap coefficient computation. Similarly, no anatomical keypoints have been set on either the static or dynamic MR images. As previously stated, the accuracy

of the keypoints placement is of critical importance in keypoints-based registration. However, the resolution of the dynamic 3D volumes is not sufficient to allow precise keypoints placement by experts.

In consideration of the available data, the registration validation process is divided into three distinct stages, progressing from a more general to a more specific analysis. The initial two stages of the process rely on quantitative metrics, while the final stage is a qualitative evaluation conducted by an expert.

Simulation validation The quality of the synthesis is evaluated through two classical metrics in no-reference image quality assessment: KID and FID. Both are detailed in Section 2.4.2. The KID reflects the shared visual similarities between two sets of data, whereas the FID is more oriented towards textures and edges.

3.2.5 Results

This section presents the results of the registration and simulation methods used to generate paired and partially paired data.

Registration The initial validation stage employs foot masks derived from both static and dynamic MR images. These masks are generated using the Otsu thresholding function [187], followed by a morphological closing. A global alignment of the feet is evaluated through a DSC computation between dynamic and registered static foot masks. This initial estimation of the registration accuracy allows for the elimination of aberrant registrations.

Without the application of bone segmentation, it is not possible to perform a fine evaluation of the overlap between structures. The second stage of the validation process employs the cross-correlation coefficient between the target and registered images. The cross-correlation coefficient is an effective similarity metric, even in a multimodal or multicontrast setup.

The final stage of the validation process is a qualitative evaluation conducted by an expert. Given the deformation field resulting in the alignment of the bones in the static MR image with those in the dynamic target image, the bone segmentations from static MR images can be registered accordingly. The registered segmentation are superposed to the dynamic MR image and submitted to the expert for evaluation. This allows the

expert to assess the quality of the registration and the degree of overlap between the structures.

The validation process yielded the following results: of the 1,554 pairs of static and dynamic MR images registered bone-to-bone, 582 were rejected and 960 were deemed satisfactory.

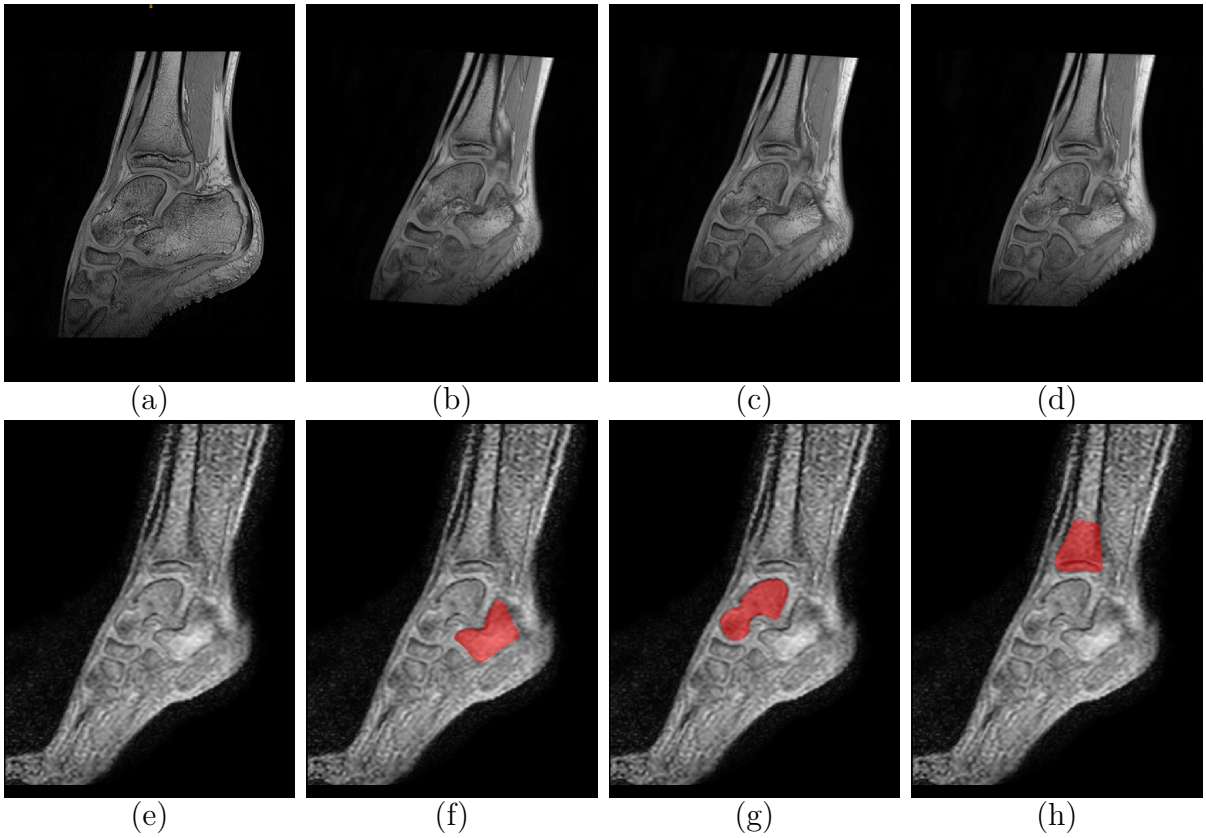


Figure 3.4 – Result of the bone-to-bone registration of a static MR image on a dynamic MR image. The registered static MR images are displayed along with the associated registered segmentations, displayed as an overlay on the dynamic MR image. (a) Original static MR image. (b) static MR image registered on the calcaneus. (c) static MR image registered on the talus. (d) static MR image registered on the tibia. (e) dynamic MR image. (f) Registered calcaneus segmentation. (g) Registered talus segmentation. (h) Registered tibia segmentation.

Learning-based simulation Figure 3.5 provides a visual assessment of the performance of the methods using partially paired data. It was observed that the method using the ResNet architecture did not generate images that exhibited realistic dynamic-like characteristics. Despite undergoing training until convergence, the ResNet architecture

	Parameters	KID ↓	FID ↓
UNet	8M	0.3	283.9
	1M	0.44	351.2
	400K	0.52	376.7
ResNet	8M	0.16	175.6
	3M	0.17	182.9
	1M	0.15	174.0

Table 3.1 – Evaluation of the quality of the synthesis of simulated dynamic MR images for different neural networks using partially paired data. The evaluation is performed using two unpaired image metrics, KID and FID.

failed to achieve a realistic level of degradation. Conversely, the UNet architecture results in a significant degradation of static images. However, the forward model thus constructed does not appear to be consistent with real data, nor does it approach the texture of real dynamic MR images. Table 3.1 provides a quantitative evaluation of each method. The results suggest that ResNet methods exhibit better performance, in contradiction to what is observed in Figure 3.5. This discrepancy may be attributed to the texture of the image, which is overly blurred in the UNet methods. Background artifacts may also degrade the quality score. In practice, none of the proposed methods for simulating dynamic MR data using partially paired data can provide a realistic estimate of degradation.

Figure 3.6 illustrates the outcomes of unpaired dynamic MRI simulation methods. As with partially paired methods, the use of a ResNet as a generator does not yield realistic results when compared with dynamic MR images. However, employing a UNet architecture as a generator allows for the modeling of a degradation that is more consistent with the data. Nevertheless, this approach does not lead to visually realistic data. Table 3.2 provides a quantitative evaluation of the methods. As previously observed, the metrics appear to be inconsistent with the perceived quality of the images in comparison to the original dynamic MR images. The models using a ResNet as a generator demonstrate superior results in this regard. These discrepancies between perceived quality and metric values can be attributed to the properties of the metrics employed. While FID is sensitive to textures and contours, it has been demonstrated to be robust to variations in brightness, contrast, and saturation [188]. The images generated by UNet, despite their visual realism, are affected by a range of artifacts, which significantly impact the contours and textures of structures. It can be postulated that the more favorable scores associated with results produced by models using a ResNet are a consequence of the absence of artifacts on the images generated, thus preserving anatomical structure and contours within the images.

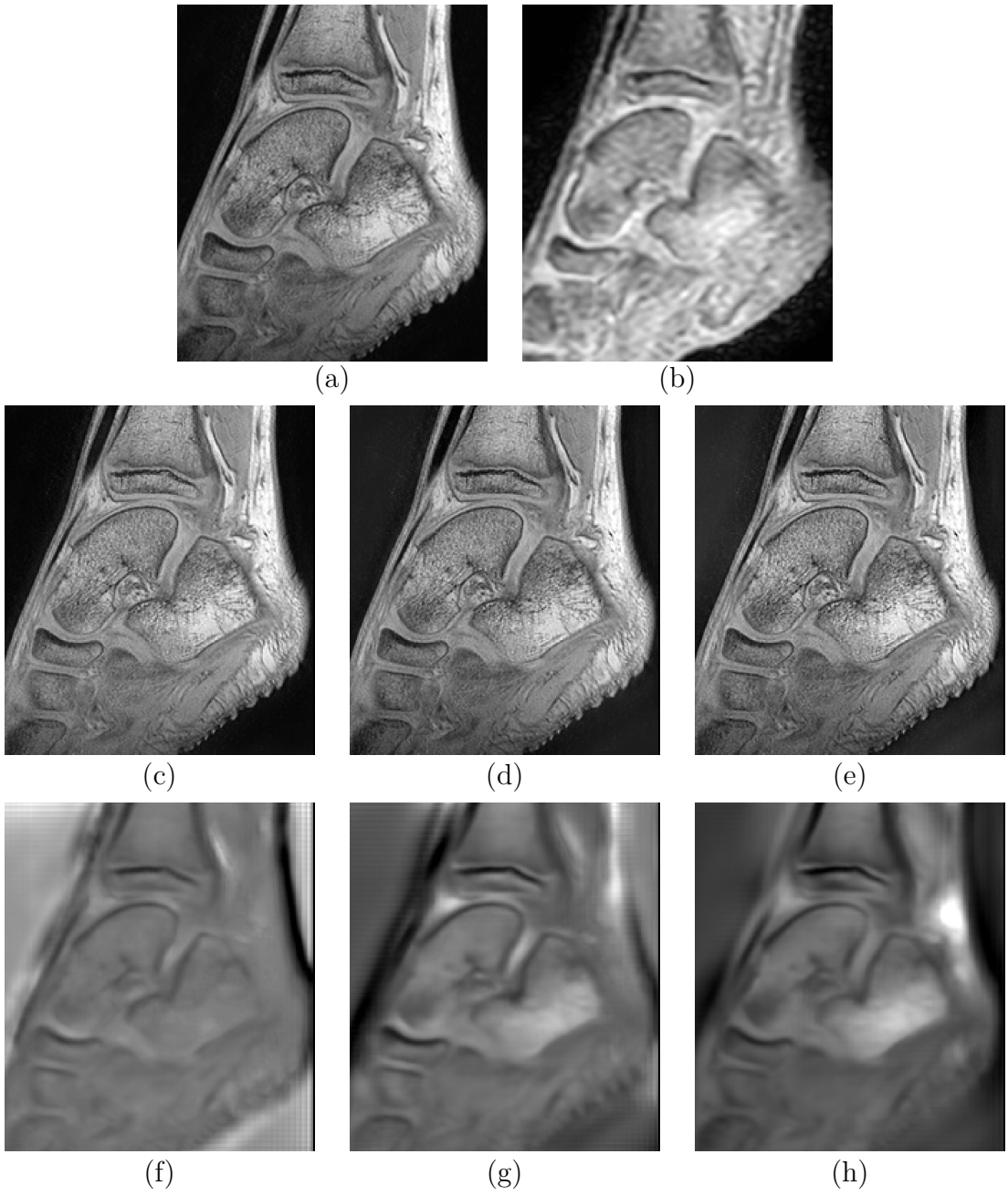


Figure 3.5 – Dynamic MR images simulation using a partially paired data. Two distinct neural network architectures are used: ResNet [186] and UNet [185]. For each network, three different sizes are studied. (a) Original static MR image (b) Corresponding registered dynamic MR image registered on the tibia (c) ResNet - 1M parameters (d) ResNet - 3M parameters (e) ResNet - 8M parameters (f) UNet - 400K parameters (g) UNet - 1M parameters (h) UNet - 8M parameters.

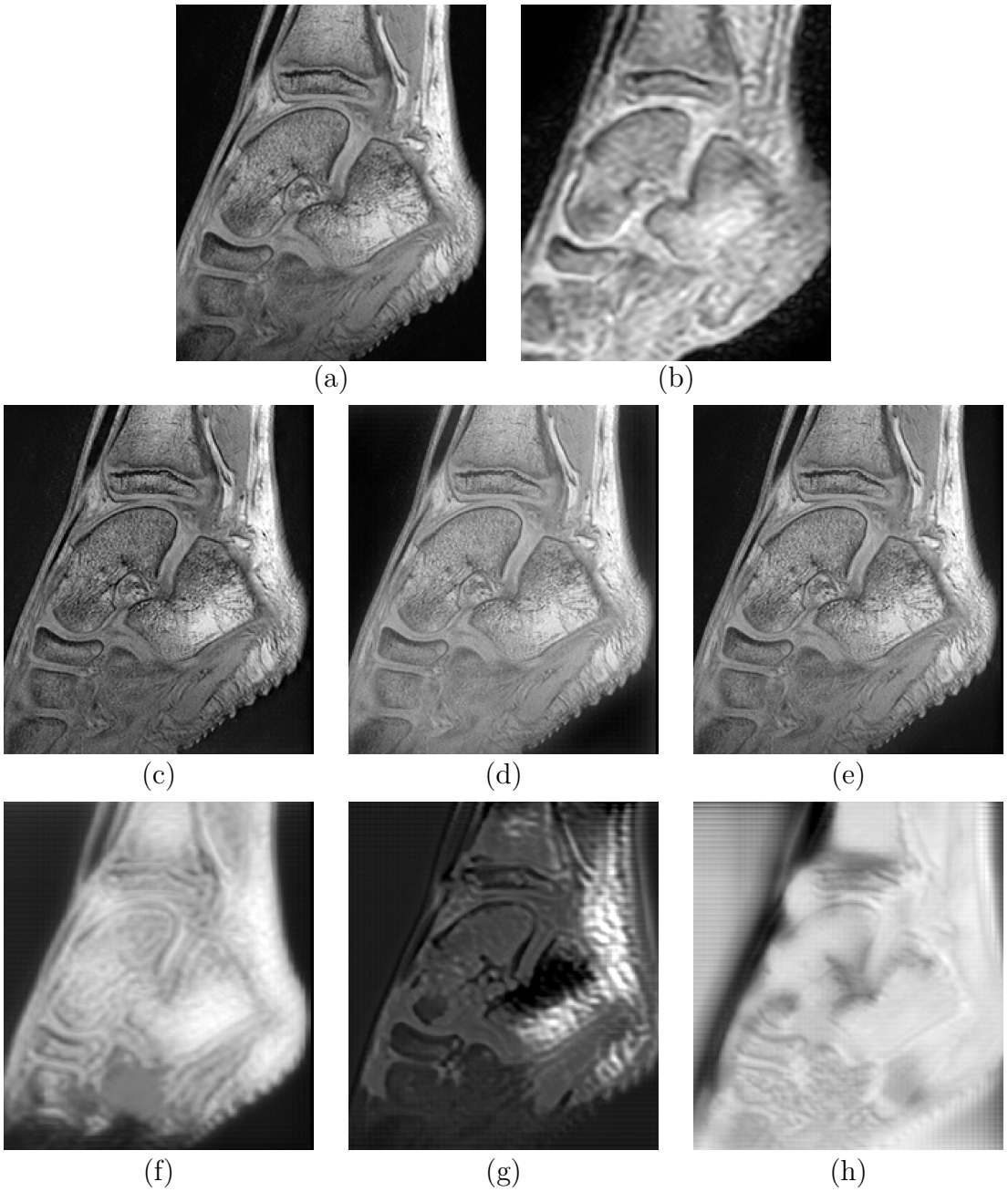
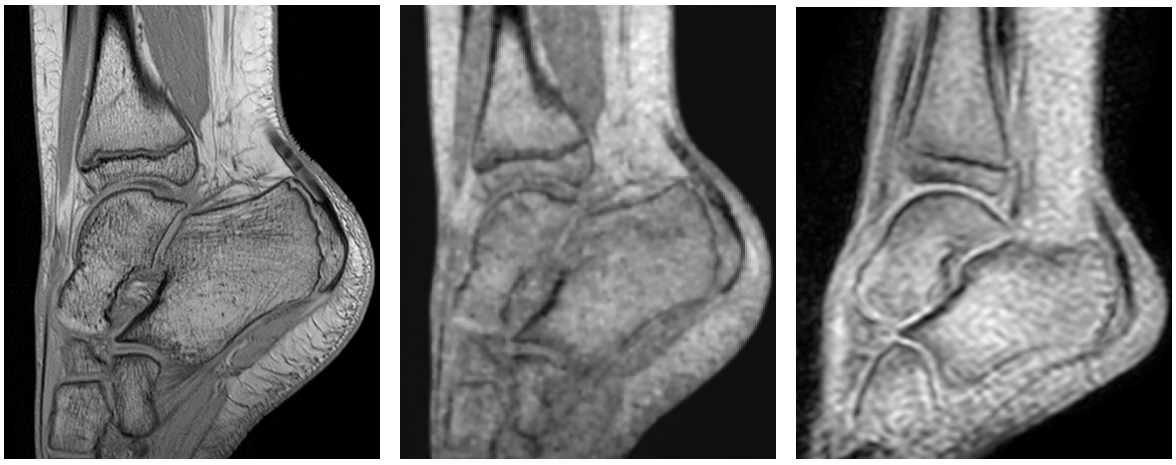


Figure 3.6 – Dynamic MR images simulation using unpaired data. Two distinct neural network architectures are used as GAN generator: ResNet [186] and UNet [185]. For each network, three different sizes are studied. (a) Original static MR image. (b) Corresponding registered dynamic MR image registered on the tibia. (c) ResNet - 1M parameters. (d) ResNet - 3M parameters. (e) ResNet - 8M parameters. (f) UNet - 400K parameters. (g) UNet - 1M parameters. (h) UNet - 8M parameters.

GAN Generator	Parameters	KID ↓	FID ↓
UNet	8M	0.55	392.1
	1M	0.18	192.52
	400K	0.22	223.75
ResNet	8M	0.19	194.04
	3M	0.19	193.7
	1M	0.18	186.6

Table 3.2 – Evaluation of the quality of the synthesis of simulated dynamic MR images for different GAN generators using unpaired data. The evaluation is performed using two unpaired image metrics, KID and FID.

Handcrafted simulation Figure 3.7 shows an example of a dynamic MR image simulated from a static MR image. The results appear to be more visually similar to real dynamic data than those produced by learning-based approaches.



(a) Original static MR image (b) Simulated dynamic MR image (c) dynamic MR image

Figure 3.7 – Dynamic MR images simulation using a handcrafted model.

3.2.6 Discussion

This section describes the process of pairing the static MR images with the dynamic MR images of the Equinus dataset. Two distinct approaches are presented in this section: a registration-based pairing and another based on a simulation process. The registration-based process resulted in a partial pairing of the static and dynamic MR images, limited to the bone region. The simulation process involves two learning-based approaches and

a handcrafted approach for estimating the forward model for simulating a dynamic MR image from a static MR image. The handcrafted approach is based on classical mathematical and imaging functions that have been selected to provide the most accurate approximation of dynamic MR images. Two learning-based approaches have been investigated: one using partially paired data and one using unpaired data. Neither approach successfully synthesized realistic simulated dynamic MR images, either qualitatively or quantitatively. The number of parameters in each neural network does not appear to have a significant influence on the quality of the synthesized images. The manual approach yielded the most realistic results of all the methods explored.

3.3 Paired high-resolution dynamic MRI synthesis

This section describes the methodologies used for the estimation of high-resolution dynamic MRI data using paired and partially paired data. This study is focused on learning-based methodologies and employs two datasets: the real dataset from the Equinus project and a simulated dataset described in Section 3.2.2. In order to assess the paired I2I task, the proposed methodologies are also evaluated on a classical paired medical imaging dataset: the HCP dataset.

3.3.1 Methods

Let Y be a set of low-resolution dynamic MRI data and let X be a set of high-resolution static MRI data. The objective is to estimate the inverse model that maps an element $y \in Y$ into an element of X. In this section, we propose to investigate the use of paired I2I translation methods to estimate this transformation. In the following section, we consider pairs of data (x, y), where y corresponds to the real or simulated input dynamic MR image and x to the static ground truth MR image.

The experiments are conducted on two distinct datasets: a real dataset with partially paired data and a simulated dataset with perfectly paired data. The experiments are conducted on a few common network architectures in I2I translation.

Paired data In the case of perfectly aligned data, as in the simulated dataset, each pixel of the predicted image can be compared to those of a target image, i.e., the ground

truth. The MSE standard loss can therefore be used as an objective function:

$$\text{MSE} = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2 \tag{3.5}$$

where $\hat{y}_i$ denotes the predicted images and y_i the ground truth images.

Partially paired data As previously stated in Section 3.2.2, the use of partially registered images to construct a paired dataset necessitates the adaptation of the MSE to exclude all image regions that are not part of the registered bone:

$$\text{MSE} = \frac{1}{N}\sum_{i}^{N}(y_i * m_i - \hat{y}_i * m_i)^2 \tag{3.6}$$

where m_i denotes the binary mask corresponding to the registered bone.

3.3.2 Experimental setup

Real dataset The real dataset is composed of pairs of static and dynamic MR images registered bone-to-bone, which process is described in Section 3.2.1.

Simulated Dataset A second set of experiments is conducted on a simulated dataset, composed of perfectly aligned image pairs. The conception of this dataset is described in detail in Section 3.2.2. Among the various methods investigated for simulating dynamic MR images, the handcrafted model was selected to generate the simulated dataset because of its greater degree of realism. The dataset comprises 18 subjects, 9 of whom with equinus and 9 with a typical development. Each subject has 20 pairs of dynamic and static MR images.

HCP dataset In order to assess the paired I2I task, a series of experiments are conducted on the HCP dataset. The HCP results from an effort to map the healthy human connectome using neuroimaging and behavioral data. The collected data are freely distributed. Each subject is provided with two paired MRI scans: one T1 and one T2. We use 400 subjects for training purposes, 4 for validation, and 45 for testing.

Implementation details For real data, the training set comprises eleven subjects, five of whom exhibit equinus and six of whom are typically developing. The validation and test

sets each include one subject, typically developing and with equinus, respectively. In the case of simulated data, 16 subjects are used for training purposes, with an equal number of subjects with equinus and those with typical development. One subject with typical development is used for validation, and one subject with equinus is used for the test. In the case of HCP data, the training set comprises 400 subjects, 4 additional subjects are used for validation, and 45 subjects are used for testing.

As previously described in Section 3.2, the real dataset includes static and dynamic MR images registered bone-to-bone, while the simulated dataset contains perfectly aligned static and simulated dynamic MR images. As in the dynamic MRI simulation model learning, and throughout all the experiments, we use 2D patches extracted from the MR data. The patch size is set to 64×64. For both the real and simulated datasets, the number of training patches is augmented to 98,000 and 96,000, respectively, through data augmentation using TorchIO [184]. Random transformations were applied to both the real and simulated data, including flips along the lateral axis, a bias field with a maximal magnitude of polynomial coefficient equal to 0.5, Gaussian noise with a standard deviation of 0.1, and affine transformations including scaling (with an amplitude of 0.2) and rotations (along the sagittal plane, with a 40° amplitude).

All models were implemented using PyTorch. Two classical architectures in paired image synthesis are employed for high-resolution dynamic MRI synthesis: UNet [185] and ResNet [186]. Three network sizes were evaluated for each architecture by adjusting the layers of each network.

For training, we used the Adam optimizer with a learning rate of 10^{-5}. The batch size is set to 32, the exponential decay rates are set to $(\beta_1, \beta_2) = (0.5, 0.999)$, and the weight decay is set to 0.0001.

Metrics Two distinct types of metrics are computed, contingent upon the dataset under consideration. In the case of the real dataset, classic no-reference image quality assessment metrics are employed, specifically, KID and FID (described in Section 2.4). For the simulated and HCP datasets, considering that the data are perfectly aligned, we rely on classical full-reference IQA metrics for the complete images: PSNR and SSIM. To provide a comparison with no-reference image quality metrics, we also compute KID and FID metrics on simulated data.

	Parameters	KID ↓	FID ↓
UNet	8M	0.41	343.5
	1M	0.37	314.4
	400K	0.45	362.4
ResNet	8M	0.16	185.1
	3M	0.17	192.1
	1M	0.15	177.6

Table 3.3 – Evaluation of the synthesis quality of high-resolution dynamic MR images on real data. The evaluation is performed using two unpaired image metrics, KID and FID.

3.3.3 Results

Real data Figure 3.8 presents the qualitative results of the synthesis of high-resolution dynamic MR images using partially paired data. It was observed that the ResNet architecture did not result in any improvement in the quality of the simulated dynamic MR image and produced results that were very similar to those of the source image. In contrast, the UNet architecture yielded results that were highly blurred and unrealistic. These results demonstrate that training on the paired portion of the image does not either permit the synthesis of a realistic texture on the bone or the generalization to the foot. Both methods were unable to synthesize realistic static-like MR images from a source dynamic one. Furthermore, the size of the network does not appear to have a significant impact on the results of the synthesis. Table 3.3 provides a quantitative evaluation of the methods. As previously observed in Section 3.2.5, the metrics appear to be inconsistent with the perceived quality of the images.

Let Y be the set of dynamic MRI data and X the set of static MRI data. The synthesis of low-resolution static MRI is equivalent to a transformation from X to Y, and the synthesis of high-resolution dynamic MRI is equivalent to a transformation from Y to X. Table 3.4 provides a comparison of the quantitative evaluations of both syntheses ($X \rightarrow Y$ and $Y \rightarrow X$). The results demonstrate that both transformations exhibit comparable performance in terms of their respective quantitative metrics, regardless of the method employed.

Simulated data Figure 3.9 presents the results of the synthesis of high-resolution dynamic MR images using the simulated dataset. As observed in the above paragraph, the ResNet architecture does not improve the quality of the input source image. In contrast, the UNet architecture demonstrated a significant improvement in the overall aspect of the

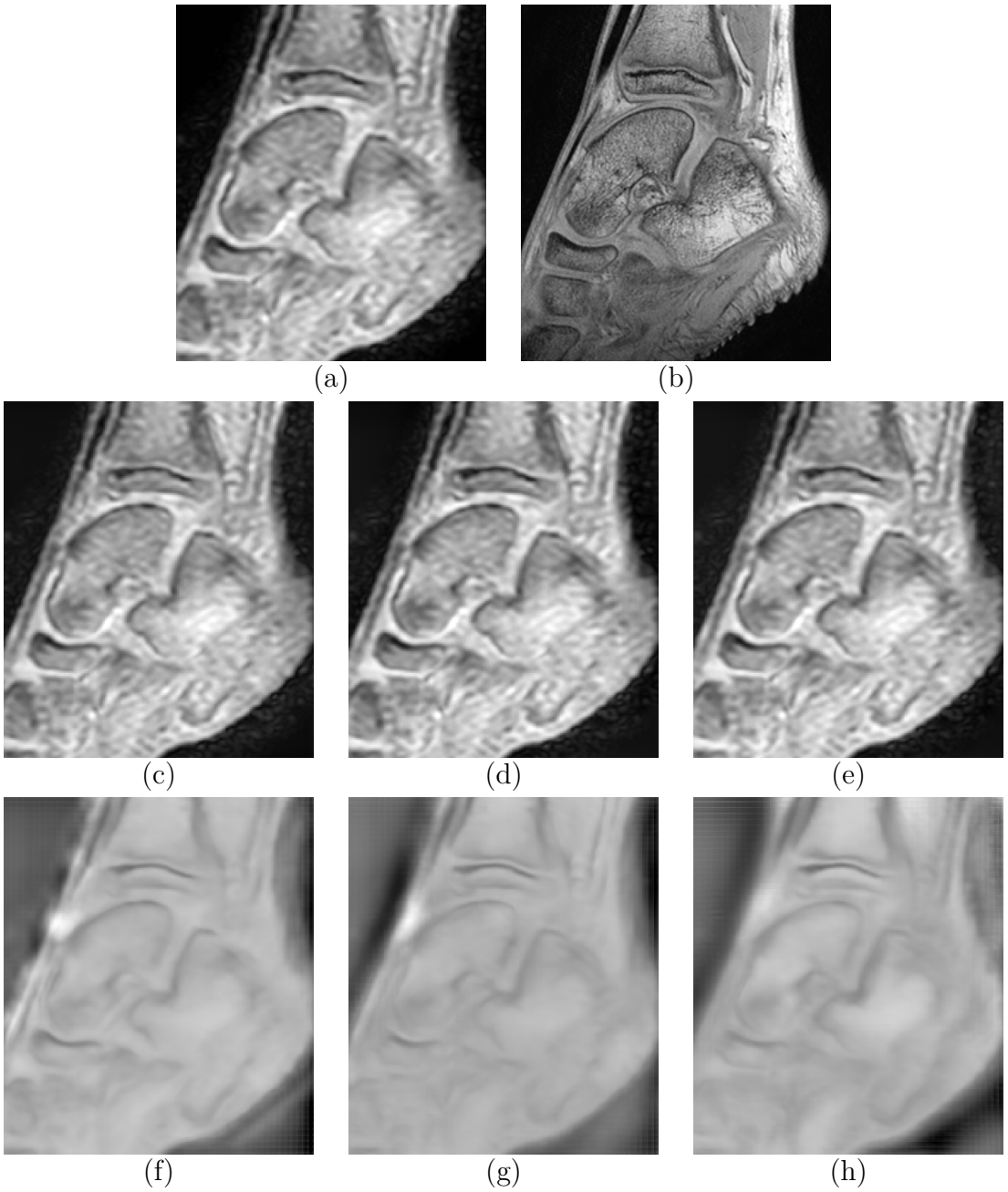


Figure 3.8 – Estimation of high-resolution MRI synthesis on real data using partially paired data. Two distinct neural network architectures are used: ResNet [186] and UNet [185]. For each network, three different sizes are studied. (a) Original dynamic MR image. (b) Corresponding registered static MR image registered on the tibia. (c) ResNet - 1M parameters. (d) ResNet - 3M parameters. (e) ResNet - 8M parameters. (f) UNet - 400K parameters. (g) UNet - 1M parameters. (h) UNet - 8M parameters.

		$\mathbf{X} \rightarrow \mathbf{Y}$		$\mathbf{Y} \rightarrow \mathbf{X}$	
	Parameters	KID ↓	FID ↓	KID ↓	FID ↓
UNet	8M	0.3	283.9	0.41	343.5
	1M	0.44	351.2	0.37	314.4
	400K	0.52	376.7	0.45	362.4
ResNet	8M	0.16	175.6	0.16	185.1
	3M	0.17	182.9	0.17	192.1
	1M	0.15	174.0	0.15	177.6

Table 3.4 – Evaluation of the quality of synthesis of high-resolution dynamic MR images in comparison to low-resolution static MR images on real data. "$X \rightarrow Y$" and "$Y \rightarrow X$" refer to the synthesis of low-resolution static MRI and high-resolution dynamic MRI, respectively. The evaluation is conducted using two unpaired image metrics: KID and FID.

Architecture	Parameters	PSNR ↑	SSIM ↑	KID ↓	FID ↓
UNet	8M	23.7	0.73	0.036	124.05
	1M	25.16	0.77	0.05	145.68
	400K	26.38	0.77	0.056	141.97
ResNet	8M	25.2	0.74	0.07	160.1
	3M	25.3	0.73	0.058	148.7
	1M	27.02	0.75	0.06	150.94

Table 3.5 – Evaluation of the synthesis quality of high-resolution dynamic MR images for different neural networks using simulated paired data. The evaluation is performed using two paired image metrics, PSNR and SSIM, and two unpaired image metrics, KID and FID.

source image. The synthesized texture appeared to be more realistic, and this method effectively removed the noisy aspect of the source data. However, the resulting image, despite the fully paired framework, did not achieve an anatomically accurate reconstruction compared to the ground truth image. It also demonstrated major inconsistencies in the fine structures, such as the bones and surrounding tissues textures. Table 3.5 presents the quantitative evaluations for each method. Despite the differing characteristics of the synthesized images, all the methods achieved comparable results for each metric, with the exception of KID, where UNet architectures exhibited superior performance. Given that KID is designed to reflect the shared visual similarities between images, we supposed that this KID improvement reflects the texture improvement produced by the network.

As better results are obtained on the simulated dataset, we seek to determine whether training in a fully-paired setup on simulated data can be generalized to real data. Figure 3.10 presents the results obtained by applying the network trained on the simulated data to the real data. It can be observed that if the ResNet does not provide any improve-

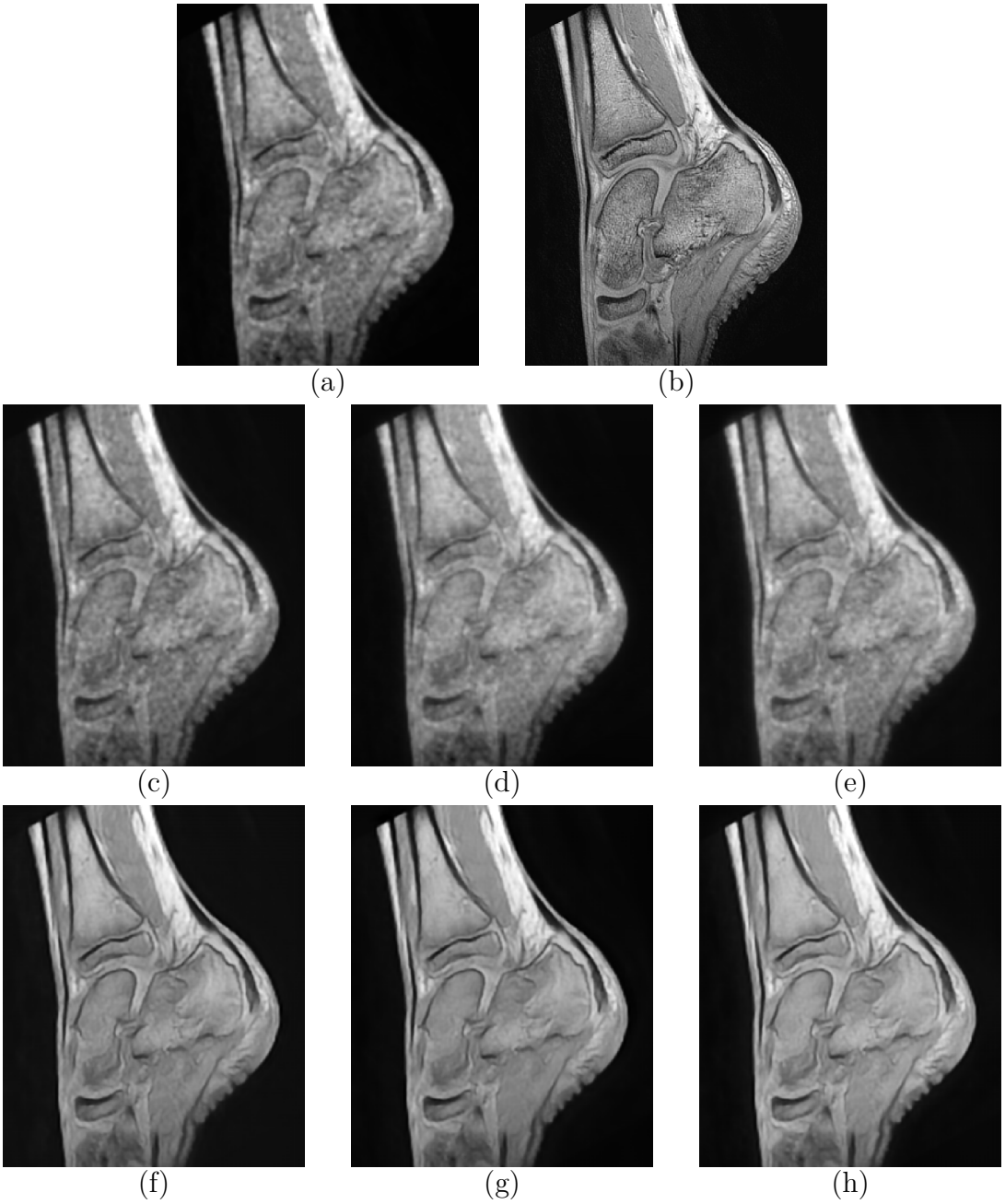


Figure 3.9 – Estimation of high-resolution MRI synthesis on simulated data using paired loss. Two distinct neural network architectures are used: ResNet [186] and UNet [185]. For each network, three different sizes are studied. (a) Simulated dynamic MR image. (b) Corresponding static MR image. (c) ResNet - 1M parameters. (d) ResNet - 3M parameters. (e) ResNet - 8M parameters. (f) UNet - 400K parameters. (g) UNet - 1M parameters. (h) UNet - 8M parameters.

ment in the quality of the source image, the UNet architecture may partially reduce the noise on the source image. However, the result remains substantially different from the intended static MR image. Furthermore, the direct application of the network trained on the simulated data on the real data does not yield comparable results between the two datasets, suggesting that the simulation model does not faithfully represent the real forward model.

As the handcrafted simulation model is not sufficiently realistic to permit a model trained on this data to generalize to real data during the test phase, we intend to evaluate the impact of transferring knowledge from a model trained on simulated data to a model adapted to real data. The model that provides the most accurate estimation of the inverse model on simulated data is then re-trained for an additional 50 epochs on real data. The results are presented in Figure 3.11. Transfer learning from the model trained on simulated data does not produce significantly more realistic results than when the model is trained from scratch on real data.

HCP data Figure 3.12 depicts the qualitative results of cross-contrast synthesis of HCP MR data. It can be observed that the performance of the architectures differs, with the UNet architecture demonstrating the most accurate estimation of the transformation. In contrast, the ResNet architecture yielded inaccurate synthesis of T2 MRI, leading to inconsistencies in contrast. Table 3.6 corroborates the aforementioned observations. These results demonstrate that in a paired setup, both architectures are able to estimate a consistent transformation from T1 MR image to T2 MR image. This evidence suggests that the inability of the networks to synthesize convincing high-resolution dynamic MR data is not intrinsic to the network itself, but rather a consequence of the data. This observation leads to the conclusion that the use of a paired setting for high-resolution dynamic MRI synthesis may not be suitable for the data in question.

3.3.4 Discussion

Two distinct neural network architectures were employed in order to estimate the inverse model and synthesize high-resolution dynamic MR images. The experiment was conducted on three datasets: a real dataset with partially paired data, a simulated dataset with fully paired data, and the HCP dataset.

Despite the promising results observed on the HCP dataset, neither the use of real data nor simulated data allows for the realistic synthesis of high-resolution dynamic MRI. The

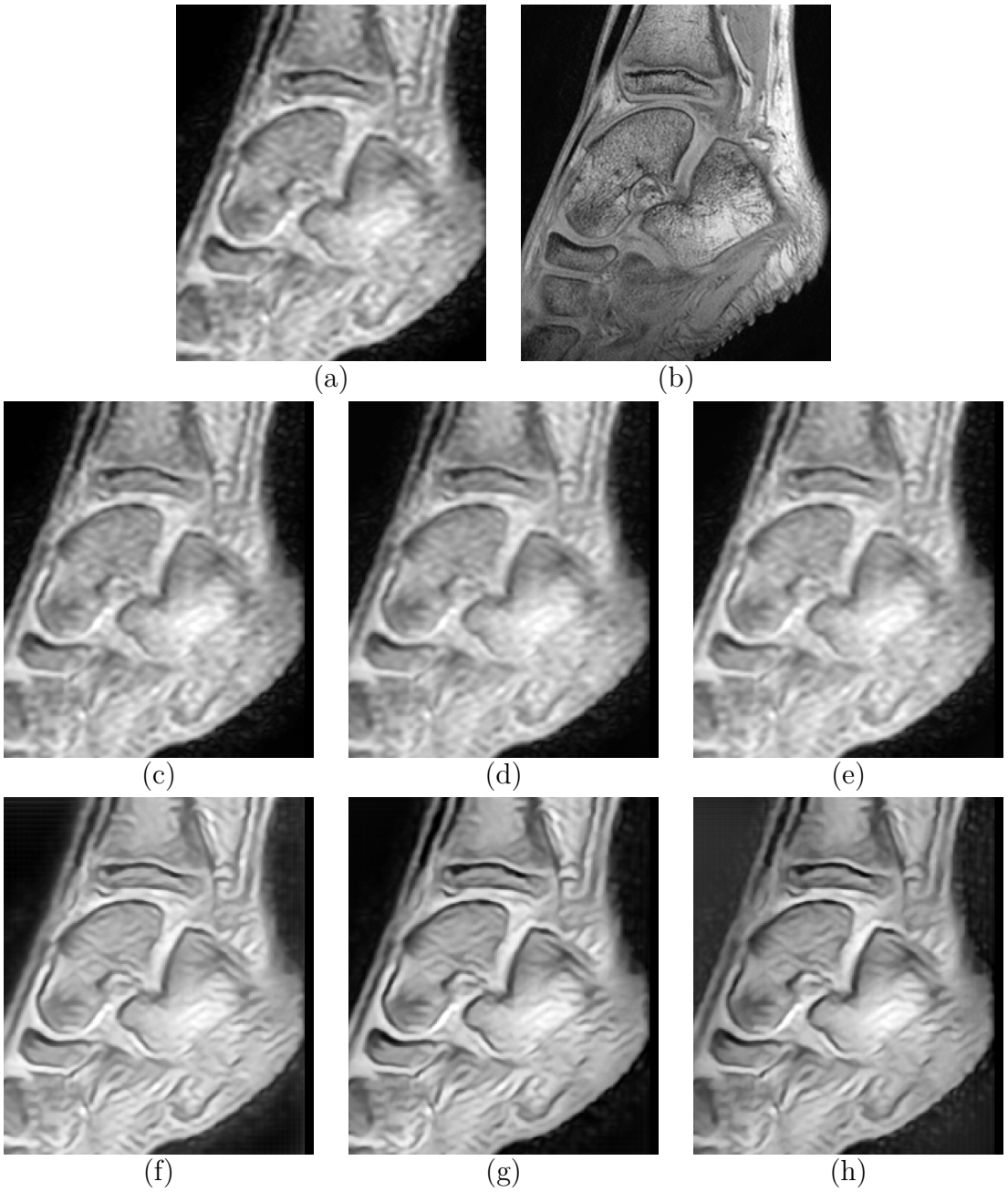
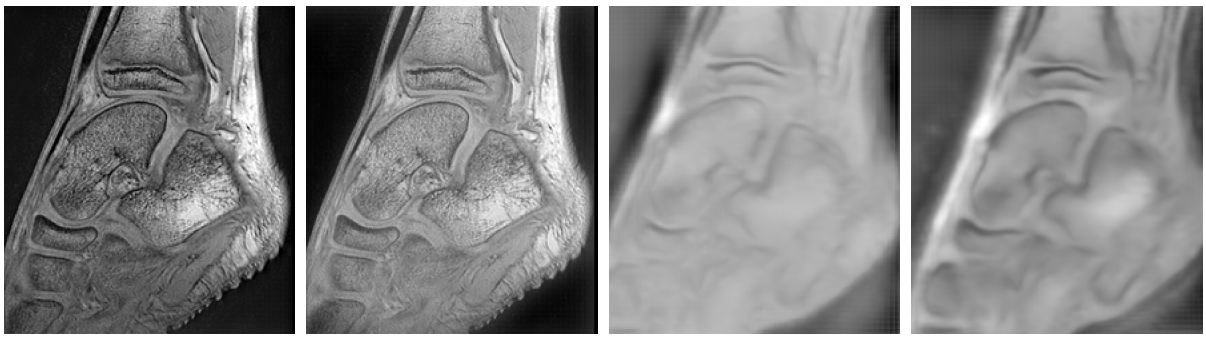


Figure 3.10 – Generalization of the high-resolution dynamic MR images synthesis model learned on simulated data to real data. Models trained on simulated data are directly applied to real data. (a) Original dynamic MR image. (b) Corresponding registered static MR image registered on the tibia. (c) ResNet - 1M parameters. (d) ResNet - 3M parameters. (e) ResNet - 8M parameters. (f) UNet - 400K parameters. (g) UNet - 1M parameters. (h) UNet - 8M parameters.



(a) (b) (c) (d)

Figure 3.11 – Estimation of high-resolution MRI synthesis on real data using transfer learning. (a) Original dynamic MR image. (b) Corresponding registered static MR image registered on the tibia. (c) Model trained from scratch. (d) Model trained for 150 epoch on simulated data followed by 50 epochs on real data.

	Parameters	PSNR ↑	SSIM ↑
UNet	8M	19.65	0.23
	1M	21.58	0.26
	400K	21.51	0.26
ResNet	8M	8.81	0.03
	3M	11.46	-0.008
	1M	12.21	0.0004

Table 3.6 – Evaluation of the synthesis quality of T2 MRI from T1 MRI on the HCP dataset. The evaluation is performed using two paired image metrics, PSNR and SSIM.

method employing the UNet architecture on the simulated dataset yielded more realistic results; however, these results have been demonstrated to be non-transferable to the real data. The size of the network did not significantly influence the synthesis performance.

3.4 Conclusions

This chapter explores the use of paired image synthesis methods for high-resolution dynamic MR synthesis. The Equinus dataset is processed in order to register the static and dynamic MR images of a subject in the bone region. The outcome of the bone-to-bone rigid registration is a partially paired dataset of static and dynamic MR images.

In order to estimate the inverse model for high-resolution dynamic MRI synthesis, two main approaches are explored. The first consists in attempting to estimate the inverse

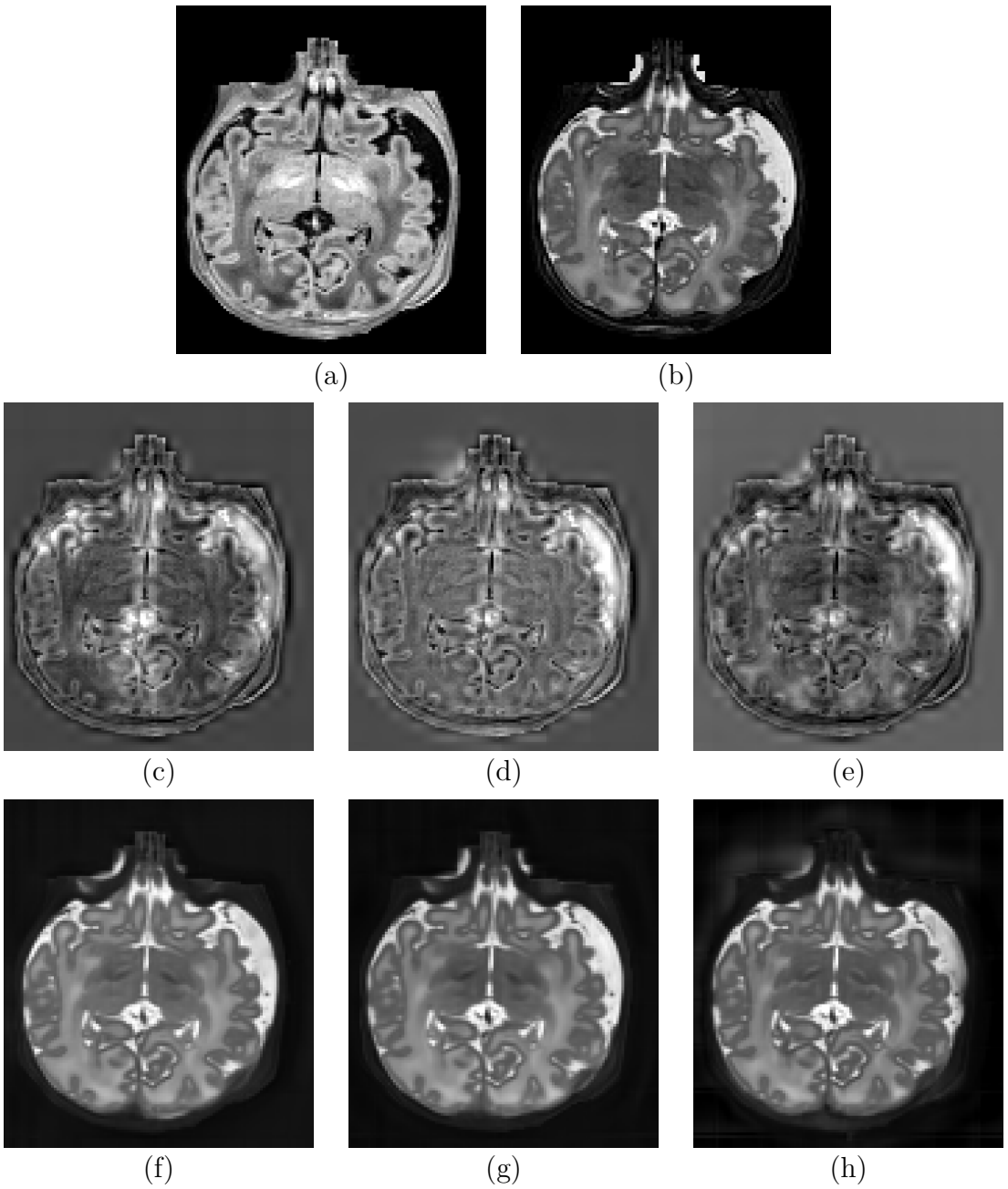


Figure 3.12 – Estimation of T2 MRI from T1 MRI on the HCP dataset. (a) Original T1 MR image. (b) Corresponding T2 MR image. (c) ResNet - 1M parameters. (d) ResNet - 3M parameters. (e) ResNet - 8M parameters. (f) UNet - 400K parameters. (g) UNet - 1M parameters. (h) UNet - 8M parameters.

model directly from the partially registered data. The second aims at solving the inverse problem by estimating the forward model to simulate dynamic data, which is theoretically easier to estimate, in order to generate new fully-paired data and ease the estimation of the inverse model.

Although the experiments conducted on the HCP dataset yielded realistic results, neither the direct estimation of the inverse model nor the use of an estimation of the forward model has produced a realistic synthesis of high-resolution dynamic MRI. This may be attributed to the quality of the registration process. Indeed, due to the large aspect difference between static and dynamic MR images, the registration process may be complex and arduous, potentially leading to small registration errors. Such errors may have a significant impact on the paired training. Furthermore, none of the learning-based methods for simulating dynamic MR data have yielded realistic synthesis results. Moreover, the handcrafted model appears to be insufficiently realistic to generalize the estimation of the inverse model to real data.

Given that paired methods for estimating high-resolution dynamic MRI from low-resolution dynamic MRI have been unsuccessful, the following chapter will explore the use of unpaired I2I methods.

Chapter 4

Unpaired synthesis of high-resolution dynamic MRI

Abstract

This chapter proposes an investigation into the potential of unpaired image-to-image translation methods for the synthesis of high-resolution dynamic MRI sequences. A significant proportion of these methods are based on disentangled representation learning. By relying on the factorization of an image into independent variation factors, this approach offers greater control over the result of the synthesis than GAN-based ones. However, disentangling latent representations is not a trivial task and may be influenced by particular inductive biases. In this chapter, we first evaluate the impact of several hyperparameters and the addition of latent space constraints, and then study the impact of both the entanglement module and the addition of a segmentation auxiliary task on the result of the synthesis and the disentanglement of the representations. Our results demonstrate that the choice of the entanglement module greatly influences the learning of a good representation, and that the addition of a segmentation auxiliary task leads to better synthesis performances. Additionally, this method yields realistic synthesis of 4D+t dynamic MRI sequences. This

4.1 Introduction

Unpaired image synthesis aims to link two distinct visual domains that do not have corresponding images. As a reminder, an unpaired image-to-image translation task is defined as an image-to-image translation task where there is no exact reference image to which the translated image can be compared (see Figure 2.6). While paired image-to-image translation tasks can rely on classical losses such as MSE or MAE, unpaired settings take advantage of specific techniques such as cyclic consistency [46]. This section provides an overview of two key frameworks in unpaired image-to-image synthesis.

4.1.1 Learning unpaired image-to-image translation

Image-to-image translation is generally defined as the process of mapping an image from an original source image domain to another particular target image domain. Image domains do not necessarily refer to visual domains in human perception. Instead, they can refer to larger image-related domains. For instance, in medical imaging, the translation between the k-space and an MR image can be considered an I2I task, since k-space data corresponds to the MR images in the frequency domain.

Two major frameworks are used for unpaired I2I translation in medical imaging: **cycle GANs**, and **Disentangled Representation Learning** (DRL) frameworks. More recently, diffusion models have gained significant interest from the research community due to their ability to generate highly realistic images. Diffusion models represent a class of generative models wherein images are progressively corrupted with an increasing level of noise, which the models then learn to reverse. This generative process is defined as the reverse of a Markovian process, whereby white noise is progressively denoised to produce an image. These models have already been applied in unpaired image synthesis in computer vision [189], [190] and medical imaging [191], [192], [193]. Diffusion-based methods are capable of achieving diverse and highly realistic image synthesis, and have been shown to outperform GAN-based methods for certain tasks. However, the sampling process is performed in a non-disentangled data domain, such as the image domain. These methods are computationally expensive during training and inference steps. Cycle GANs refer to GANs that use cycle-consistency loss to achieve I2I translation, as previously described in Section 2.3.3. The original CycleGAN paper was released in 2017 by Zhu et al. [46] for unpaired image-to-image translation. Figure 4.1 provides an overview of the model. Considering two domains X and Y, the authors propose to use two GANs to form a cycle:

the generator of the first GAN aims at mapping images from X to Y, while the corresponding discriminator aims at distinguishing the original image from domain Y from those generated by the generator. The second GAN operates on the same principle for the translation from domain Y to X.

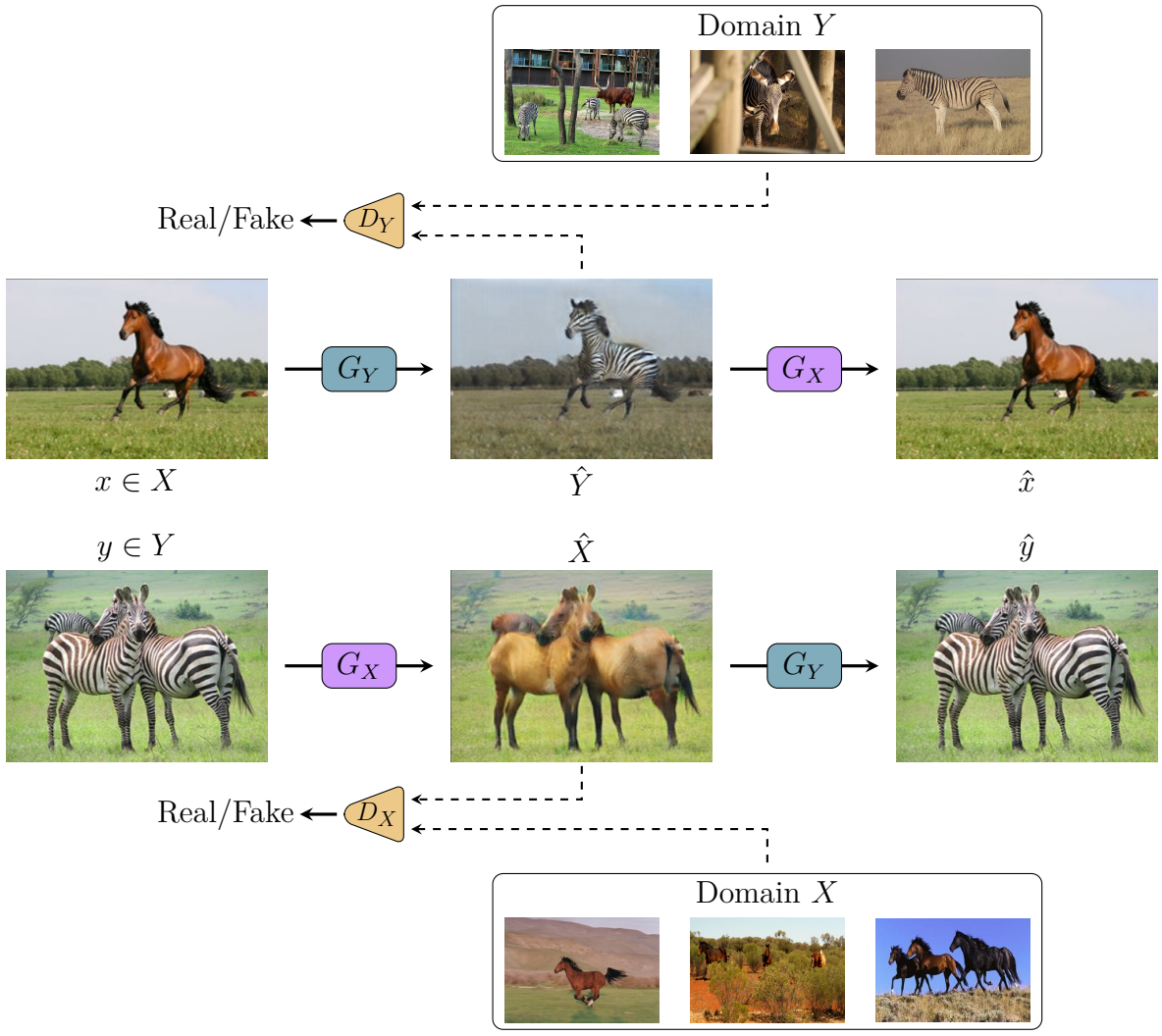


Figure 4.1 – CycleGAN [46] overview. The CycleGAN framework employs two GANs working in a cycle to achieve image-to-image translation between domains X and Y. The model consists of two mapping functions $G_X : Y \to X$ and $G_Y : X \to Y$ and the corresponding adversarial discriminators D_X and D_Y. D_X encourages G_X to map the elements of Y to elements that are indistinguishable from those of X and vice versa for D_Y. A cycle-consistency loss computes the difference between x and $\hat{x}$ (and respectively y and $\hat{y}$), expressing the intuitive property that the translation from one domain to another followed by the inverse transformation should output the original input image. Images extracted from [46] and ImageNet dataset [108].

Cycle GANs are known for their ability to generate realistic and high-quality images for various tasks. However, they are also known for producing structural inconsistencies between the original image and the synthesized image [194]. Additionally, the structure of Cycle GANs provides limited control over the synthesis results, which may lead to inaccurate synthesis results. In medical image synthesis, it is critical to develop AI models that are trustworthy and accurate in their predictions for diagnosis and follow-up processes. The models should aim for anatomically accurate predictions and a more explainable synthesis process.

4.1.2 Disentangled representation learning

Although deep learning models provide state-of-the-art performance in many image synthesis tasks, their exact decision process is often intractable. Therefore, trustworthy AI is gaining popularity, especially in the medical imaging synthesis community, where transparency in the decision-making process is crucial. Disentangled representation learning is a step towards this objective by constructing a latent space representation that is both human-interpretable and manageable.

In deep learning, representation learning is the process of learning a data representation to ease information extraction for downstream tasks [195]. As a result, finding a "good" representation for a particular task is a critical issue in deep learning. Disentangled representation learning aims to find the representation that captures the underlying factors of variation in the data distribution. Ideally, each of these factors represents an independent, semantically meaningful aspect of the data that is humanly understandable. Such a representation then allows to generate an image with a control over each semantic aspect. As an example, consider a set of images of human faces. Each face can be described by a set of features such as eye color, face shape, eyeglasses, hair color, and freckles. A disentangled representation that is aligned with these factors of variation may provide a way to generate an entirely new face by adjusting each corresponding factor in a generation process. Figure 4.2 provides some examples of disentangled factors.

Thanks to their flexible structure, disentangled representations are applicable across various modalities and tasks [125]. In image processing, they are commonly used for image generation [196] (registration [197][198], cross-modality synthesis [199][200], causal synthesis [201], face generation [202], [203], colorization [204] or domain adaptation [205]), image segmentation (single-modal [97], [206], [207] or multi-modal [208], [209]), image classification [210], [211], [212], [213] or anomaly detection [214], [215], [216].

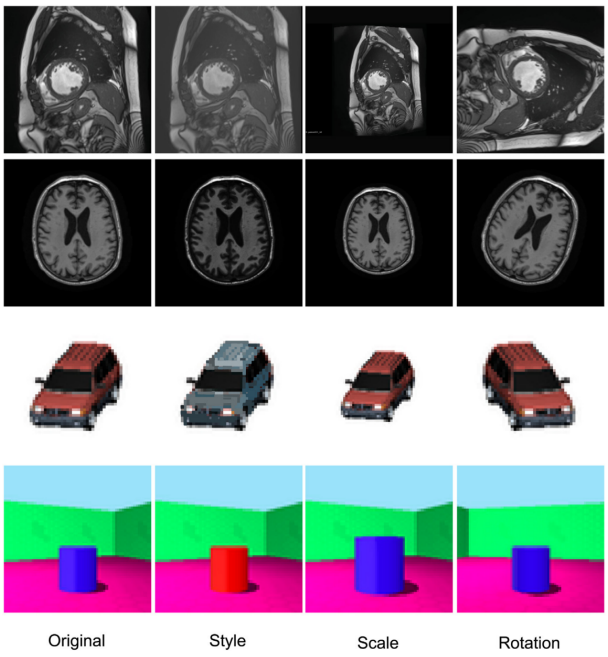


Figure 4.2 – Examples of factors of variation in four different applications: cardiac and brain scans, cars and 3D shapes. For each application, style, scale and rotation factors are considered. Image reproduced from [121].

Definition A representation is typically considered 'disentangled' if an alteration of a single factor results in a specific change only along the corresponding mode of variation within the data, without affecting the other characteristics [124]. However, this definition is subjective and has been criticized for being ineffective in dealing with the large number of possibilities. As there is no clear consensus on the definition of disentanglement, several approaches have been proposed in the last decade to formalize the concept [123], [195], [217], [218], [219]. For instance, Do and Tran [217] proposed a definition based on information theory that identifies three key properties of disentangled representations:

— *Informativeness*: The mutual information between a datum x and its corresponding latent code z^i should be large, since z^i is supposed to carry information about x

— *Separability and independence*: Given two latent codes z^i and z^j extracted from x, the multivariate mutual information between x, z^i, and z^j should ideally be zero, meaning that z^i and z^j contain no redundant information about x.

— *Interpretability*: Given a ground truth g^i corresponding to each z^i, the entropy of z^i, the entropy of g^i, and the mutual information between z^i and g^i should be equal.

This definition characterizes a perfect disentangled representation when ground truth factors of variations of the data are known. However, in most practical applications, these ground truth factors are not available to guide the disentanglement. Therefore, achieving interpretability without explicit guidance for the latent codes encoding the factors of variation is a major challenge.

Four main frameworks are used to learn disentangled representations: Variational Autoencoders (VAE) [220], Normalizing Flows (NFs) [221], GANs, and more recently, Content-Style disentanglement frameworks [95]. All of them are illustrated in Figure 4.3. The review by Liu et al. [121] provide additional information on each framework for disentangled representation learning. While VAE, NFs, and GANs encode the disentangled factors in the different dimensions of their latent vector representation, content-style disentanglement frameworks encode their disentangled factors in distinct latent variables. This approach is further explained in the following paragraph.

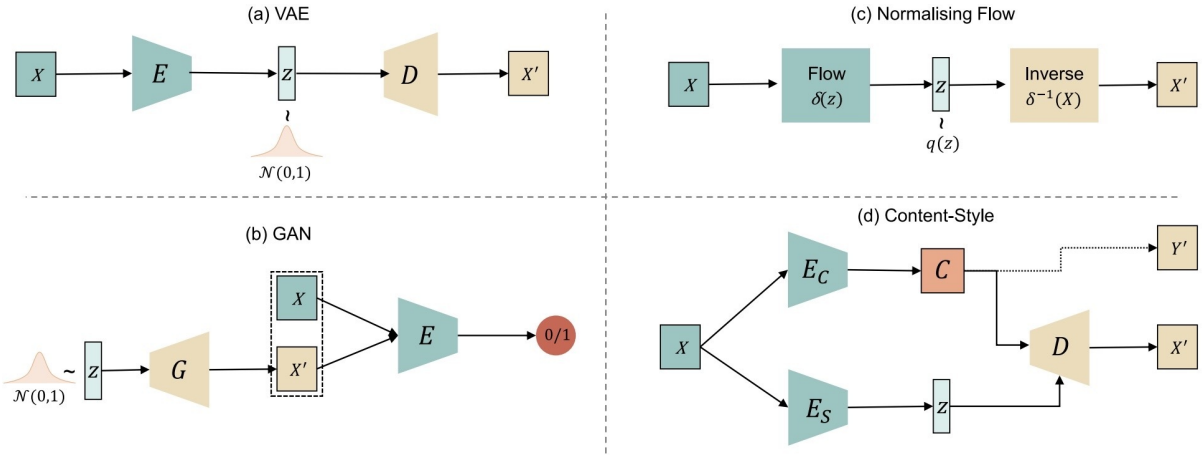


Figure 4.3 – Main frameworks in Disentangled Representations Learning. (a) VAE (b) GAN (c) Normalizing Flows (d) Content-style disentanglement. X and X' are respectively input and reconstructed images. z and C are the latent representations while C represent spatial tensor variable (e.g. image content) and z represents a vector latent variable. The dashed line in (d) denotes the use of C for learning a representation Y' for a parallel equivariant task (e.g. semantic segmentation). Finally, $\mathcal{N}$ denotes the normal distribution with zero mean and unit variance, whilst $q(z)$ can be any prior distribution. Image reproduced from [121].

Content-style disentanglement A well-known framework for learning disentangled representations is content-style disentanglement. It has been widely used in medical imaging and computer vision for domain adaptation [127], style transfer [95] and segmentation [97]. This type of disentanglement encodes the factors of variation into distinct latent

variables. In typical content-style disentanglement frameworks, the image is decomposed into a domain-specific 'style' and a domain-invariant 'content'. This framework is mostly applied in image processing, where content and style are quite intuitive, but it has also been applied in other modalities such as speech synthesis [222]. In I2I translation, the content latent code is usually encoded into a tensor variable that can preserve the spatial correlations within the data. Such decomposition allows for the reuse of the latent variable in equivariant tasks.

Disentangling style and content is a non-trivial process without any ground truth for the latent variable to be extracted, and especially in unpaired I2I. To ensure the correct information is extracted, a set of constraints can be applied. A good disentanglement may be favored by specific inductive biases, such as architectures, latent space manipulations, or learning schemes. In practice, various aspects of the framework can be exploited to improve disentanglement, such as the choice of entanglement module, the addition of latent space constraints, specific learning schemes, and, in the case of unpaired image-to-image translation, the use of an equivariant task as regularization. The following paragraphs provide a detailed explanation of these methods.

Entanglement module The entanglement module serves the purpose of merging various latent representations within a generator to produce a new image. The effectiveness of this merging, or entanglement, is critical in disentangled representations. The following paragraphs use the terminology of style transfer. The term 'content' refers to domain-invariant spatial information, while 'style' refers to domain-specific information.

Many popular techniques for merging information in recent literature are based on Conditional Normalization (CN). CN is a concept derived from normalization layers, such as Batch Normalization [223], where the affine parameters of the transformation are learned from conditioning information. These methods are designed to modulate the intermediate activations of the network in a domain-specific manner. The most popular CN methods include Conditional Instance Normalization (CIN) [224], Feature-wise Linear Modulation [225], and Adaptative Instance Normalization (AdaIN) [118]. Initially, the FiLM layer is introduced as a conditioning layer for visual reasoning tasks and is proposed as a generalization of CIN. The FiLM module applies a channel-wise affine transformation to the intermediate activations of a neural network. Two multi-layer perceptrons (MLPs) learn the parameters of this transformation from an arbitrary input condition. The applications of this technique range from medical image segmentation [97], soundscape editing

[226], to vision language control for robotics [227]. The AdaIN module was originally designed for arbitrary style transfer but has since been widely used in literature for various conditional generation tasks, including text-to-image generation [228], motion synthesis [229], colorization from audio [230], text-to-speech synthesis [231] and face generation [202]. Similar to the FiLM layer, AdaIN acts as an affine transformation in the feature spaces. However, unlike FiLM, the affine transformation parameters are not learned but extracted from the input condition through the first and second-order statistics. The transformation is achieved by adjusting the channel-wise mean and standard deviation of the network's intermediate activations based on those from the condition across spatial locations. Originally designed for arbitrary style transfer in real-time, this technique breaks with the previously introduced approaches by using a non-learnable transformation that dispenses with the iterative optimization process of the other modules.

Both FiLM and AdaIN are commonly used in content-style disentanglement frameworks. In these applications, the content is taken as input by the generator and the style is used as conditional input. Although [232] used a concatenation between style and content, this technique may hinder the correct disentanglement between content and style [121]. Besides these well-known entanglement modules, some methods have designed their own module to fuse style and content such as [126]. To the best of our knowledge, no studies have been conducted to measure the impact of the entanglement module on the synthesized images in content-style disentanglement.

Promote and measure disentanglement Learning disentangled representations is still a challenging task, especially in unpaired frameworks. A successful framework for disentangled representation learning relies on the independence between the latent representations. However, most common applications of such frameworks do not provide any ground truth for the latent codes. One of the major challenges in these methods is to promote the disentanglement, and thus the independence, between the latent representations of the variation factors without knowing the exact realizations of these factors. Therefore, additional strategies are introduced to enforce the disentanglement of the latent representations without explicit supervision of the latent variables.

Most of the time, unpaired I2I translation methods based on disentangled representation rely on a cycle consistency loss to guide the translation process. This objective, introduced by Zhu et al. [46] constitutes the state-of-the-art in unpaired I2I translation. This objective has proven to be efficient in guiding inter-domain translation, providing

strong regularization, and stabilizing training [233]. However, cycle-consistency by itself cannot guarantee correct disentanglement of latent representations. Recently, some approaches have introduced constraints in the latent spaces to further control the latent representations themselves and their disentanglement, rather than working solely in the image spaces.

These techniques fall into two main categories: those that manipulate the structure of the latent space, and those that manipulate its content. Those working on the structure of the latent representation typically aim to condition the distribution of data points in the latent space. The most popular approach was initially introduced in the VAE architecture [220], where the latent space is constrained to fit a normal distribution with a zero mean and unit variance. In generative models such as VAE, where new data samples can be generated from a sample latent code, a structured latent space has a significant advantage. Indeed, sampling a latent code of a VAE capable of generating realistic data is much easier if the latent space follows a normal distribution, which is well known in mathematics. Additionally, this type of modeling has been shown to encourage the unsupervised disentanglement of the factors of variation. Recently, many works have proposed the use of contrastive learning [199], [234], [235], [236]. Contrastive learning is a particular technique working on the latent representation of the data. The underlying idea is to favor the proximity between similar data in latent spaces and to push away the data which are different. This technique, particularly applied in self-supervised representation learning, is described in more detail further on in this chapter. These methods can encourage disentanglement by promoting a human-interpretable structure and favoring certain properties of the latent representations. The second category of methods focuses on constraining the content of the latent representations. The challenge is how to favor certain semantic properties in the latent representations without knowledge of any realization of the latent codes. As previously mentioned, the content representation in content-style disentanglement is particularly suitable for equivariant tasks. In imaging tasks, it mostly concerns semantic segmentation. Requiring the content to convey the necessary anatomical properties for segmenting anatomical structures has been shown to improve the learned representation, as demonstrated by Chartsias et al. [97]. On the other hand, Lee et al. [126] introduce a content discriminator to encourage the content representations to be indistinguishable between the two modalities. This type of prior enforces domain-invariance, allowing the style to convey domain-specific information.

Although there is a growing number of approaches that focus on learning disentangled

representations, the metrics for quantifying this disentanglement are still relatively scarce. Most of the existing metrics are applicable either when the ground truths corresponding to the variation factors are known, or when the disentanglement of the factors occurs within a single latent vector variable (e.g., GAN or VAE) [203], [237], [238], [239], [240]. Content-style disentanglement often presumes that the ground truths for the variation factors are inaccessible. Additionally, the different representations are encoded as multiple latent variables of various dimensionalities (style is often represented as a vector, while content is often represented as a spatial tensor). In practice, only a few disentanglement metrics are designed to overcome the need for ground truth.

4.1.3 Application of Unpaired Image Synthesis to high-resolution dynamic MRI synthesis

This work aims to estimate the transformation for synthesizing a high-resolution dynamic MRI data from a low-resolution dynamic MRI data and a high-resolution static MRI in an unpaired setting. The approach is based on DRIT++, an unpaired image synthesis framework exploiting disentangled representation learning proposed by [126]. This chapter proposes an evaluation of this framework for high-resolution dynamic MRI synthesis in an unpaired setting. The main contributions are divided into two categories. The first focuses on the model structure and hyperparameters, while the second examines ways to improve the disentanglement of latent representations. These contributions are listed below:

— We conduct an evaluation of the influence of some architectural hyperparameters on the synthesis result. Specifically, we examine the impact of model size and discriminator architecture.
— We demonstrate the influence of the patch size on the synthesis accuracy.
— We evaluate the impact of the entanglement module on the learning of disentangled representations. We investigate several popular modules in the literature and examine their effects on the application case of unpaired image synthesis.
— We demonstrate the impact of incorporating an auxiliary segmentation task on the synthesis result.
— A quantitative evaluation of the disentanglement is conducted, and a subsequent analysis of the latent space is provided.
— We evaluate the impact of latent space constraints on the synthesis result.

— We apply the framework to dynamic MRI sequences, resulting in the synthesis of high-resolution dynamic MRI 4D+t sequences.

This chapter describes the synthesis method and methodological contributions. Details on the experimental context, including datasets used, metrics, and implementation details, are then provided. The results are presented, followed by a discussion.

4.2 Methods

Given a set of dynamic low-resolution MRI data Y and a set of static high-resolution MRI data X. The goal is to estimate the transformation that synthesizes a high-resolution dynamic MRI data from a low-resolution dynamic MRI data, meaning to estimate the transformation mapping an element $y \in Y$ into an element of X. We inspired from DRIT++, an unpaired image synthesis framework based on disentangled representation learning, proposed by [126].

Starting from this framework, we conduct an experiment to study the impact of several hyperparameters on the synthesis result. We experimented with different architectural hyperparameters, such as the size and architecture of the network. Then, we conduct a series of experiments to study the impact of DRL framework components on the synthesis accuracy. The study examines three main ideas: the impact of the entanglement module, the contribution of segmentation as a task prior, and the impact of latent space constraints on the synthesis result. The next section is organized as follows: first, we introduce the disentanglement learning framework used in this work. Second, we describe the techniques implemented to promote disentanglement.

4.2.1 Unpaired Image Synthesis

Consider a pair of MR images acquired from the same subject, denoted as $x \in X$ and $y \in Y$. This work aims to disentangle the content, referred to as "anatomy" in medical imaging terminology, from its representation, which aligns with the concept of "style" in broader literature. This representation, denoted as "modality," refers to the MRI acquisition parameters specific to each image and embodies modality-specific attributes. Typically represented as a vector, it does not encode spatial information but rather the rendering of the anatomical structures. Conversely, the anatomy represents the modality-invariant, spatial information within the image. Encoded as a tensor, this

representation preserves the spatial correlations of the original images, making it suitable for tasks requiring equivariance.

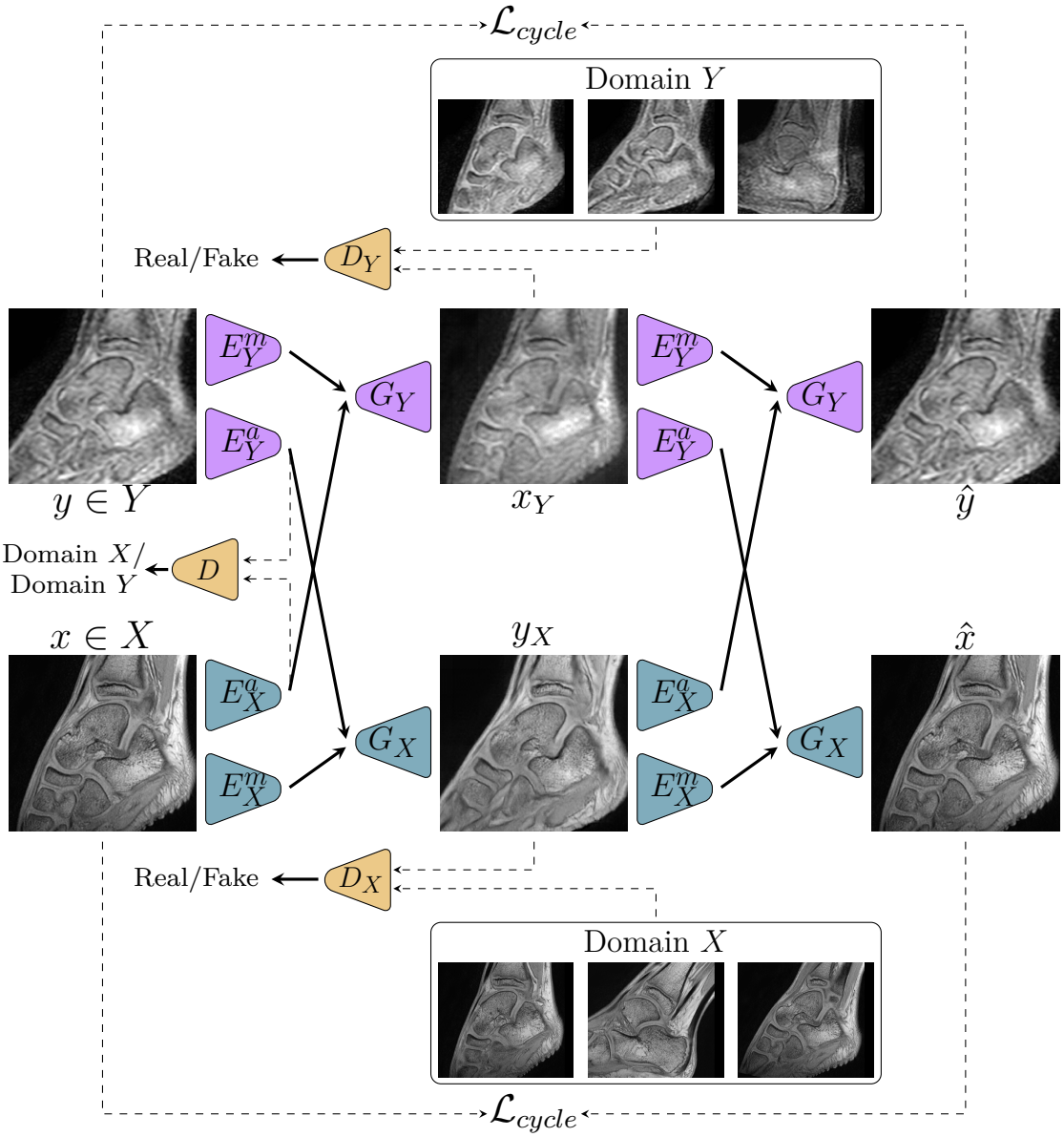


Figure 4.4 – Overview of the unpaired I2I synthesis process. Each image $x \in X$ and $y \in Y$ is factorized into anatomy and modality latent codes by E^a and E^m. Generators G recombine the latent codes and generate a new image according to the input latent codes. Once the cross-modality synthesis is complete, the inverse operation is performed to recover the original images and ensure the cycle consistency.

For each image, dedicated modality-specific, fully-convolutional encoders extract two latent codes representing the disentangled factors. We denote by E_X^a and E_Y^a the anatomy

encoders and E_X^m and E_Y^m the modality encoders. The anatomy encoders aims at mapping images into a modality-invariant, shared latent space while the modality encoders aims at mapping the images into modality specific, independent latent spaces. The extracted latent codes are denoted z^m for the modality latent code and z^a for the anatomy latent code. The disentanglement process is described as follows:

$$\begin{cases} x = (z_x^a, z_x^m) = (E_X^a(x), E_X^m(x)) \\ y = (z_y^a, z_y^m) = (E_Y^a(y), E_Y^m(y)) \end{cases} \tag{4.1}$$

Since each pair (x, y) consists of images acquired from the same subject, they are expected to share the same anatomical properties despite differing modalities. Disentanglement is enforced through a discriminator work in the anatomy latent space and shared weights in the networks. This discriminator is denoted by D and aims to identify the source modality of a given anatomy latent code. This setup encourages the anatomy representation to be modality invariant. The corresponding adversarial loss is expressed as follows:

$$\begin{aligned} \mathcal{L}_{\text{adv}}^a(E_X^a, E_Y^a, D) = \mathbb{E}_x[\frac{1}{2}\log D(E_X^a(x)) + \frac{1}{2}\log(1 - D(E_X^a(x)))]+ \\ \mathbb{E}_y[\frac{1}{2}\log D(E_Y^a(y)) + \frac{1}{2}\log(1 - D(E_Y^a(y)))] \end{aligned} \tag{4.2}$$

Two generators, one for each modality, are conditioned on both anatomy and modality latent codes to generate a new image. G_X stands for the generator in the X domain while G_Y stands for those in the Y domain. A cross-modality synthesis from a source modality to a target modality is achieved by providing to a generator the anatomy latent code of the source image and the modality latent code of the image from the target modality. We denote x_Y and y_X the synthesized cross-modality images. The generation process is described by the following equation:

$$\begin{cases} x_Y = G_Y(z_x^a, z_y^m) \\ y_X = G_X(z_y^a, z_x^m) \end{cases} \tag{4.3}$$

To address the unpaired setting, the model is conditioned on both image and latent spaces. Similar to the CycleGAN framework, we first introduce an adversarial loss to constrain the generated images' data distribution to approximate that of the target modality's real images. Two discriminators, one for each modality and denoted by D_X and D_Y are

introduced to this end. The corresponding adversarial loss, including both discriminators is expressed by Equation 4.4. Given x' and y', two "real" data from X and Y, respectively:

$$\begin{aligned}\mathcal{L}^m_{\text{adv}}(G_X, D_X, G_Y, D_Y) &= \mathbb{E}_{x' \in X}[\log D_X(x')] + \mathbb{E}_{x \in X, y \in Y}[\log(1 - D_X(G_X(z^a_y, z^m_x))] \\ &\quad + \mathbb{E}_{y' \in Y}[\log D_Y(y')] + \mathbb{E}_{x \in X, y \in Y}[\log(1 - D_Y(G_Y(z^a_x, z^m_y))] \\ &= \mathbb{E}_{x' \in X}[\log D_X(x')] + \mathbb{E}_{x \in X, y \in Y}[\log(1 - D_X(y_X)] \\ &\quad + \mathbb{E}_{y' \in Y}[\log D_Y(y')] + \mathbb{E}_{x \in X, y \in Y}[\log(1 - D_Y(x_Y)]\end{aligned} \tag{4.4}$$

Secondary, the framework leverages a cycle-consistency loss to stabilize the training, introduce a regularization, and enforce content preservation through the domain translation. $\hat{x}$ and $\hat{y}$ (expressed by Equation 4.5) denote the images obtained after two inter-modality translations:

$$\begin{cases} \hat{x} = G_X(z^a_{x_Y}, z^m_{y_X}) \\ \hat{y} = G_Y(z^a_{y_X}, z^m_{x_Y}) \end{cases} \tag{4.5}$$

The cycle-consistency loss is then expressed by the following equation:

$$\mathcal{L}_{\text{cycle}} = ||x - \hat{x}||_1 + ||y - \hat{y}||_1 \tag{4.6}$$

In addition to the previously introduced objective functions, several others are used in order to ease the training. A self-reconstruction loss compares the original image with its reconstruction obtained by using latent codes from the original image:

$$\mathcal{L}_{\text{self}} = ||x - G_X(z^a_x, z^m_x)||_1 + ||y - G_Y(z^a_y, z^m_y)||_1 \tag{4.7}$$

A latent regression loss enforces an invertible mapping between image and modality latent spaces, forcing images to contain the information encoded in the modality representation [127]. Given z^m_{rdn} a random modality latent code sampled from $\mathcal{N}(0, 1)$, the latent regression loss is expressed by Equation 4.8:

$$\mathcal{L}_{\text{latent}} = \frac{1}{n}\sum_{i=1}^{n} \left| z^m_{\text{rdn}} - E^m_X\left(G_X(z^a_y, z^m_{\text{rdn}})\right)\right| + \frac{1}{n}\sum_{i=1}^{n} \left| z^m_{\text{rdn}} - E^m_Y\left(G_Y(z^a_x, z^m_{\text{rdn}})\right)\right| \tag{4.8}$$

The global cost function is written:

$$\mathcal{L} = \lambda^a_{\text{adv}}\mathcal{L}^a_{\text{adv}} + \lambda^m_{\text{adv}}\mathcal{L}^m_{\text{adv}} + \lambda_{\text{latent}}\mathcal{L}_{\text{latent}} + \lambda_{\text{self}}\mathcal{L}_{\text{self}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}} \tag{4.9}$$

where λ^a_{adv}, λ^m_{adv}, λ_{latent}, λ_{self}, and λ_{cycle} are the dedicated weighting for each particular objective function. An overview of the method is proposed on Figure 4.4.

4.2.2 Latent space constraints

When using disentangled representations in an unpaired framework, there is a need to impose constraints on latent spaces without direct ground truth for the extracted latent codes. In the original DRIT++ framework [126], a constraint on the content space is introduced by the discriminator D. The discriminator encourages the projection of content latent codes from both MRI sequences into a common space, making them indistinguishable according to the discriminator. Both discriminators D_X and D_Y, on the other hand, operate in image space. The style space is only constrained by the function $\mathcal{L}_{latent}$, which enforces an invertible transformation between image and style space.

This study investigates the integration of two additional constraints on the latent codes. The first constraint involves a contrastive cost function applied in the modality space. The purpose of contrastive loss is to increase the similarity between an anchor (the current sample) and similar samples (i.e., those from the same class, referred "positive" examples) within a latent space, while simultaneously enforcing distance between the anchor and samples from different classes (referred to as "negative" samples). The purpose of using a contrastive cost function in the modality space is to encourage the model to learn representations in which similar modalities are clustered together while dissimilar modalities are pushed apart. Since the modalities of each image are known, a supervised function can be employed, as proposed by Khosla et al. [241]. By defining positive and negative examples based on the classes of the modalities, the model learns to map each modality to a distinct region in the latent space, thereby facilitating better control over the generation process. The supervised contrastive loss is defined as follows:

$$\mathcal{L}_{contrast} = \sum_{n \in [\![1,N]\!]} \frac{-1}{|P_n|} \sum_{p \in P_n} \log \frac{\exp(z^m_n \cdot z^m_p / \tau)}{\sum\limits_{a \in A_n} \exp(z^m_n \cdot z^m_a / \tau)} \tag{4.10}$$

Considering a set of N samples, let y_n denote the associated class (dynamic or static) with a given modality latent code z^m_n, $A_n = \{a \in [\![1, N]\!] \mid a \neq n\}$ represent the set of indices distinct from n, $P_n = \{p \in A_n \mid y_p = y_n\}$ the set of positive examples associated with the anchor z^m_n, and $\tau \in \mathbb{R}^+$ a scalar temperature parameter. The (.) symbol denotes the inner dot product.

The second constraint involves a cyclic cost function applied in the latent space of anatomy and modality. This cost encourages a one-to-one transformation between the different modalities, similar to a cyclic cost in image space.

$$\mathcal{L}^{a}_{cycle} = ||z^{a}_{x} - z^{a}_{x_Y}||_1 + ||z^{a}_{y} - z^{a}_{y_X}||_1 \tag{4.11}$$

$$\mathcal{L}^{m}_{cycle} = ||z^{m}_{x} - z^{m}_{y_X}||_1 + ||z^{m}_{y} - z^{m}_{x_Y}||_1 \tag{4.12}$$

By incorporating these additional constraints on the latent vectors, we aim to enhance the disentanglement of representations in the model. This will lead to more robust and interpretable image generation and manipulation capabilities in unpaired settings.

4.2.3 Entanglement module

Let consider the anatomy latent code $z^a \in \mathbb{R}^{N\times C\times H\times W}$ and the modality latent code $z^m \in \mathbb{R}^{N\times h}$, where N stands for the batch size, (C, H, W) refers to the spatial dimensions of the anatomy latent code and h to the last dimension of the modality latent code.

FiLM The FiLM module [225] learns the mapping function from z^m to the affine parameters γ and β. Both mapping functions are modeled by a MLP. γ and β are computed channel-wise and across spatial locations and aim at modulating intermediate activations of a neural network. The FiLM operator can be expressed by Equation 4.13 where (.) stands for the element-wise multiplication and (+) for the element-wise addition:

$$\text{FiLM}(z^{a}_{n,c}, z^{m}_{n,c}) = \gamma_{n,c}(z^m).z^{a}_{n,c} + \beta_{n,c}(z^m) \tag{4.13}$$

where $n \in [1, N]$, $c \in [1, C]$ and $\gamma, \beta \in \mathbb{R}^{N,C}$.

AdaIN AdaIN [118] was initially designed for arbitrary style transfer and also acts as an affine transformation in the feature space. Unlike FiLM, the AdaIN module has no learnable parameters and the parameters of the affine transformation are obtained from z^m statistics. While both FiLM and AdaIN modules are used for information merging, their approaches differ. The FiLM module directly applies an affine transformation to the intermediate activations derived from z^a. In contrast, the AdaIN module focuses on aligning the channel-wise mean and standard deviation of each intermediate activation

sample with the corresponding values from z^m:

$$\text{AdaIN}(z^a_{n,c}, z^m_{n,c}) = \sigma_{n,c}(z^m)\left(\frac{z^a_{n,c} - \mu_{n,c}(z^a)}{\sigma_{n,c}(z^a)}\right) + \mu_{n,c}(z^m) \tag{4.14}$$

Conv The method used in the DRIT++ framework [126] relies on successive concatenations between z^a and z^m, followed by convolution operations to merge both data. K_i denotes the $i \times i$ convolution kernel and $(\oplus)$ the concatenation operator:

$$\text{Conv}(z^a_{n,c}, z^m_{n,c}) = F(\text{IN}(F(z^a_{n,c}, z^m_{n,c}) * K_3), z^m_{n,c}) \tag{4.15}$$

where $F(f, z) = (a((f \oplus z) * K_1) * K_1)$, a is an activation function (e.g. ReLU) and IN is the Instance Normalization function.

4.2.4 Supervised Segmentation

This section details the introduction of a segmentation module into the framework described in Section 4.2.1. An overview of the proposed method is illustrated in Figure 4.5. As discussed in Section 4.1.2, auxiliary task such as segmentation provides a way to enforce disentanglement of the latent representations by encouraging particular properties in the content representation. The content representation in I2I translation is particularly suitable for equivariant tasks such as segmentation [97]. To this end, we leverage the ankle joint bones segmentation on the static MR images to introduce a supervised segmentation task from the anatomy representation.

This segmentation net takes as input the anatomy representation of the static MR image and outputs the corresponding segmentation of the ankle joint bones. Given segmentation maps of static data, this task is performed in a supervised manner. The goal is to constrain the anatomical information to be encoded into the anatomy representation in order to perform the segmentation while the modality-specific information remains in charge of the modality representation. As an objective function, we use a cross-entropy loss between the estimation provided by the segmentation net and the ground truth. This paired setting for the segmentation offers an additional constraint to guide the extraction of the latent representation.

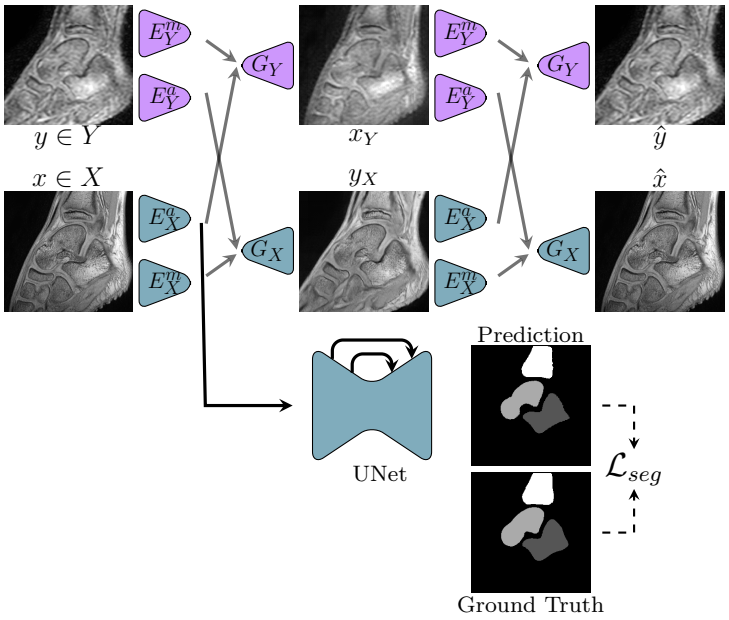


Figure 4.5 – Overview of the segmentation process. Each image, $x \in X$ and $y \in Y$, is factorized into anatomy and modality latent codes by E^a and E^m. As the static MRI is the only modality for which ground truth segmentation is provided, the segmentation training is conducted on this single modality. The anatomy latent code, denoted by z_X^a, is fed to the UNet model, which outputs the corresponding segmentation. The estimated segmentation is then compared to the ground truth through a cross-entropy loss function.

4.3 Experimental setup

This section provides a description of the experimental setup, including a reminder of the data used in the experiments in Section 4.3.1, the metrics considered for the evaluation the models (see Section 4.3.3), and the implementation details in Section 4.3.2.

4.3.1 Dataset

All the experiments are conducted on images from the Equinus dataset, described in Section 1.4.

4.3.2 Implementation details

The training set includes eleven subjects, five with equinus and six typically developing. The validation and test sets each include one subject, respectively typically developing and with equinus. All MR images were resampled to an intermediate resolution of $0.41 \times 0.41 \times 8$ mm. The resolution in the sagittal plane was chosen to be halfway between the static and dynamic resolutions to limit information loss in static images and interpolation artifacts in dynamic images. Finally, to limit the proportion of interpolated images in the training set and given the significant differences in resolution between static and dynamic images, the resolution in the frontal axis is kept the same as that of the dynamic images. The static MR images are preprocessed to be roughly registered to the dynamic images. This registration enables the use of image patches in the disentangled representation learning framework in an unpaired setting without using too dissimilar image patches for each modality. For example, a patch sampled from the foot in the dynamic image and a patch sampled from the background in the static image should not be used in parallel. As explained in Chapter 3, we employed 2D patches derived from MRI sequences for training. The use of 3D patches entails a significant increase in computational times and a slowing down of the training process. Moreover, the strong anisotropy observed in dynamic MR images between the resolution in the sagittal plane and along the frontal axis represents a significant limiting factor in terms of accuracy and robustness in volumetric image processing. To increase the quantity of training data, data augmentation was performed using the TorchIO library [184]. TorchIO is an open-source library designed for medical imaging. It provides tools for loading, preprocessing, and augmenting data. It enables the generation of MRI-specific artifacts, such as magnetic field inhomogeneity and motion

artifacts, as well as typical computer vision augmentations. All selected transformations can be applied with a specific probability and can be composed with others. Random transformations were applied to the data, including flips along the lateral axis, a bias field with a maximal magnitude of polynomial coefficient equal to 0.5, Gaussian noise with a standard deviation of 0.1, and affine transformations including scaling (with an amplitude of 0.2) and rotations (along the sagittal plane, with a 40° amplitude). The number of extracted patches was set to 95,400.

The model has been implemented using PyTorch. The framework is inspired by that of [126] but using lighter network architectures. Each encoder is fully-convolutional. The anatomy encoders, denoted E^a, consist of three convolutional layers followed by two residual blocks. The modality encoders, denoted E^m, consist of five convolutional layers followed by an average pooling layer and a final convolutional layer. We set $z^m \in \mathbb{R}^8$ and $z^a \in \mathbb{R}^{256\times32\times32}$. Similar to [202], each generator shares a common structure of a fully-connected mapping network that takes modality latent code z^m as input, followed by two residual blocks and three fractionally-strided convolutions. Each residual block shares the same fully-convolutional structure, except for the entanglement module. The content discriminator D is fully-convolutional and comprises three convolutional layers. Both modality-specific discriminators, D_X and D_Y, are PatchGAN discriminators introduced in CycleGAN [46], each set with three layers and Instance normalization. The segmentation network is a UNet [185], followed by fractionally-strided convolutions, with approximately 26K parameters. This model is referred as mDRIT++ hereafter.

In Section 4.4.2 the patch size is set to 64 × 64, and a reduced DRIT++ architecture is employed (referred as rDRIT++ hereafter). The model contains approximately 24 millions of trainable parameters. From Section 4.4.3 until the end of the chapter, the patch size is set to 128×128, and the mDRIT++ model is used, comprising approximately 26 millions of trainable parameters, in comparison to the original model's 87 millions.

For training, we use Adam optimizer, with a learning rate of 10^{-5}. The batch size is set to 32, exponential decay rates to $(\beta_1, \beta_2) = (0.5, 0.999)$ and weight decay to 0.0001. For all experiments, we set $\lambda_{\text{cycle}} = 10$, $\lambda_{\text{latent}} = 10$, $\lambda_{\text{reg}} = 0.01$, $\lambda_{\text{self}} = 10$. For the segmentation task, we set $\lambda_{\text{seg}} = 10$. If not specified, all remaining weights are set to 0. The source code for this project is available at `https://github.com/cScavinner/Unpaired_image_synthesis`. As a baseline, we use the CycleGAN model [46].

4.3.3 Metrics

A great part of the literature on image synthesis makes use of full-reference image quality assessment metrics such as PSNR, MSE or SSIM. These metrics rely on a strong physical foundation and have proven their efficiency and relevance. They are often low in term of computational cost and easy to set up and use. However, these metrics are not suitable for unpaired datasets since they compare a synthesized image with its corresponding ground truth. To evaluate the synthesis accuracy, we use two classical metrics in NR-IQA: FID and KID. Both are detailed in Section 2.4.2. We do not consider LPIPS metric as it takes into account the diversity of the generated images [242]. We compute the KID and the FID between the dynamic sequences of the test subject and the corresponding roughly registered static images. Each 3D image is split into a set of 2D images in the sagittal plane.

Despite the interest demonstrated in content-style disentanglement frameworks, very few metrics are designed to evaluate such disentanglement. Liu et al. [243] focus on quantifying the disentanglement in the case of style-content disentanglement. They propose using distance correlation [244] and a metric called 'Information over Bias' to measure correlation and informativeness. Distance correlation (DC) measures the degree of dependence between two random variables of arbitrary dimensions, and unlike Pearson correlation, DC is bounded in the interval [0,1]. A DC=0 indicates that the two random variables are independent. In the case of disentangling style and content, let X and Y be two matrices with N rows corresponding to N examples and an arbitrary number of columns. X and Y can refer to style, content, or linked images. As variables corresponding to content or images are spatial tensors, they are reformatted as vectors, resulting in a 2D matrix. The distance correlation is written as follows:

$$\mathrm{DC}(X,Y) = \frac{\mathrm{dCov}(X,Y)}{\mathrm{dVar}(X)\mathrm{dVar}(Y)} \tag{4.16}$$

with $\mathrm{dCov}(X,Y) = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{A_{i,j}B_{i,j}}{N^2}}$ the distance covariance between X and Y and $\mathrm{dVar}(X) = \mathrm{dCov}(X,X)$ the distance variance. A and B are the respective distance matrix of X and Y. The DC can be computed between the anatomy and the modality latent codes. DC values closer to 0 indicate a higher disentanglement. The DC can also be computed between those latent codes and the resulting image, indicating the level of dependence between the image and its generating factor. However, such metric alone cannot retrieve the disentanglement ability of a system. As indicated in [243],

$\mathrm{DC}(X, Y) = 0$ may indicate either X and Y encode unrelated information, or one of those encodes all the information and the other encodes noise, indicating full entanglement and posterior collapse. To this end, the authors introduced the Information over Bias (IoB) metric, which aims to measure the informativeness of generating factors relative to the corresponding image. IoB aims at comparing a reconstruction accuracy of particular images from uninformative representation and the estimated representation done by the disentangle representation learning model. Given z an estimated representation, $\mathbb{1}$ an uniformative constant tensor, and G_{θ_n} a neural network defined by its parameters θ_n aiming to reconstruct images I given a representation, IoB is defined as the expectation over the test images of the ratio between MSE obtained after the training of G_{θ_n} on the uninformative representation and the informative one:

$$\mathrm{IoB}(I, z) = \mathbb{E}_i \left[\frac{\mathrm{MSE}\Big(I_i, G_{\theta_1}(\mathbb{1})\Big)}{\mathrm{MSE}\Big(I_i, G_{\theta_2}(z_i)\Big)} \right] \tag{4.17}$$

When learning from the uninformative representation $\mathbb{1}$, G_{θ_n} only learns the dataset bias which can be modeled by θ_n, so, higher values of IoB can be associated with a higher amount of information encoded into the representation z. An IoB value of 1 means that no information about images I are encoded into the representation z.

4.4 Results

4.4.1 DRIT++ architecture

This section presents a comparative analysis of the various architectural modifications proposed for the original DRIT++ architecture, proposed in [126]. The objective is to both reduce the model size in order to save computational time on training and improve the synthesis performance through architectural adaptations. The original model is incrementally modified. Initially, the overall size of the model is reduced by decreasing the number of layers in each network. This model is designated as rDRIT++. Subsequently, the PatchGAN discriminator [56], also used in the CycleGAN model [46], [92], is incorporated instead of each modality encoder. This architecture is the one previously introduced as mDRIT++. Finally, the effect of patch size on the synthesis result is evaluated. Qualitative results are shown in Table 4.1. The efficacy of the methods is evaluated using the KID and FID metrics. Figure 4.6 provides a visual evaluation of the results of the

different methods, compared to the original static and dynamic images.

The quantitative evaluation shows that the model configuration with the PatchGAN discriminator and the larger patch size demonstrates the best performance according to both FID and KID. These results are corroborated by Figure 4.6. Each successive modification resulted in an improvement in the quality of the original synthesis, while maintaining a much smaller number of parameters. Both quantitative and qualitative analysis demonstrated a significant improvement in the quality of the synthesis with the introduction of the PatchGAN discriminator and the larger patch size. The quantitative evaluation indicated a greater improvement in the quality of the synthesis from the original model to the reduced one.

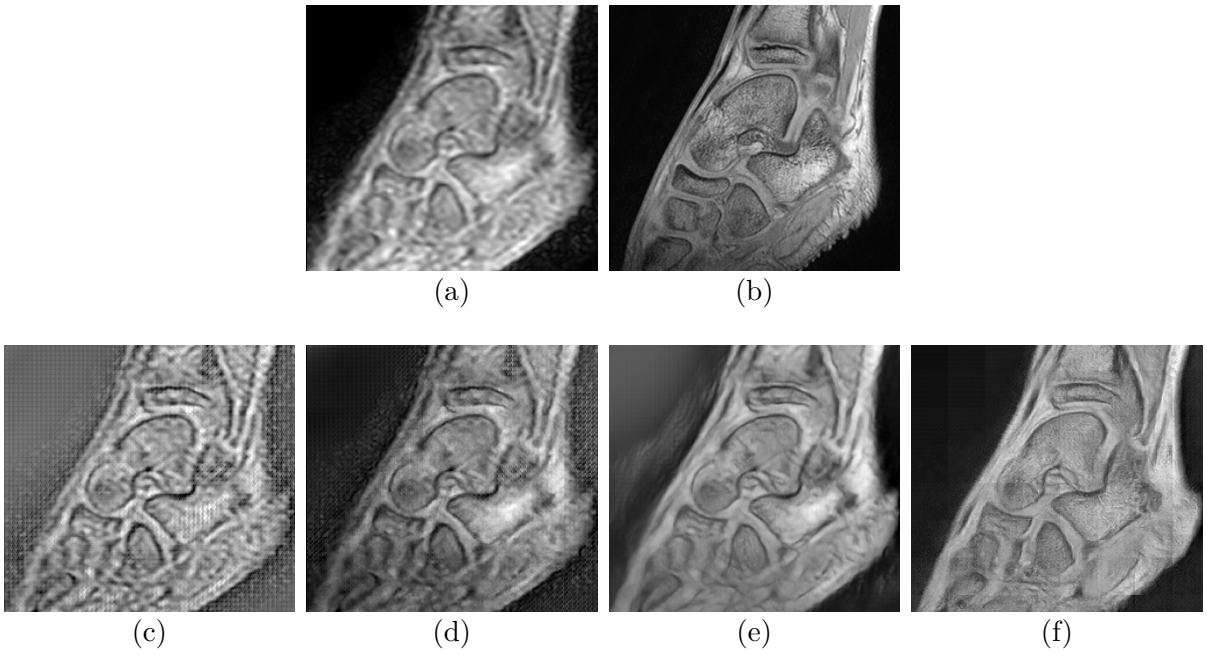


Figure 4.6 – High resolution dynamic MRI synthesis with different variants of the DRIT++ framework. (a) Original dynamic MR image. (b) Static MR image from the same subject. (c) Synthesis using the original DRIT++ architecture. (d) Reduced architecture (rDRIT++). (e) Reduced architecture combined with a PatchGAN discriminator [56] (mDRIT++). (f) Reduced architecture with PatchGAN discriminator (mDRIT++) and a patch size of 128×128 instead of 64×64. The best results are obtained using the latter configuration.

Architecture	Number of parameters	Discriminator	Patch size	FID ↓	KID ↓
Original	90M	original	64	380.99	0.54
Reduced	24M	original	64	275.56	0.33
Reduced	26M	PatchGAN	64	220.05	0.22
Reduced	26M	PatchGAN	128	**164.13**	**0.12**

Table 4.1 – Evaluation of the synthesis quality between different network architectures of the DRIT++ framework. Four settings were compared: the original network architecture presented in [126], a lightweight version of the same architecture, the lightweight version using a PatchGAN discriminator architecture proposed by [56] for D_X and D_Y, and a larger patch size.

4.4.2 Latent space constraints

Each method is trained on the unpaired Equinus dataset. The disentangled representation learning framework is compared with CycleGAN [46] whose network architecture includes around 28 million parameters. For a fair comparison, and in particular to avoid a bias in the results related to the number of network parameters, we compared it to the reduced version of the DRIT++ framework (rDRIT++), with a similar number of parameters to CycleGAN (about 24 million parameters instead of 87 million for the original DRIT++ code). In this implementation, the AdaIN module [118], is employed as the entanglement module within the generator architecture.

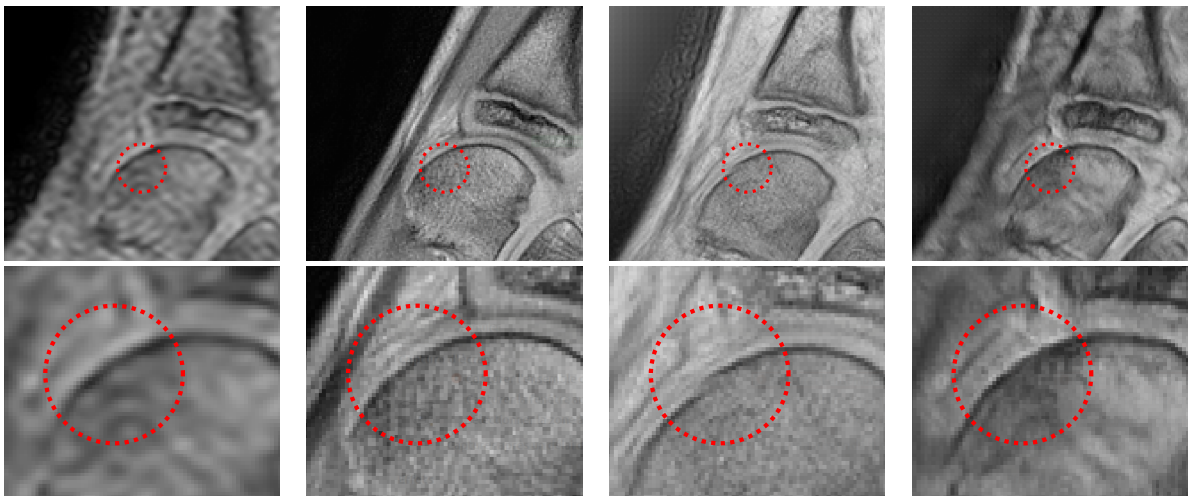
Quantitative results are shown in Table 4.2. According to the two quantitative metrics KID and FID, CycleGAN provides the best results, followed by the rDRIT++ with the cyclic constraint on the anatomy latent code. The other two constraints (contrastive loss and cyclic constraint on the modality latent code) do not provide any improvements. Visual evaluation is provided by Figure 4.7 showing the result of the cross-modality synthesis for CycleGAN and rDRIT++. Overall, as shown also in [189], we can see that rDRIT++ is not able to reproduce a realistic image (i.e. with a texture similar to the high-resolution data) compared to CycleGAN. However, the images produced by CycleGAN can demonstrate some structural inconsistencies, such as modifications to the bone structure, which may result in anatomically incorrect images.

4.4.3 Entanglement module

The disentangled representation learning framework and its variations are compared to CycleGAN. Figure 4.8 shows the results of using the three entanglement modules with and without the segmentation module for high-resolution dynamic MRI synthesis. Table 4.3 provides a quantitative evaluation of the synthesis in terms of KID and FID.

	$\mathcal{L}^m_{cycle}$	$\mathcal{L}^a_{cycle}$	$\mathcal{L}_{contrast}$	KID ↓	FID ↓
rDRIT++	✗	✗	✗	0.1818	209.64
	✓	✗	✗	0.2317	234.48
	✗	✓	✗	*0.1667*	*198.35*
	✓	✓	✗	0.2042	219.00
	✓	✓	✓	0.4248	309.46
CycleGAN				**0.1035**	**154.81**

Table 4.2 – Quantitative evaluation using KID and FID. Best performance is indicated in bold, and second best performance in italic.



(a) (b) (c) (d)

Figure 4.7 – High-resolution dynamic sequence synthesis results using CycleGAN and rDRIT++. For each sub-figure, the second row shows a zoom of the highlighted part of the image above. (a) Dynamic image. (b) Static image. (c) Estimation with CycleGAN. (d) Estimation with rDRIT++.

The Dice Similarity Coefficient (DSC) is included for the methods that incorporate the segmentation module.

DRL methods that use segmentation as a task prior demonstrate better performance than those without segmentation in terms of both KID and FID. The use of segmentation as a task prior globally improved the synthesis performance for each entanglement module. The best results were obtained using the AdaIN entanglement module in combination with the segmentation task. The FiLM entanglement module demonstrated the poorer synthesis performance. Figure 4.8 provides a visual assessment of the synthesis quality. The source dynamic and static images are provided as a reference. It can be observed that the combination of the AdaIN module with the segmentation network provides the best trade-off between image quality and anatomical accuracy. Although the DRIT++ entanglement module seems to provide clean images with and without segmentation, it appears to be more prone to artifacts and produces less regular edges. The AdaIN module without segmentation produces a realistic synthesis of textures but generates inconsistent edges and lacks realistic bone and cartilage shapes.

Figure 4.9 shows the results obtained for low-resolution static image synthesis. The majority of the methods employed in this study were found to generate images that were visually realistic. The only method that fails to generate a realistic texture is the one that employs the Conv module and the segmentation task. Table 4.4 provides a quantitative assessment of each methodology. As illustrated in Figure 4.9, the method combining the Conv module and the segmentation task exhibits the poorest performance in terms of KID and FID. In contrast to high-resolution dynamic image synthesis, the addition of the segmentation module has a detrimental effect on the performance of low-resolution static image synthesis on average. The best performance is obtained using the AdaIN module without the segmentation task, followed by the method combining the FiLM module and the segmentation task.

Sequence reconstruction Figures 4.10, 4.11, and 4.12 show the synthesis results for the DRL method for each entanglement module, respectively with and without segmentation as an auxiliary task, and the synthesis results for CycleGAN, compared to the method offering the best synthesis performance among the disentangled representation methods. As illustrated in Figure 4.8, the methods incorporating the AdaIN module appear to provide the best performance on temporal sequence. In contrast, the FiLM entanglement module shows the worst performance, with images exhibiting both artifacts

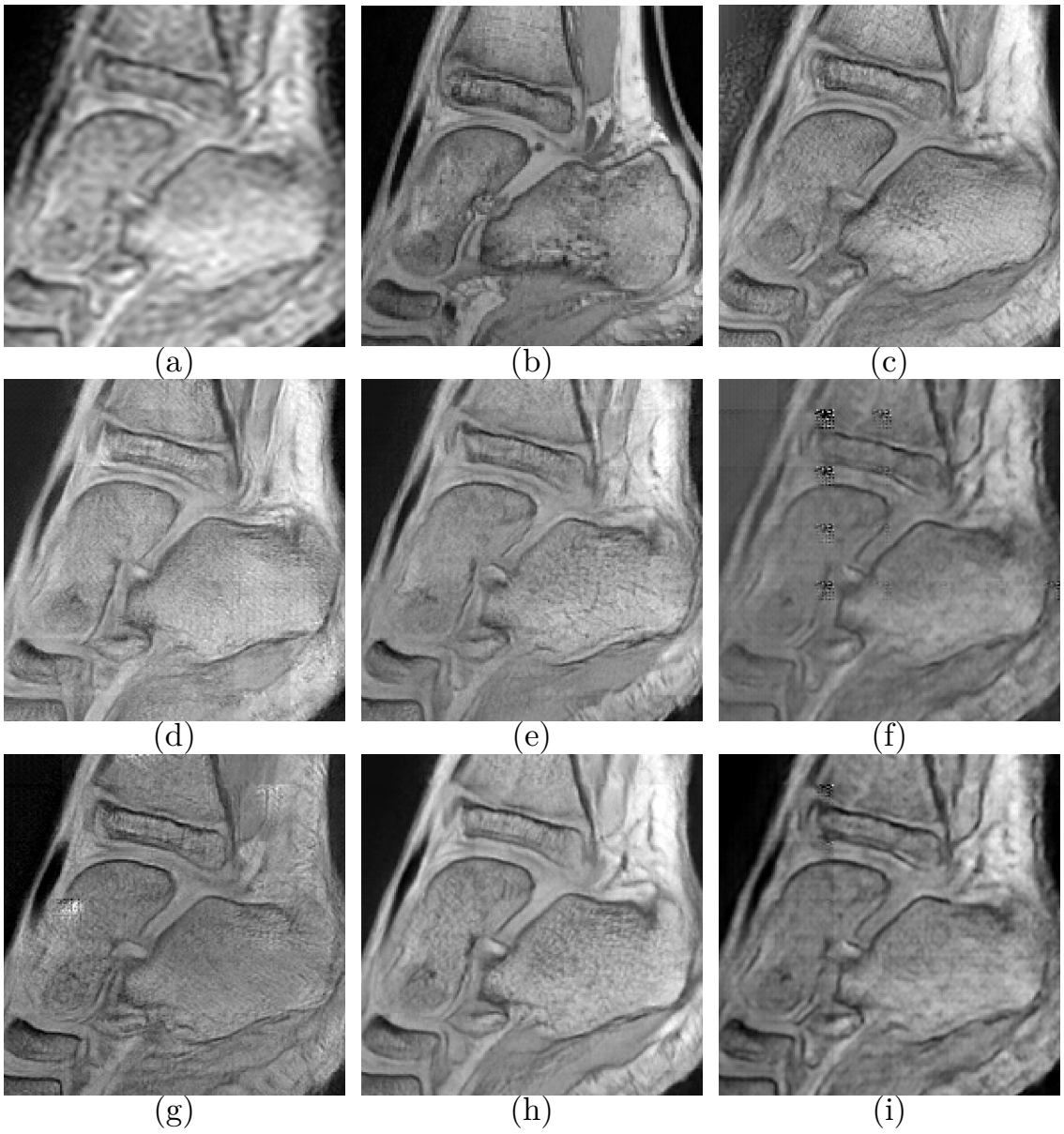


Figure 4.8 – High-resolution dynamic sequence synthesis results using CycleGAN and mDRIT++. (a) Dynamic image (b) Static image (c) Estimation with CycleGAN (d) Estimation with mDRIT++ - Conv (e) Estimation with mDRIT++ - AdaIN (f) Estimation with mDRIT++ - FiLM (g) Estimation with mDRIT++ - Conv with segmentation as an auxiliary task (h) Estimation with mDRIT++ - AdaIN with segmentation (i) Estimation with mDRIT++ - FiLM with segmentation.

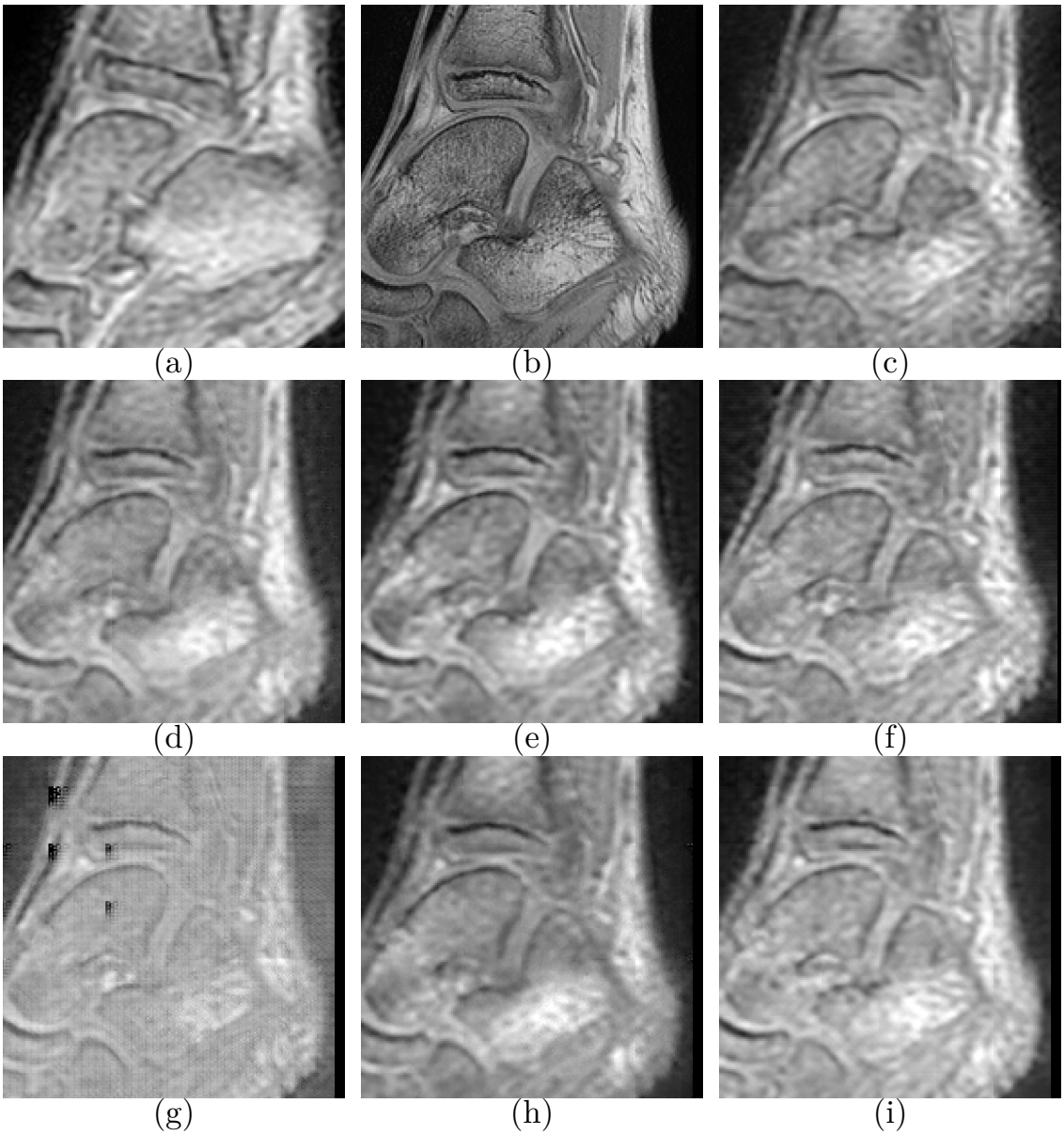


Figure 4.9 – Low-resolution static image synthesis results using CycleGAN and mDRIT++. (a) Dynamic image (b) Static image (c) Estimation with CycleGAN (d) Estimation with mDRIT++ - Conv (e) Estimation with mDRIT++ - AdaIN (f) Estimation with mDRIT++ - FiLM (g) Estimation with mDRIT++ - Conv with segmentation as an auxiliary task (h) Estimation with mDRIT++ - AdaIN with segmentation (i) Estimation with mDRIT++ - FiLM with segmentation.

	$\mathcal{L}_{seg}$	Entanglement	FID ↓	KID ↓	DSC ↑
DRL	✗	Conv	164.13	0.12	/
		AdaIN	195.5	0.17	/
		FiLM	367.5	0.44	/
DRL	✓	Conv	162.54	0.12	0.959 ± 0.003
		AdaIN	**134.8**	**0.076**	0.959 ± 0.004
		FiLM	201.4	0.15	0.963 ± 0.004
CycleGAN			154.81	0.10	/

Table 4.3 – Evaluation of the synthesis quality of high-resolution dynamic images for different entanglement modules, with and without segmentation, compared to CycleGAN. The evaluation is performed using two unpaired image metrics, KID and FID, and a DSC computed on synthesized image segmentations. The best performance is indicated in bold.

	$\mathcal{L}_{seg}$	Entanglement	FID ↓	KID ↓
DRL	✗	Conv	105.43	0.05
		AdaIN	**93.28**	**0.04**
		FiLM	101.67	0.05
DRL	✓	Conv	289.59	0.29
		AdaIN	119.06	0.07
		FiLM	94.88	0.044
CycleGAN			107.47	0.063

Table 4.4 – Evaluation of the synthesis quality of low-resolution static images different entanglement modules, with and without segmentation, compared to CycleGAN. The evaluation is performed using two unpaired image metrics, KID and FID. The best performance is indicated in bold.

and poor reconstruction of details. The Conv module still demonstrates its ability to synthesize textures realistically, but seems prone to artifacts. Figure 4.10 illustrates the presence of variable artifacts along the temporal sequence in the sequence reconstructed using the Conv module. In contrast, the method using the AdaIN module produces the most regular and realistic results throughout the sequence. Furthermore, unlike Conv, AdaIN does not amplify artifacts naturally found in the source images (see the fifth image in the sequence). CycleGAN provides a realistic texture synthesis but amplifies existing artifacts.

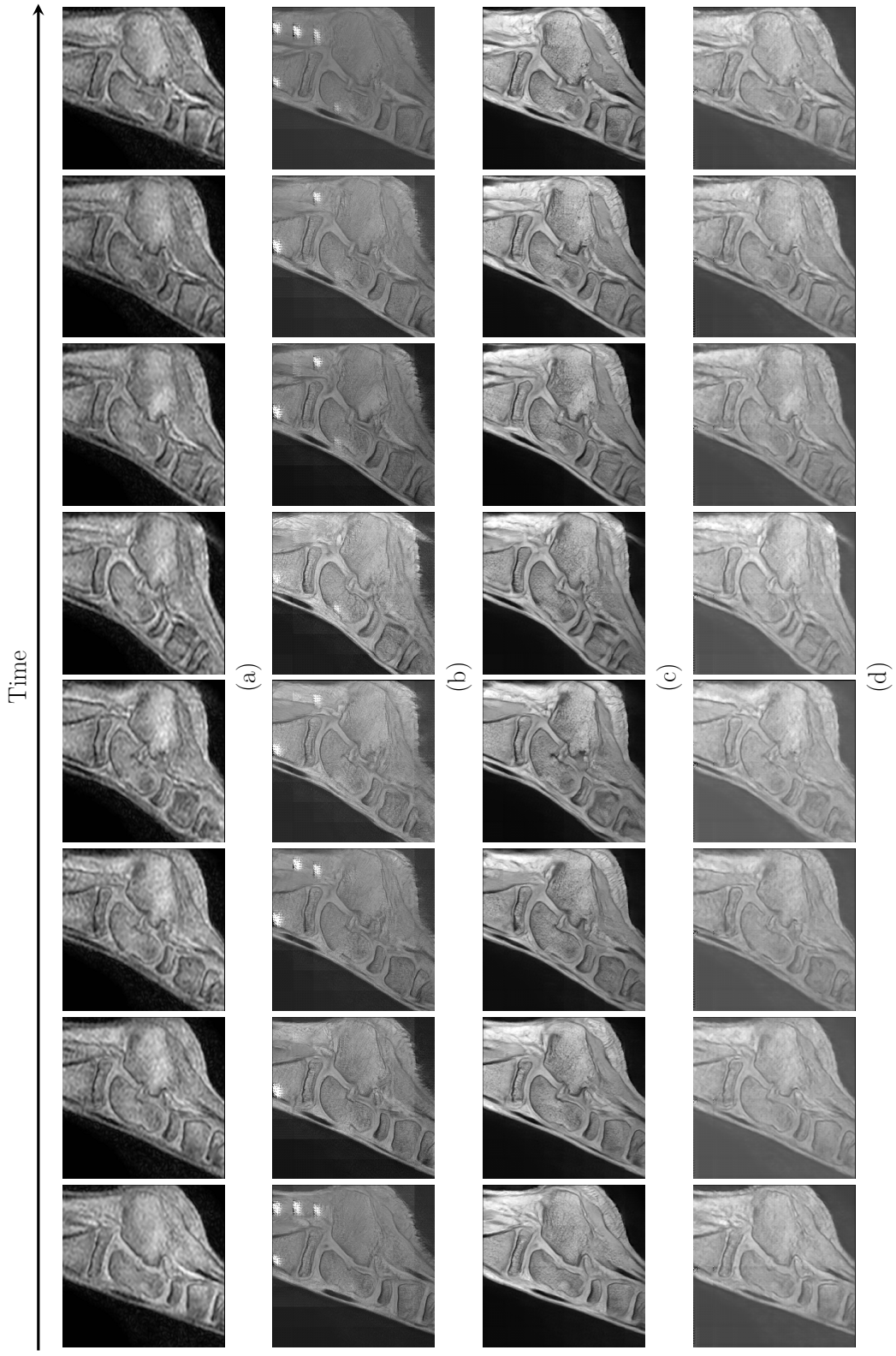


Figure 4.10 – High-resolution dynamic sequence synthesis for each entanglement module *with* segmentation as an auxiliary task. (a) Original dynamic sequence (b) Synthesis using Conv (c) Synthesis using AdaIN (d) Synthesis usng FiLM.

Figure 4.11 – High-resolution dynamic sequence synthesis for each entanglement module without segmentation as an auxiliary task. (a) Original dynamic sequence (b) Synthesis using Conv (c) Synthesis using AdaIN (d) Synthesis usng FiLM.

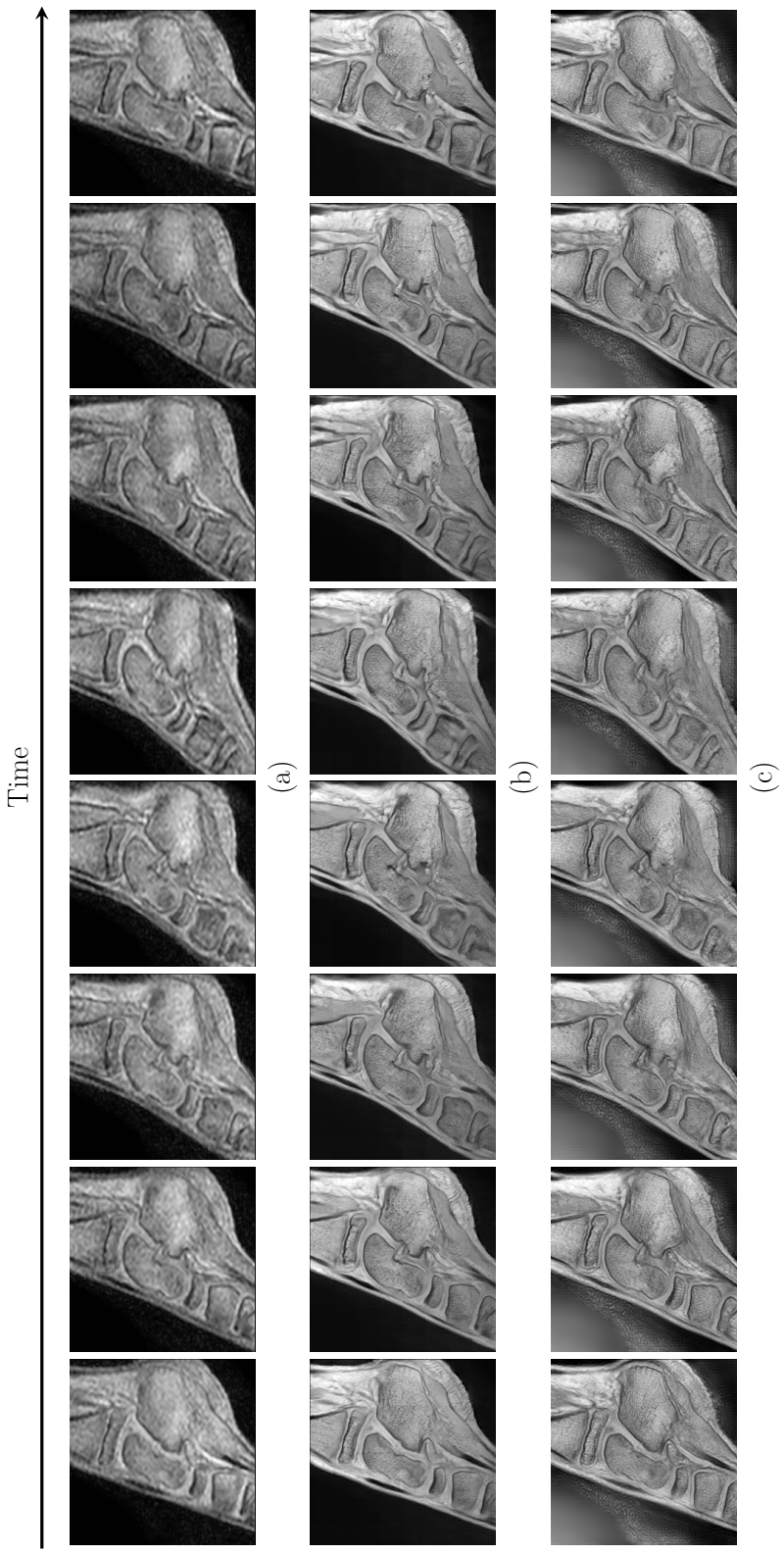


Figure 4.12 – High-resolution dynamic sequence synthesis for CycleGAN and the most successful DRL approach (with AdaIN and the segmentation as an auxiliary task). (a) Original dynamic sequence (b) Synthesis using AdaIN and segmentation (c) Synthesis using CycleGAN.

Anatomy conservation To demonstrate the framework's stability with respect to anatomy preservation, we synthesize three high-resolution dynamic images from a single dynamic source image and three different static images. The images are generated using the DRL method, which provides the optimal synthesis results, combining the AdaIN module and the segmentation task. Each synthesized image is generated from the anatomy representation of the dynamic image and the modality representation of the corresponding static image. Figure 4.13 illustrates the results obtained for each combination. Despite the differences between each of the static images, due to in-plane and out-of-plane rotations and translations, each of the synthesized images preserves the anatomical structure of the dynamic image and achieves similar textures. We observe that the mean image, derived from the three synthesized images, exhibits no blurring, indicating a high degree of similarity between each synthesized image. Using the first synthesized image as a reference, it is possible to evaluate the differences between the different images generated. Therefore, the average PSNR between each image is 32, and the average SSIM is 0.99. These observations indicate that the level of disentanglement between the representations learned by the model is sufficient for a perturbation in the anatomical representation of the static image to have no perceptible repercussions on the synthesized image.

4.4.4 Segmentation

This section presents the segmentation results obtained with each entanglement module. The segmentation accuracy is evaluated using the DSC. The quantitative results for the segmentation task are presented in Table 4.3. Despite the disparity in terms of synthesis quality, all the methods demonstrate similar segmentation performances, as shown by the DSC. The segmentation results are presented in Figure 4.14. The estimated segmentations are displayed as overlays on the source static MRI in the first row. The ground truth is indicated as a reference on the left. The second row shows the absolute difference between the ground truth and the estimated segmentations. It could be observed that the segmentation is globally similar across the different entanglement modules. The absolute error was found to be evenly distributed along bone contours, regardless of whether the bone edges were smooth or sharp.

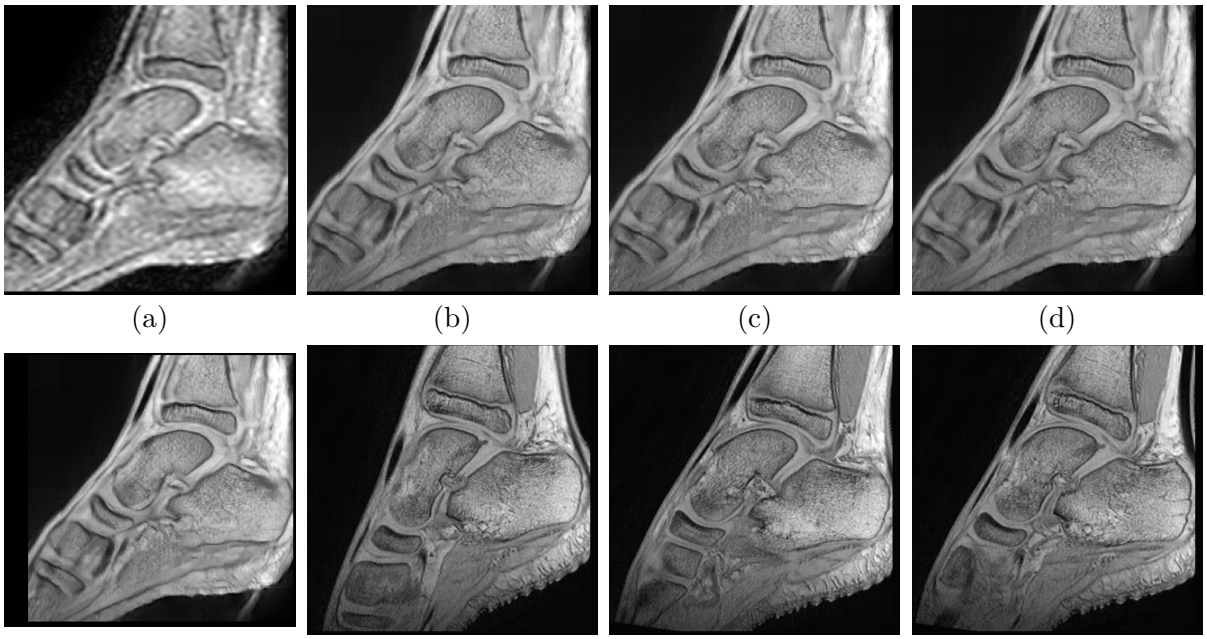


Figure 4.13 – Anatomy preservation in the DRL framework in an unpaired setting. A comparison is made between high-resolution dynamic images synthesized from one dynamic image and three different static images. All source images are from the same subject, and all the static images are identical, except for an affine transformation. (a) Dynamic source image, (b) Synthesized image from the dynamic and static S1 source images, (c) Synthesized image from the dynamic and static S2 source images, (d) Synthesized image from the dynamic and static S3 source images, (e) Mean image between the three synthesized dynamic images (f) Static source image S1, (g) Static source image S2, (h) Static source image S3.

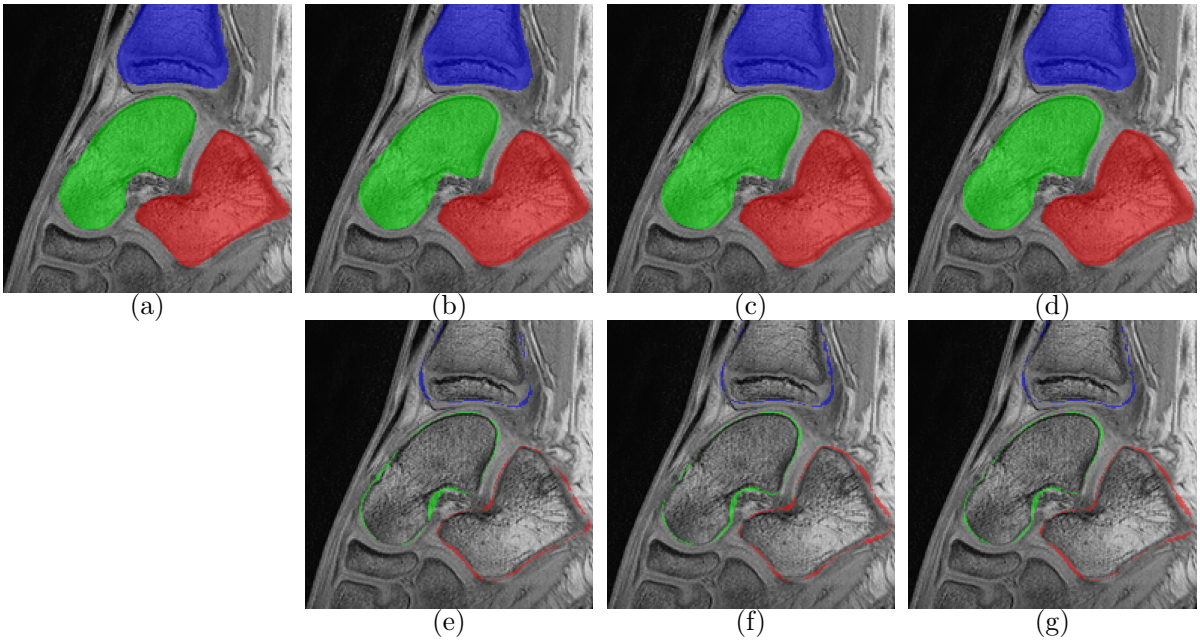


Figure 4.14 – Segmentation results on real data using mDRIT++. Each segmentation is displayed as overlay on the corresponding static image. (a) Ground truth segmentation (b) Estimation with mDRIT++ - Conv (c) Estimation with mDRIT++ - AdaIN (d) Estimation with mDRIT++ - FiLM (e) (f) and (g) correspond to the difference between the ground truth and the above result.

4.4.5 Uncertainty

Data augmentation is commonly used during the training process, though it can also be used during the test time. The purpose is to generate multiple variations of a single image and then combine the predictions made by the model after reversing the transformation. Typically used as a way to improve the prediction of neural networks during training, especially in segmentation tasks, it also provides a measure of the uncertainty of the model with respect to particular transformations. In this case, we use it to evaluate the uncertainty of the high-resolution dynamic image synthesis process.

The transformations are randomly sampled from a set of rotations and scaling transformations. The amplitude of scaling is set to 0.2, and the rotation amplitude is set to 40° on the sagittal plane. A total of ten images are used, including the original images and nine randomly transformed images. High-resolution synthesis is performed for all ten images, and the inverse transformation is then applied to restore the original configuration. Statistics are then computed based on the ten resulting images, with a focus on the mean and standard deviation images.

Figure 4.15 illustrates the results for each method. The first and second rows of this figure correspond to the mean image and standard deviation image, respectively. Each mean image and standard deviation image shares the same contrast dynamic. A blurred mean image indicates greater variability in the image synthesis, whereas a sharper one indicates greater stability of the reconstructions. The standard deviation image reflects the amount of variability present in a given pixel. First, we observe that all the mean images from the mDRIT++ methods seem to exhibit greater blurring than the CycleGAN one. However, while the mean image from the CycleGAN appears sharper than the others, the texture of the bones is quite homogeneous, and the standard deviation within that area is considerably greater than in the other methods. The standard deviation image of the CycleGAN methods demonstrates a global variability in the image, while the other methods show high standard deviation values concentrated around the bones' edges. Secondly, it can be observed that the texture of the bone is rendered more accurately by the mean images generated by the AdaIN methods (with and whithout segmentation) than by those produced by the other methods. This is particularly visible on the calcaneus. The enhanced detail observed in the mean images produced by the AdaIN methods suggests that these methods yield more robust results than the other methods. Additionally, we observe that the mean image of each method using the segmentation as an auxiliary task demonstrates significantly richer textures in terms of details, and a reduced variability,

according to the standard deviation images. These observations suggest that the addition of the segmentation reduces the variability among the synthesized images, producing a more robust model. The standard deviation images demonstrate the presence of artifacts, which are predominantly observed in the images generated by the methods that employ the Conv and FiLM modules.

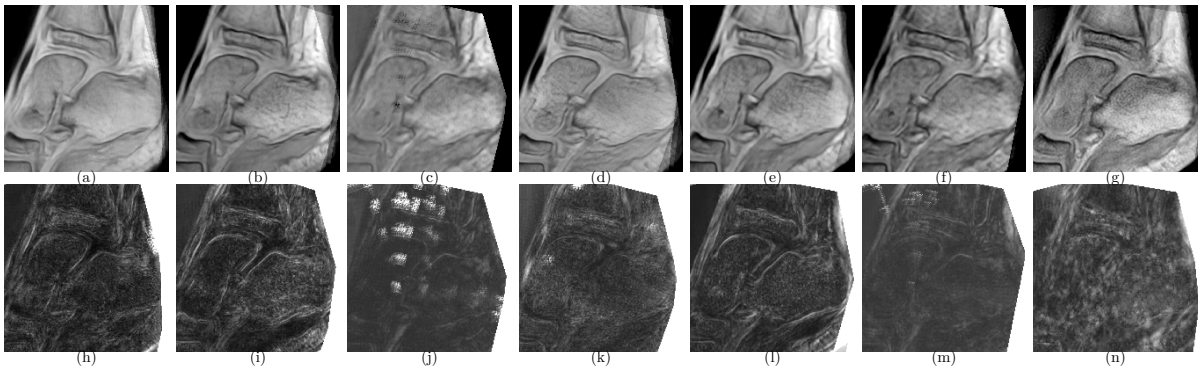


Figure 4.15 – Results of the test-time augmentation for n=10. The first row corresponds to the mean images and the second corresponds to the standard deviation image. (a)(h) mDRIT++ - Conv. (b)(i) mDRIT++ - AdaIN. (c)(j) mDRIT++ - FiLM. (d)(k) mDRIT++ - Conv with segmentation. (e)(l) mDRIT++ - AdaIN with seg. (f)(m) mDRIT++ - FiLM with seg. (g)(n) CycleGAN.

4.4.6 Evaluation of the disentanglement

Measuring the disentanglement This part provides an approach to evaluate the disentanglement of the anatomy and modality latent representations for each of the disentangled representation learning scenarios previously introduced. The evaluation uses the DC and IoB metrics introduced in Section 4.3.3 and the quantitative results are provided in tables 4.5 and 4.6. The disentanglement of the latent representations and their correlation with the source images were evaluated using both DC. DC was calculated between the two latent representations and between each representation and the corresponding source image. IoB was computed between each representation and the source image to evaluate their level of informativeness regarding the source image. The results for a dynamic image are presented in Table 4.5, and for a static image in Table 4.6.

The method using the AdaIN entanglement module consistently exhibits a higher correlation between the source image and the extracted anatomy representation for both

static and dynamic images. The DC between the image and the modality representation is globally even across all variations of the DRL framework and for both dynamic and static images. The segmentation module enhances the decorrelation between the representations of anatomy and modality for both static and dynamic images, suggesting a better disentanglement. The DC between the two representations is lower in static images, while both representations are more informative, indicating a greater amount of information carried by the latent representations of the static images and a better disentanglement.

	without segmentation			segmentation		
	Conv	AdaIN	FiLM	Conv	AdaIN	FiLM
$\mathrm{DC}(z_x^a, z_x^m) \downarrow$	**0.56 ± 0.03**	0.57 ± 0.03	**0.56 ± 0.02**	**0.52 ± 0.03**	0.56 ± 0.03	0.57 ± 0.02
$\mathrm{DC}(z_x^a, x) \uparrow$	0.77 ± 0.02	**0.85 ± 0.03**	0.81 ± 0.03	0.76 ± 0.02	**0.83 ± 0.02**	0.8 ± 0.03
$\mathrm{DC}(z_x^m, x) \uparrow$	**0.90 ± 0.01**	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.02
$\mathrm{IoB}(z_x^a, x) \uparrow$	1.33 ± 0.03	1.33 ± 0.02	1.33 ± 0.02	1.32 ± 0.03	1.33 ± 0.03	**1.34 ± 0.03**
$\mathrm{IoB}(z_x^m, x) \uparrow$	**1.08 ± 0.06**	1.06 ± 0.02	1.03 ± 0.05	1.05 ± 0.02	**1.06 ± 0.03**	1.1 ± 0.02

Table 4.5 – Evaluation of the disentanglement on a dynamic image. We use Distance Correlation (DC) and Information over Bias introduced in [243]. The best performance is indicated in bold.

	wo segmentation			segmentation		
	Conv	AdaIN	FiLM	Conv	AdaIN	FiLM
$\mathrm{DC}(z_y^a, z_y^m) \downarrow$	**0.48 ± 0.03**	0.58 ± 0.04	0.54 ± 0.01	**0.47 ± 0.03**	0.54 ± 0.04	0.56 ± 0.04
$\mathrm{DC}(z_y^a, y) \uparrow$	0.71 ± 0.04	**0.83 ± 0.05**	0.78 ± 0.04	0.71 ± 0.04	**0.8 ± 0.05**	0.78 ± 0.05
$\mathrm{DC}(z_y^m, y) \uparrow$	**0.91 ± 0.03**	0.9 ± 0.03	0.77 ± 0.06	0.89 ± 0.03	0.9 ± 0.03	**0.92 ± 0.03**
$\mathrm{IoB}(z_y^a, y) \uparrow$	1.36 ± 0.04	**1.37 ± 0.07**	1.35 ± 0.05	**1.35 ± 0.04**	1.34 ± 0.06	**1.35 ± 0.05**
$\mathrm{IoB}(z_y^m, y) \uparrow$	1.12 ± 0.05	**1.13 ± 0.05**	**1.13 ± 0.04**	**1.12 ± 0.04**	1.08 ± 0.03	1.09 ± 0.04

Table 4.6 – Evaluation of the disentanglement on a static image. We use Distance Correlation (DC) and Information over Bias introduced in [243]. The best performance is indicated in bold.

Latent space interpretation In order to gain a deeper understanding of the impact of architectural features on the latent representations, it would be beneficial to observe the latent representations themselves. The latent spaces are of high dimensions, with 8 and 262144, respectively, and thus cannot be visualized in their original form. The Uniform Manifold Approximation and Projection (UMAP) library [245] is a dimension reduction technique (linear and non-linear) based on Riemannian geometry and algebraic topology. The UMAP algorithm is based on the manipulation of weighted graphs, placing it within the class of k-neighbor based graph learning algorithms. The UMAP algorithm initially generates a high-dimensional graph representation of the data, and then optimizes a low-dimensional graph to maximize its structural similarity with the high-dimensional graph.

This algorithm is applied to both modality and anatomy representations, resulting in a mapping of both in a two-dimensional space. Each data point corresponds to a particular patch of the source image.

Figures 4.16 and 4.17 provide visualizations of the latent representations generated by UMAP. The latent representations for both real and fake images for each MR sequence are displayed for each DRL method. The color map is adapted according to the latent representation under consideration, with the objective of highlighting representations that are presumed to be similar. For the anatomy representation, the data points corresponding to the dynamic MR images y are represented in blue, and similarly for those corresponding to the fake static MR image y_X. It is assumed that both images share a common anatomy representation. Similarly, the static and fake dynamic images x and x_Y are displayed in red and pink, respectively. Equations 4.18 and 4.19 serve as a reminder of the disentanglement and generation processes. Equation 4.18 describes the disentanglement of the source image, while Equation 4.19 describes te generation process of the fake images, translated in the other modality.

$$\begin{cases} x = (z_x^a, z_x^m) \\ y = (z_y^a, z_y^m) \end{cases} \tag{4.18}$$

$$\begin{cases} x_Y = G_Y(z_x^a, z_y^m) \\ y_X = G_X(z_y^a, z_x^m) \end{cases} \tag{4.19}$$

In order to quantitatively characterize the differences and similarities between the anatomy and modality representations, and in complement to Figures 4.16 and 4.17, we compute the MSE between corresponding latent representations. This study is conducted for both anatomy and modality latent representations. Table 4.7 displays the results obtained for each latent representation. For each representation, three different parameters are compared. First, for each patch, the distance between the representation (anatomy or modality) of the patch in the static image and the same patch in the dynamic image is evaluated. It should be noted that since the images from a single subject are partially registered, a similar anatomical content may be found on corresponding latent patches at the same location on the static and dynamic MR images. However, since the registration is relatively coarse, it is not possible to assume that the anatomical content between corresponding static and dynamic images patches is identical for all images. Consequently, the comparison between static and dynamic patches is inherently unfair. The other two comparisons are made between the representation of source image patches and the corre-

sponding representation in fake images, which are assumed to be similar. The objective is to assess how well the fake images preserve the representations they are constructed from. In the case of anatomy representation, we compare the anatomy representation between x and x_Y, and between y and y_X. It is assumed that these two sets of images share a common anatomy representation (see Equations 4.18 and 4.19). In the case of modality representation, the comparison is conducted between the modality representation of x and y_X, and between y and x_Y. All the MSE calculations are performed on the low-dimensional latent space in order to preserve maximal coherence between the measures and the corresponding figures.

Figure 4.16 illustrates the anatomy representations for each DRL method. For each method, the low-dimensional graph is constructed based on anatomical representations of both real and fake images for each modality: z^a_x, z^a_y, $z^a_{x_Y}$ and $z^a_{y_X}$. Optimizing the low-dimensional graph using all these data enables the comparison of all these representations according to a common reference frame. In accordance with the model definition, the two anatomy representations derived from the source image x (resp. y) and the corresponding image translated in the other modality x_Y (resp. y_X) are expected to be similar, since both images share a common anatomical content. Moreover, the source images x and y are expected to share similar anatomy representations, given that both images belong to the same subject.

The method using the FiLM module without the segmentation task demonstrates a significant discrepancy between the anatomy representations of dynamic and fake static images, in contrast with the other methods. As illustrated in Figure 4.8, the fake static image corresponding to this method exhibits the worst results among all the methods, and fails to achieve a realistic synthesis of textures. In a reduced way, the difference between z^a_x and $z^a_{x_Y}$ can also be observed for the method that includes both the FiLM module and the segmentation task. This method also demonstrates lower synthesis performance compared to the other methods (see Table 4.3 for the quantitative evaluation of each method). The other methods, however, exhibiting higher synthesis performances, all demonstrate a visually higher proximity between the anatomy representations of the source images and the corresponding synthesized images. These observations are corroborated by Table 4.7. These results suggest that a higher similarity between latent anatomy representations of real images and corresponding fake images leads to better synthesis performance. These results are consistent with those presented in Section 4.4.2, where the addition of a cycle-consistency constraint on the anatomy latent space was shown to improve the

synthesis performance of the model. We observe that the anatomy representations of the methods using the segmentation module are characterized by a higher overlap than those which are not using the segmentation module. This observation suggests that the segmentation module promotes a common representation of the anatomy between the dynamic and static MRI sequences. A quantitative evaluation of the similarity between corresponding anatomy representations is provided by Table 4.7. Although the full point cloud analysis is more relevant for the comparison between the anatomy representations of static and dynamic images due to the coarse registration between image pairs, we also observe in Table 4.7 that the average error is relatively low in comparison to the extent of the point cloud. This tendency is further accentuated when the segmentation task is added. These results demonstrate that static and dynamic images from the same subject share similar anatomical representations. The AdaIN module, when combined with the segmentation module, exhibits the lowest error between the anatomical representations of the source images and those of the corresponding synthesized images, for both types of MRI sequences. As previously stated, the FiLM module, in the absence of the segmentation module (and to a lesser extent, with the segmentation module), exhibits the greatest discrepancy between the anatomical representation of the source image and that of the corresponding synthesized image, particularly between dynamic and fake static images. These observations demonstrate a failure to preserve, in the synthesized image, the anatomy representation from which it was constructed. These errors are also present, to a lesser extent, for the Conv module with segmentation and AdaIN module without segmentation task, between the anatomical representations of static images and those of fake dynamic images. The larger distance between the anatomy representations of the static and dynamic images compared to the other measurements is due to the unpaired patches.

Figure 4.17 illustrates the modality representations for each DRL method. As for the anatomy representation, the low-dimensional graph is optimized from modality representations of both real and fake images of both types of MR sequences: z_x^m, z_y^m, $z_{x_Y}^m$ and $z_{y_X}^m$. According to the model definition, it is expected that the modality representations of x and y (resp. x_Y and y_X) will be separable from each other. Moreover, it is expected that z_x^m and $z_{y_X}^m$ (resp. z_y^m and $z_{x_Y}^m$) will be similar, as each pair is supposed to share a common modality representation. The color map of the figure has been adapted to reflect these similarities. Consequently, the modality representations of x and y_X are represented in red and pink, while those of y and x_Y are displayed in blue.

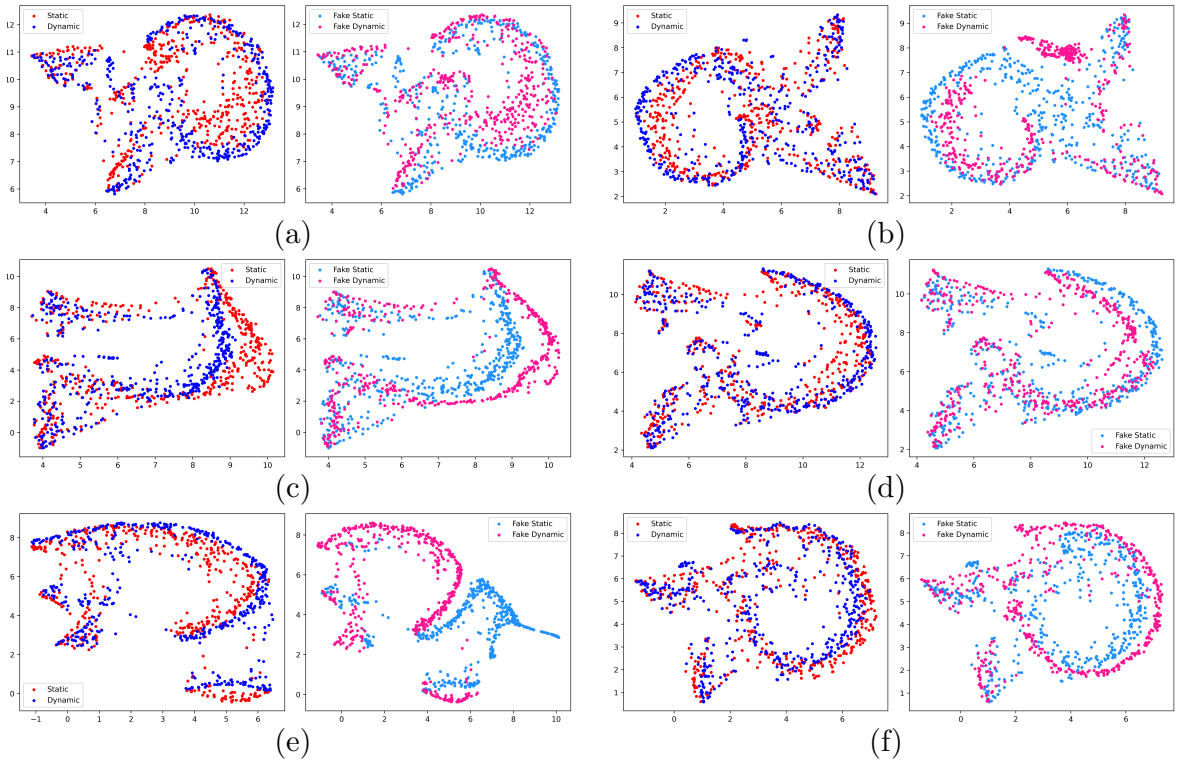


Figure 4.16 – Two-dimensional UMAP visualization of the anatomy latent codes for each DRL method. For each method, the latent representations for both real and fake images are examined. (a) Conv module without segmentation, (b) Conv module with segmentation task, (c) AdaIN module without segmentation, (d) AdaIN module with segmentation task, (e) FiLM module without segmentation, and (f) FiLM module with segmentation task.

As anticipated, the modality representations corresponding to the two types of MRI sequences constitute two distinct and linearly separable clusters for all methods. As with the anatomy representation, the method combining the FiLM module with the segmentation task, which produced the poorest synthesis performance compared with the other methods studied, fails to reconstruct the modality representation of the source image in the fake images. This reconstruction error can be observed for both types of MRI sequences, in contrast to its anatomical representations, where the phenomenon was circumscribed to a single source modality. In general, and for all methods, the reconstruction of modality representation in fake images compared to real images is less efficient than that observed for anatomical representation. These observations are corroborated by Table 4.7. For methods employing the Conv and AdaIN modules, this trend is further accentuated with the incorporation of the segmentation module, which yields a higher modality representation reconstruction error than methods devoid of a segmentation module. However, Table 4.7 shows that the segmentation module significantly increases, for methods using the FiLM and AdaIN modules, and moderately, for the method using the Conv module, the discrepancy between modality representations of images from static and dynamic sequences. Finally, Figure 4.17 shows that modality representations are inhomogeneous clusters for all methods. Consequently, it can be assumed that the modality representation is not homogeneous within the image. In order to improve understanding of latent representations, the following paragraph proposes to relate the data points of latent representations to the corresponding patches on source images.

In order to enhance the comprehension of the latent representation spaces, we provide a correspondence between the anatomy representations and the corresponding image patch. This correspondence enables the visualization of the construction of the latent space as a function of the input image. We use the latent representations of the method that employs the AdaIN module in conjunction with the segmentation task, which provides the optimal synthesis results among the methods that were evaluated. Figure 4.18 shows the image patches corresponding to the anatomical representations. We have chosen to show the corresponding patches only on dynamic images, as the distribution of static patches is very similar. A continuous distribution of representations can be seen on the right-hand side of the latent space representation. As we progress along this structure, a continuous progression is observed along the foot. Images with a higher proportion of background are predominantly grouped in the upper part of the structure, while those with a high proportion of tissue are grouped towards the lower part. The transition between the two

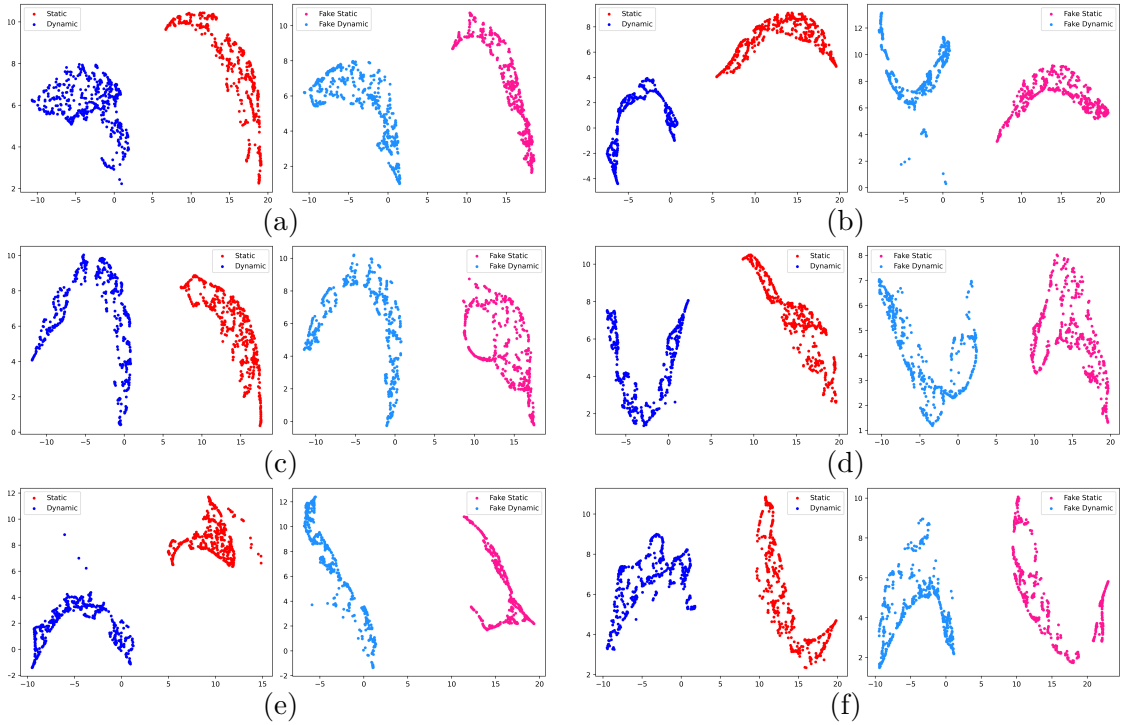


Figure 4.17 – Two-dimensional UMAP visualization of the modality latent codes for each DRL method. For each method, the latent representations for both real and fake images are examined. (a) Conv module without segmentation, (b) Conv module with segmentation task, (c) AdaIN module without segmentation, (d) AdaIN module with segmentation task, (e) FiLM module without segmentation, and (f) FiLM module with segmentation task.

	wo segmentation			segmentation		
	Conv	AdaIN	FiLM	Conv	AdaIN	FiLM
MSE$(z_x^a, z_y^a) \downarrow$	0.51	0.32	1.05	0.26	0.46	0.38
MSE$(z_x^a, z_{x_Y}^a) \downarrow$	0.011	0.015	12.49	0.041	0.0039	0.05
MSE$(z_y^a, z_{y_X}^a) \downarrow$	0.032	0.14	0.046	0.78	0.01	0.31
MSE$(z_x^m, z_y^m) \uparrow$	159.57	160.05	122.02	161.45	221.96	160.28
MSE$(z_x^m, z_{y_X}^m) \downarrow$	1.58	0.72	24.48	43.91	2.64	3.70
MSE$(z_y^m, z_{x_Y}^m) \downarrow$	1.81	1.84	28.94	2.15	4.77	5.21

Table 4.7 – Quantification of the similarity between corresponding latent representations. To this end, we employ the MSE distance on the UMAP projections to compare the latent representations. For the anatomy latent representation, we compare representations from static and dynamic images, as well as those from dynamic and fake static images and static and fake dynamic images. The first pair of images is presumed to demonstrate similar representations in the latent space, given that each pair of static and dynamic images is from the same subject. In addition, the two latter representations are presumed to be nearly identical, given that the dynamic image (resp. static) is assumed to share the same anatomy as the fake static images (resp. fake dynamic). The second section of the table is dedicated to the modality latent representation. We compare the latent representations from static and dynamic MR images, which are presumed to be disparate, and those from static and fake static MR images and dynamic and fake dynamic images, which are presumed to be nearly identical.

extremes is continuous along the structure. Similar or closely related areas of the foot are grouped together in the same portion of the structure. The latent space representation also shows two clusters of data points. The two clusters exhibit a less clear interpretation than the first structure and do not appear to be characteristic of any particular anatomical structure or image.

Figure 4.19 shows the patches from the source images corresponding to the different modality representations. It can be observed that the representations for each type of MRI sequence appear to be partly dependent on the position of the patch in the image. Indeed, patches centered on the foot are typically grouped together in one part of the cluster, while those centered on the junction between the foot and the background are grouped together at another end of the cluster. The sets of data points corresponding to the junction between the foot and the background also correspond to the location where the two clusters of each modality are closest. This proximity between the two clusters is consistent with the significant presence of the background, which is the same for both modalities, without the characteristic texture of dynamic images. As with the anatomy representations, the data points are grouped by anatomical proximity, but in a less structured pattern. This similar grouping may be due to incomplete disentanglement, which allows anatomy to infuse into the modality representation. Alternatively, it may be a logical consequence of the modality representation being similar for close anatomical

areas. However, if anatomy infused information into modality, the synthesis of two images from the same content of a dynamic image and the modality of two different static images would produce different results. Nevertheless, as shown in section 4.4.3, the synthesis of images from the same anatomy and different modality representations corresponding to two static images in different spatial positions produces highly similar results.

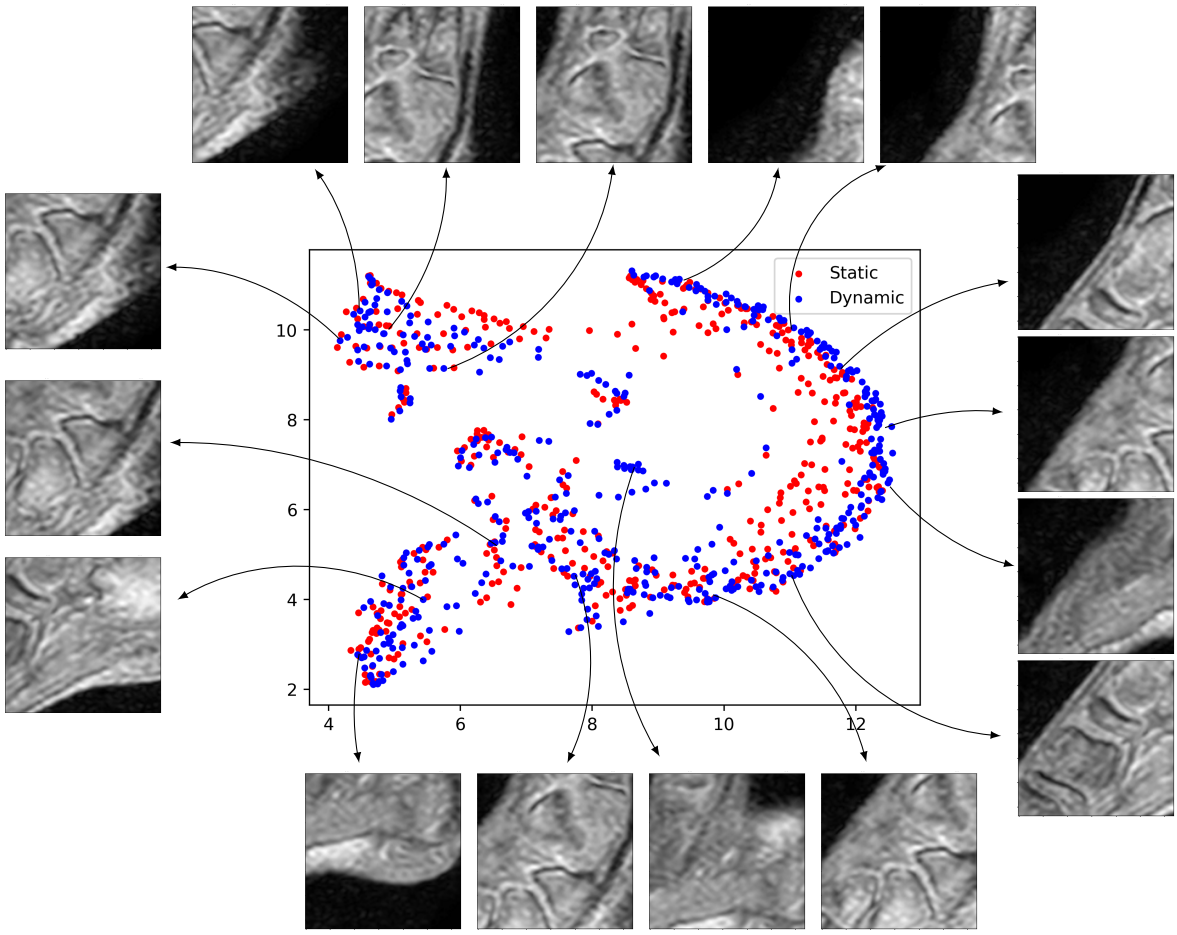


Figure 4.18 – Anatomy latent representation *versus* corresponding image patch.

4.5 Discussion

Three types of constraints on latent spaces are studied. Experiments show that the addition of a cyclic constraint on the anatomy representation improves synthesis quality.

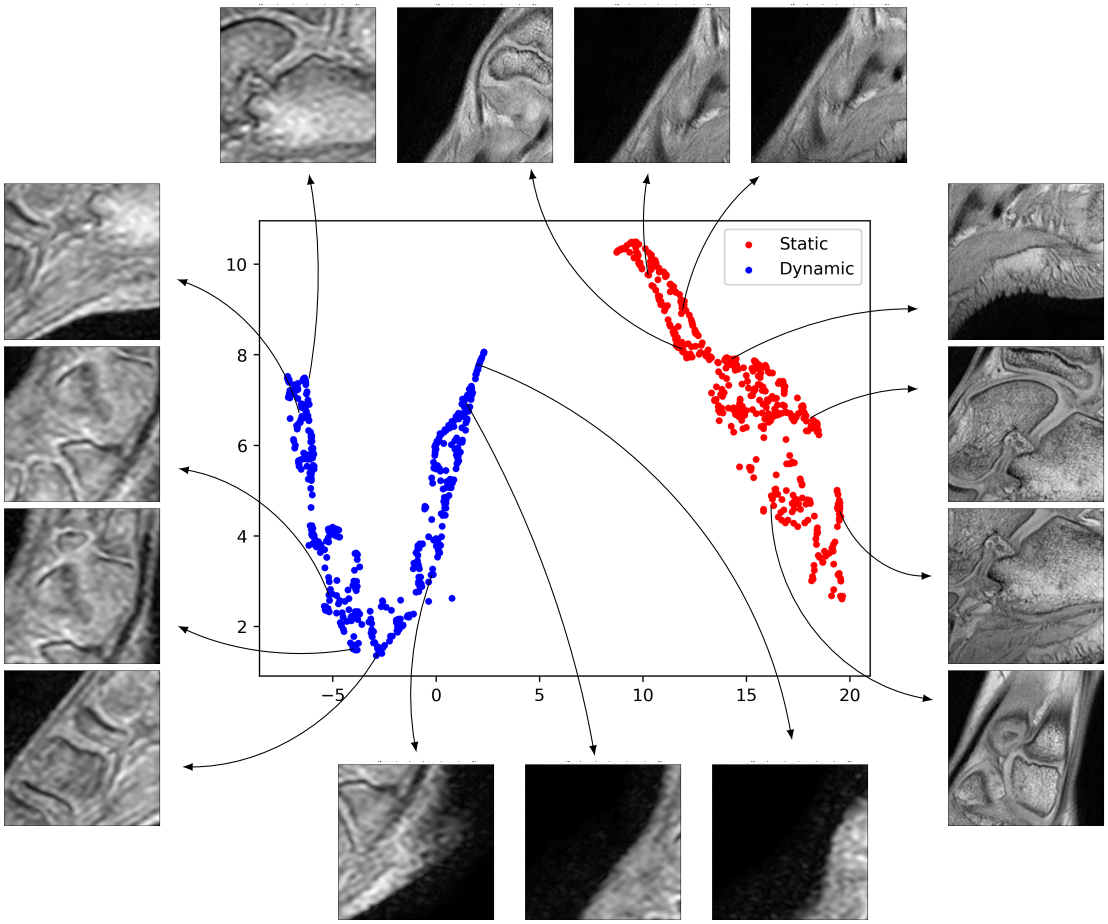


Figure 4.19 – Modality latent representation *versus* corresponding image patch.

Three entanglement modules are compared, and the impact of an auxiliary segmentation module is investigated. The experiments demonstrate that the addition of an auxiliary segmentation task significantly improves synthesis quality of high-resolution dynamic images, improves the robustness of the model to spatial transformations, and leads to a better disentanglement between the latent representations. While both the Conv and the AdaIN-based DRL framework produce visually realistic images, the method using the AdaIN module appears to be less prone to artifacts and produces more regular edges. If the methods using the AdaIN module exhibit a systematic improvement in the correlation between the anatomy latent representation and the source image, the Conv-based methods consistently demonstrate higher disentanglement between the two latent representations. However, we show the stability of the model regarding spatial affine perturbation in the image from the target modality. The CycleGAN framework generates a visually realistic texture of the bones, but appears to be more susceptible to irregular anatomical structures and artifacts. Furthermore, it has been demonstrated to be less robust than the DRL approaches. The proposed DRL approaches are capable of generating high-resolution 3D+t dynamic sequences. The optimal results are achieved when the AdaIN module is coupled with the segmentation task.

4.6 Conclusion

This chapter explores the use of disentangled representations for the unpaired synthesis of high-resolution dynamic MRI. Disentangled representations are a class of methods based on the assumption that a data distribution can be described by a set of independent, semantically meaningful factors. The objective of disentangled representations is to identify these factors and encode them in distinct dimensions.

This work makes two main contributions. First, we propose the addition of constraints on latent spaces to improve the quality of synthesis. Secondly, we propose to study the influence of the entanglement module on synthesis performance. To our knowledge, no study has investigated its influence on disentangled representations. We are also investigating the impact of adding a segmentation module from the latent anatomy representation on synthesis results. In this second study, we also propose an evaluation of disentanglement and an analysis of latent spaces. Our results demonstrate that the entanglement modulus has a significant impact on the performance of a DRL method, and notably on the disentanglement rate of the method.

This study demonstrates the relevance of using DRL methods for unpaired I2I synthesis applied to dynamic MRI. By applying the proposed method to a dynamic sequence, we succeeded in generating a high-resolution 3D+t dynamic sequence. Furthermore, this method offers comparable or even superior performance to CycleGAN, the state-of-the-art in unpaired I2I synthesis methods, as well as superior stability with respect to affine transformations. Furthermore, this study illustrates the potential of using DRL methods for unpaired image-to-image synthesis tasks, even in scenarios where the dataset is limited.

The contributions presented in this chapter have resulted in two publications: a journal article [246] and a conference proceeding [247].

Conclusion

Conclusion

Dynamic MR imaging is a medical imaging modality used to observe physiological dynamics *in vivo*. This imaging modality, which is widely employed to investigate cardiac muscle or joint biomechanics, is non-ionizing and non-invasive, providing excellent visualization of anatomical structures. However, the acquisition of high-resolution dynamic MRI sequences is a time-consuming process that requires repeated movements at a constant speed. To reduce the strain on patients with musculoskeletal disorders, the acquisition time of the sequences can be reduced at the cost of spatial resolution. This thesis examines the potential of deep learning methods for the synthesis of high-resolution dynamic MRI sequences.

One of the most significant challenges in this work lies in the pairing of data. Indeed, the data acquired as part of the Equinus project are subject to non-rigid deformations between different images of the same subject, resulting in unaligned static and dynamic MR images. These deformations are a consequence of the dynamic nature of the study and the movements required of the subjects. These data guided the choice of approaches studied for the synthesis of high-resolution dynamic MRI sequences.

The scientific questions that arise from this study are as follows: 1) Can paired I2I synthesis methods be used to synthesize high-resolution dynamic MRI sequences? 2) Do unpaired I2I methods offer superior performance? 3) What is the impact of the entanglement module and the incorporation of an auxiliary segmentation task on the learning of disentangled representations and the quality of synthesis? 4) Does the introduction of constraints in the latent space of a disentangled representation model affect the reconstruction quality?

A first approach to high-resolution dynamic sequence synthesis was conducted to investigate the potential of I2I methods employing paired data. We first describe two distinct approaches for data pairing: the first based on a registration process and the second based on data simulation. Using these data, we explore several architectures for the synthesis of high-resolution dynamic MRI sequences. After validating the proposed approaches on

a paired medical imaging dataset, we demonstrate the inability of these same methods to synthesize visually realistic high-resolution dynamic images.

The second approach investigated focused on I2I synthesis methods using unpaired data for high-resolution dynamic MRI synthesis. In particular, the research concentrated on methods based on disentangled representations, which offer greater control over the synthesis result than approaches based on GANs, such as CycleGAN. The results demonstrated that this approach provides a realistic synthesis of a high-resolution dynamic MR image. The method was applied to a dynamic MRI sequence, resulting in the successful synthesis of a high-resolution 3D+t dynamic MRI sequence. Furthermore, the approach allows for the simulation of images from dynamic sequences from static images.

Building upon this approach, we investigate the impact of various parameters on the learning of disentangled representations and synthesis quality. In particular, we examine the influence of the entanglement module and the integration of an auxiliary segmentation task from an anatomy representation. The entanglement module is a critical element in the learning of disentangled representations, enabling the merging of different representations within the model. Our analysis is informed by an evaluation of disentanglement for all these methods. Our results show that these two parameters have a significant impact on both the quality of the synthesis and the disentanglement rate of the method. Furthermore, we demonstrate that the introduction of constraints on latent spaces can influence the synthesis process and the quality of the images generated.

The method employed in this study demonstrates comparable or even superior performance to CycleGAN, a state-of-the-art method, as well as enhanced stability with respect to affine transformations. Furthermore, the method is shown to preserve anatomical content within the framework, a crucial element in medical imaging.

Perspectives

The presented works have demonstrated the relevance of using disentangled representations learning for the synthesis of high-resolution 3D+t dynamic MRI sequences. We propose a few potential avenues for improvement, and new applications.

Method

Super-resolution out of the sagittal plane

The approaches presented in this manuscript focus on super-resolution within the sagittal plane. The strong anisotropy within dynamic images limits both the accuracy and robustness of volumetric image processing, and offers only a very limited number of sagittal slices for joint motion analysis.

A possible avenue of development would be to integrate super-resolution for the third spatial axis. While the resolution within the sagittal plane is initially 0.57 mm for each axis, it is 8 mm for the orthogonal axis (in comparison, the resolution in the sagittal plane is 0.26 mm in each direction and 0.5 mm out of plane for static images). This strong anisotropy within dynamic images significantly increases the complexity of the task, due to the large number of slices that need to be estimated.

The implementation of super-resolution in the orthogonal axis would facilitate the development of 3D-based approaches, thereby promoting homogeneity between the different slices.

Pose integration

Previous studies did not consider the dynamic aspect of MRI data in the synthesis model. Instead, the different temporal frames are processed independently, which may result in visual inconsistencies between successive frames. This section presents an approach to leverage the dynamic information of MRI sequences within the context of disentangled representation learning.

This concept is inspired by the work of Denton and Birodkar on the learning of disentangled representations from videos [248]. In order to capture the dynamic aspect of the data, an additional latent representation is introduced, denoted z^p.

This approach is first presented in a monomodal setup, i.e., without including the two modalities and thus the modality representation. Two representations are considered: the anatomy z^a and the pose z^p. In a subsequent stage, this framework may be extended to encompass three representations, with the addition of the modality representation z^m. Following the work from Denton and Birodkar, three objective functions are employed to guide the pose extraction. First, a temporal reconstruction loss enforces the correct extraction of the temporal-invariant anatomy representation and the temporal-dependent pose representation. Given $\{x^0, ..., x^T\}$ an MRI sequence of $T + 1$ times frames, the

disentanglement process is described by the equation 4.20.

$$x^t = (z_t^a, z_t^p) = (E^a(x), E^p(x)) \tag{4.20}$$

where z_t^p describes the pose representation from the t time frame and z_t^a the anatomy representation. E^p and E^a stand for the pose and anatomy encoders. The temporal reconstruction loss between two time frames t and t' is expressed by the following equation:

$$\mathcal{L}^p_{reconstruction} = ||x^{t'} - G(z_t^a, z_{t'}^p)||_2^2 \tag{4.21}$$

where G is the generator that takes as input the pose and the anatomy representations and outputs the corresponding image. This temporal reconstruction loss assumes that the image at a time t' could be described by the anatomy at a time t and the pose representation of the time t', enforcing the anatomy representation to convey only the pose-invariant information. The second objective function penalizes errors between anatomy representations from different time frames, to enforce its temporal invariance:

$$\mathcal{L}^p_{sim} = ||z_t^a - z_{t'}^a||_2^2 \tag{4.22}$$

Assuming that the pose representation should be invariant between two different sequences, unlike the anatomy, a discriminator D_p is introduced in the framework. This discriminator attempts to identify pose latent representations from being from the same dynamic sequence or not. This adversarial constraint $\mathcal{L}^p_{adv}$ on the pose enforces the disentanglement of the pose and anatomy by penalizing the pose representations carrying anatomy-specific information.

Limitations The temporal reconstruction loss assumption may be overly strong for dynamic data, given the presence of noise between successive frames. To demonstrate the efficacy of the concept, it is possible to conduct experiments on two datasets: a real one and a simulated one. A simulated one can be generated from a set of static MR images for which transformations can be applied to simulate a dynamic movement. In order to simulate the most complete transformation possible, random rotation and elastic deformation can be employed (see Figure 4.20).

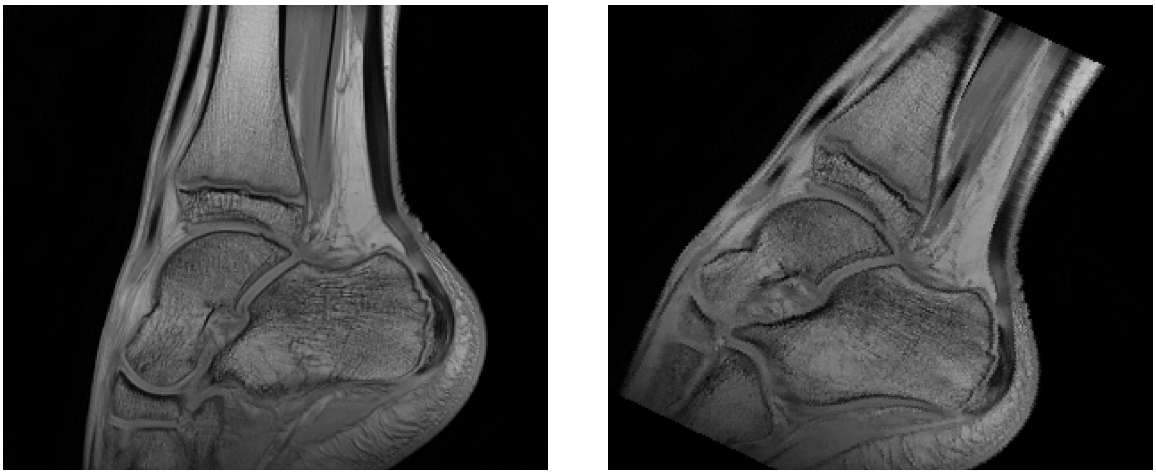


Figure 4.20 – Simulation of the dynamic movement of the ankle joint. The movement was simulated using a random rotation in the sagittal plane, with a range of $-30°$ to $30°$.

Applications

The method developed in this thesis aimed to enable the synthesis of high-resolution dynamic MRI sequences to improve the management of equinus in pediatric patients. The models developed in this thesis demonstrated their capacity to generate realistic high-resolution dynamic MRI sequences. One potential avenue for future research would be to apply this method and study biomechanical parameters that affect equinus, such as tendons or contact surfaces between the bones of the ankle joint.

Finally, the method developed in this thesis has demonstrated its ability to generate realistic results from reduced, unpaired datasets. These properties are particularly advantageous for applications in medical imaging, where datasets are frequently characterized by limited size and unpaired data. Consequently, this methodology may be applicable to other medical imaging datasets in different applications.

COMMUNICATIONS

Journal

— Scavinner-Dorval, C., Bailly, R., Borotikar, B., Brochard, S., Ben Salem, D., Rousseau, F. (2025). Learning disentangled representations for unpaired synthesis of high-resolution dynamic MRI. *Machine Learning for Biomedical Imaging*, 3(February 2025 issue), 16-37.

International conference

— Scavinner-Dorval, C., Bailly, R., Borotikar, B., Brochard, S., Ben Salem, D., Rousseau, F. (2024, April). Analysis of disentangled representation learning for high-resolution dynamic MRI synthesis. In *Medical Imaging 2024: Image Processing* (Vol. 12926, pp. 694-699). SPIE.

National conference

— Scavinner-Dorval, C., Bailly, R., Borotikar, B., Brochard, S., Ben Salem, D., Rousseau, F. *Image Synthesis for Dynamic MRI.* "Recherche en Imagerie et Technologies pour la Santé" symposium, 2022, in Brest - Oral

BIBLIOGRAPHY

[1] H. H. Banks and W. T. Green, « The Correction of Equinus Deformity in Cerebral Palsy », en-US, *JBJS*, vol. 40, *6*, p. 1359, Dec. 1958, ISSN: 0021-9355. Accessed: Jun. 22, 2023. [Online]. Available: https://journals.lww.com/jbjsjournal/Abstract/1958/40060/The_Correction_of_Equinus_Deformity_in_Cerebral.13.aspx.

[2] M. Krupiński, A. Borowski, and M. Synder, « Long Term Follow-up of Subcutaneous Achilles Tendon Lengthening in the Treatment of Spastic Equinus Foot in Patients with Cerebral Palsy », eng, *Ortopedia, traumatologia, rehabilitacja*, vol. 17, *2*, pp. 155–161, Mar. 2015, ISSN: 2084-4336. DOI: 10.5604/15093492.1157092. Accessed: Aug. 16, 2023. [Online]. Available: https://doi.org/10.5604/15093492.1157092.

[3] G. B. Firth et al., « Multilevel Surgery for Equinus Gait in Children with Spastic Diplegic Cerebral Palsy: Medium-Term Follow-up with Gait Analysis », en-US, *JBJS*, vol. 95, *10*, p. 931, May 2013, ISSN: 0021-9355. DOI: 10.2106/JBJS.K.01542. Accessed: Aug. 16, 2023. [Online]. Available: https://journals.lww.com/jbjsjournal/abstract/2013/05150/multilevel_surgery_for_equinus_gait_in_children.10.aspx.

[4] S. Y. Joo, D. N. Knowtharapu, K. J. Rogers, L. Holmes, and F. Miller, « Recurrence after surgery for equinus foot deformity in children with cerebral palsy: Assessment of predisposing factors for recurrence in a long-term follow-up study », en, *Journal of Children's Orthopaedics*, vol. 5, *4*, pp. 289–296, Aug. 2011, Publisher: SAGE Publications, ISSN: 1863-2521. DOI: 10.1007/s11832-011-0352-4. Accessed: Aug. 16, 2023. [Online]. Available: https://doi.org/10.1007/s11832-011-0352-4.

[5] C. Y. Chung et al., « Recurrence of Equinus Foot Deformity After Tendo-Achilles Lengthening in Patients With Cerebral Palsy », en-US, *Journal of Pediatric Orthopaedics*, vol. 35, *4*, p. 419, Jun. 2015, ISSN: 0271-6798. DOI: 10.1097/BPO.0000000000000278. Accessed: Aug. 16, 2023. [Online]. Available: https://journals.

lww.com/pedorthopaedics/abstract/2015/06000/recurrence_of_equinus_foot_deformity_after.20.aspx.

[6] S. K. Zhou et al., « 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. DOI: 10.1109/JPROC.2021.3054390.

[7] H. K. Graham et al., « Musculoskeletal Pathology in Cerebral Palsy: A Classification System and Reliability Study », en, *Children*, vol. 8, *3*, p. 252, Mar. 2021, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2227-9067. DOI: 10.3390/children8030252. Accessed: Sep. 12, 2022. [Online]. Available: https://www.mdpi.com/2227-9067/8/3/252.

[8] D. R. Patel, M. Neelakantan, K. Pandher, and J. Merrick, « Cerebral palsy in children: a clinical overview », *Translational Pediatrics*, vol. 9, *Suppl 1*, S125–S135, Feb. 2020, ISSN: 2224-4336. DOI: 10.21037/tp.2020.01.01. Accessed: Sep. 12, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7082248/.

[9] *La Fondation Paralysie Cérébrale | Recherche sur la paralysie cérébrale*. Accessed: Jun. 22, 2023. [Online]. Available: https://www.fondationparalysiecerebrale.org/.

[10] *Dictionnaire de l'Académie Nationale de Médecine*. Accessed: Jun. 22, 2023. [Online]. Available: http://dictionnaire.academie-medecine.fr/.

[11] A. Horsch, M. C. M. Klotz, H. Platzer, S. Seide, N. Zeaiter, and M. Ghandour, « Is the Prevalence of Equinus Foot in Cerebral Palsy Overestimated? Results from a Meta-Analysis of 4814 Feet », en, *Journal of Clinical Medicine*, vol. 10, *18*, p. 4128, Jan. 2021, Number: 18 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2077-0383. DOI: 10.3390/jcm10184128. Accessed: Sep. 12, 2022. [Online]. Available: https://www.mdpi.com/2077-0383/10/18/4128.

[12] M. C. Gourdine-Shaw, B. M. Lamm, J. E. Herzenberg, and A. Bhave, « Equinus Deformity in the Pediatric Patient: Causes, Evaluation, and Management », English, *Clinics in Podiatric Medicine and Surgery*, vol. 27, *1*, pp. 25–42, Jan. 2010, Publisher: Elsevier, ISSN: 0891-8422, 1558-2302. DOI: 10.1016/j.cpm.2009.10.003. Accessed: Aug. 16, 2023. [Online]. Available: https://www.podiatric.theclinics.com/article/S0891-8422(09)00105-0/fulltext.

[13] University Hospital, Brest, « In Vivo Dynamic Evaluation of Ankle Joint and Muscle Mechanics in Children With Spastic Equinus Deformity Due to Cerebral Palsy: Implications for Recurrent Equinus. », clinicaltrials.gov, Clinical trial registration NCT02814786, Dec. 2021, submitted: February 10, 2016. Accessed: Aug. 15, 2023. [Online]. Available: https://clinicaltrials.gov/study/NCT02814786.

[14] L. Servier, *Lateral collateral ligament of ankle joint*, en, Apr. 2020. Accessed: Mar. 14, 2024. [Online]. Available: https://commons.wikimedia.org/wiki/File:Lateral_collateral_ligament_of_ankle_joint.png.

[15] *Ankle Feet Joints*, fr, May 2013. Accessed: Mar. 14, 2024. [Online]. Available: https://commons.wikimedia.org/wiki/File:919_Ankle_Feet_Joints.jpg.

[16] C. L. Brockett and G. J. Chapman, « Biomechanics of the ankle », en, *Orthopaedics and Trauma*, vol. 30, *3*, pp. 232–238, Jun. 2016, ISSN: 1877-1327. DOI: 10.1016/j.mporth.2016.04.015. Accessed: Jun. 19, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877132716300483.

[17] G. Cobeljic, M. Bumbasirevic, A. Lesic, and Z. Bajin, « The management of spastic equinus in cerebral palsy », en, *Orthopaedics and Trauma*, vol. 23, *3*, pp. 201–209, Jun. 2009, ISSN: 1877-1327. DOI: 10.1016/j.mporth.2009.05.003. Accessed: May 30, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877132709000761.

[18] S. Yagel, S. M. Cohen, I. Shapiro, and D. V. Valsky, « 3D and 4D ultrasound in fetal cardiac scanning: a new look at the fetal heart », en, *Ultrasound in Obstetrics & Gynecology*, vol. 29, *1*, pp. 81–95, 2007, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ ISSN: 1469-0705. DOI: 10.1002/uog.3912. Accessed: Jun. 20, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/uog.3912.

[19] R. Guillin, A. J. Marchand, A. Roux, E. Niederberger, and R. Duvauferrier, « Imaging of snapping phenomena », *The British Journal of Radiology*, vol. 85, *1018*, pp. 1343–1353, Oct. 2012, Publisher: The British Institute of Radiology, ISSN: 0007-1285. DOI: 10.1259/bjr/52009417. Accessed: Jun. 20, 2023. [Online]. Available: https://www.birpublications.org/doi/full/10.1259/bjr/52009417.

[20] E. D. Brandner et al., « Abdominal organ motion measured using 4D CT », en, *International Journal of Radiation Oncology*Biology*Physics*, vol. 65, *2*, pp. 554–560, Jun. 2006, ISSN: 0360-3016. DOI: 10.1016/j.ijrobp.2005.12.042. Accessed:

Jun. 20, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360301606000794.

[21] Y. Kwong, A. O. Mel, G. Wheeler, and J. M. Troupis, « Four-dimensional computed tomography (4DCT): A review of the current status and applications », en, *Journal of Medical Imaging and Radiation Oncology*, vol. 59, *5*, pp. 545–554, 2015, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1754-9485.12326, ISSN: 1754-9485. DOI: 10.1111/1754-9485.12326. Accessed: Jun. 20, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1754-9485.12326.

[22] W. C. Bae, T. Ruangchaijatuporn, and C. B. Chung, « New Techniques in MR Imaging of the Ankle and Foot », English, *Magnetic Resonance Imaging Clinics*, vol. 25, *1*, pp. 211–225, Feb. 2017, Publisher: Elsevier, ISSN: 1064-9689, 1557-9786. DOI: 10.1016/j.mric.2016.08.009. Accessed: Sep. 12, 2022. [Online]. Available: https://www.mri.theclinics.com/article/S1064-9689(16)30065-4/fulltext.

[23] H. H. Quick et al., « Real-time MRI of joint movement with trueFISP », en, *Journal of Magnetic Resonance Imaging*, vol. 15, *6*, pp. 710–715, 2002, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.10120, ISSN: 1522-2586. DOI: 10.1002/jmri.10120. Accessed: Jun. 20, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.10120.

[24] R. I. Pettigrew, « Dynamic Cardiac MR Imaging Techniques and Applications », en, *Radiologic Clinics of North America*, vol. 27, *6*, pp. 1183–1203, Nov. 1989, ISSN: 0033-8389. DOI: 10.1016/S0033-8389(22)01205-2. Accessed: Sep. 12, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0033838922012052.

[25] M. A. Clark, M. Douglas, and J. Choi, *Muscle des jambes*, Sep. 2023. Accessed: Mar. 14, 2024. [Online]. Available: https://commons.wikimedia.org/wiki/File:Muscle_jambe.png?uselang=fr.

[26] E. L. Hahn, « Spin Echoes », *Physical Review*, vol. 80, *4*, pp. 580–594, Nov. 1950, Publisher: American Physical Society. DOI: 10.1103/PhysRev.80.580. Accessed: Jun. 22, 2023. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.80.580.

[27] 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, May 2020, ISSN: 1869-4101. DOI: 10.1186/s13244-020-00868-5. Accessed: Jun. 20, 2023. [Online]. Available: https://doi.org/10.1186/s13244-020-00868-5.

[28] M. Conconi, F. De Carli, M. Berni, N. Sancisi, V. Parenti-Castelli, and G. Monetti, « In-Vivo Quantification of Knee Deep-Flexion in Physiological Loading Condition trough Dynamic MRI », en, *Applied Sciences*, vol. 13, *1*, p. 629, Jan. 2023, Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-3417. DOI: 10.3390/app13010629. Accessed: Jun. 22, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/1/629.

[29] F. T. Sheehan, A. R. Seisler, and K. E. Alter, « Three-dimensional in vivo quantification of knee kinematics in cerebral palsy », eng, *Clinical Orthopaedics and Related Research*, vol. 466, *2*, pp. 450–458, Feb. 2008, ISSN: 0009-921X. DOI: 10.1007/s11999-007-0004-7.

[30] 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 », en, *PLOS ONE*, vol. 12, *12*, e0189587, Dec. 2017, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0189587. Accessed: Jun. 26, 2023. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0189587.

[31] K. Makki, « Development of dynamic MRI to study the musculoskeletal system during motion », en, Ph.D. dissertation, Ecole nationale supérieure Mines-Télécom Atlantique, Oct. 2019. Accessed: Sep. 12, 2022. [Online]. Available: https://tel.archives-ouvertes.fr/tel-02414706.

[32] P. A. Yushkevich et al., « User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability », en, *NeuroImage*, vol. 31, *3*, pp. 1116–1128, Jul. 2006, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2006.01.015. Accessed: Jun. 26, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811906000632.

[33] Y. Cheng et al., « Comprehensive personalized ankle joint shape analysis of children with cerebral palsy from pediatric MRI », en, *Frontiers in Bioengineering and Biotechnology*, vol. 10, Nov. 2022, ISSN: 2296-4185. DOI: 10.3389/fbioe.2022.1059129. Accessed: Jun. 26, 2023. [Online]. Available: https://doi.org/10.3389/fbioe.2022.1059129.

[34] K. Makki et al., « 4D in vivo quantification of ankle joint space width using dynamic MRI », eng, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2019, pp. 2115–2118, Jul. 2019, ISSN: 2694-0604. DOI: 10.1109/EMBC.2019.8856687.

[35] K. Makki, B. Borotikar, M. Garetier, S. Brochard, D. Ben Salem, and F. Rousseau, « In vivo ankle joint kinematics from dynamic magnetic resonance imaging using a registration-based framework », eng, *Journal of Biomechanics*, vol. 86, pp. 193–203, Mar. 2019, ISSN: 1873-2380. DOI: 10.1016/j.jbiomech.2019.02.007.

[36] K. Makki, B. Borotikar, M. Garetier, S. Brochard, D. Ben Salem, and F. Rousseau, « Temporal resolution enhancement of dynamic MRI sequences within a motion-based framework », eng, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2019, pp. 4004–4007, Jul. 2019, ISSN: 2694-0604. DOI: 10.1109/EMBC.2019.8857749.

[37] K. Makki, B. Borotikar, M. Garetier, S. Brochard, D. B. Salem, and F. Rousseau, « HIGH-RESOLUTION TEMPORAL RECONSTRUCTION OF ANKLE JOINT FROM DYNAMIC MRI », en, IEEE, Apr. 2018. DOI: 10.1109/ISBI.2018.8363809. Accessed: Jun. 26, 2023. [Online]. Available: https://imt-atlantique.hal.science/hal-02286013.

[38] A. M. Turing, « Computing Machinery and Intelligence », en, *in Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, R. Epstein, G. Roberts, and G. Beber, Eds., Dordrecht: Springer Netherlands, 2009, pp. 23–65, ISBN: 9781402067105. DOI: 10.1007/978-1-4020-6710-5_3. Accessed: Aug. 1, 2023. [Online]. Available: https://doi.org/10.1007/978-1-4020-6710-5_3.

[39] Conseil de l'Europe, *Histoire de l'intelligence artificielle*. Accessed: Aug. 1, 2023. [Online]. Available: https://www.coe.int/fr/web/artificial-intelligence/history-of-ai.

[40] C. Janiesch, P. Zschech, and K. Heinrich, « Machine learning and deep learning », en, *Electronic Markets*, vol. 31, *3*, pp. 685–695, Sep. 2021, ISSN: 1422-8890. DOI: 10.1007/s12525-021-00475-2. Accessed: Jul. 31, 2023. [Online]. Available: https://doi.org/10.1007/s12525-021-00475-2.

[41] C. M. Bishop and H. Bishop, *Deep Learning: Foundations and Concepts*, en. Cham: Springer International Publishing, 2024, ISBN: 978-3-031-45467-7 978-3-031-45468-4. DOI: 10.1007/978-3-031-45468-4. Accessed: Mar. 4, 2024. [Online]. Available: https://link.springer.com/10.1007/978-3-031-45468-4.

[42] L. Ruthotto and E. Haber, « An introduction to deep generative modeling », en, *GAMM-Mitteilungen*, vol. 44, *2*, e202100008, 2021, _eprint: https://onlinelibrary.wiley.com/do ISSN: 1522-2608. DOI: 10.1002/gamm.202100008. Accessed: Mar. 4, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/gamm.202100008.

[43] Y. LeCun, Y. Bengio, and G. Hinton, « Deep learning | Nature », *Nature*, vol. 521, *7553*, pp. 436–444, May 2015, ISSN: 1476-4687. DOI: 10.1038/nature14539. Accessed: Jul. 31, 2023. [Online]. Available: https://www.nature.com/articles/nature14539.

[44] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, « A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects », *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, *12*, pp. 6999–7019, Dec. 2022, Conference Name: IEEE Transactions on Neural Networks and Learning Systems, ISSN: 2162-2388. DOI: 10.1109/TNNLS.2021.3084827. Accessed: Feb. 21, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9451544.

[45] I. Goodfellow et al., « Generative adversarial networks », *Communications of the ACM*, vol. 63, *11*, pp. 139–144, Oct. 2020, ISSN: 0001-0782. DOI: 10.1145/3422622. Accessed: Sep. 12, 2022. [Online]. Available: https://doi.org/10.1145/3422622.

[46] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, « Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks », *in Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232. Accessed:

Sep. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html.

[47] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, « Contrastive Learning for Unpaired Image-to-Image Translation », en, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 319–345, ISBN: 9783030585457. DOI: 10.1007/978-3-030-58545-7_19.

[48] Y. Taigman, A. Polyak, and L. Wolf, « Unsupervised Cross-Domain Image Generation », en, *in International Conference on Learning Representations*, Jul. 2022. Accessed: Apr. 4, 2024. [Online]. Available: https://openreview.net/forum?id=Sk2Im59ex.

[49] Z. Wang, X. Tang, W. Luo, and S. Gao, « Face Aging With Identity-Preserved Conditional Generative Adversarial Networks », 2018, pp. 7939–7947. Accessed: Sep. 18, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Face_Aging_With_CVPR_2018_paper.html.

[50] H. Zhang, V. Sindagi, and V. M. Patel, « Image De-Raining Using a Conditional Generative Adversarial Network », *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, *11*, pp. 3943–3956, Nov. 2020, ISSN: 1558-2205. DOI: 10.1109/TCSVT.2019.2920407.

[51] K. E. Smith and A. O. Smith, *Conditional GAN for timeseries generation*, arXiv:2006.16477 [cs, stat], Jun. 2020. DOI: 10.48550/arXiv.2006.16477. Accessed: Sep. 18, 2023. [Online]. Available: http://arxiv.org/abs/2006.16477.

[52] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju, « Improving Image Captioning with Conditional Generative Adversarial Nets », en, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, *01*, pp. 8142–8150, Jul. 2019, ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.33018142. Accessed: Sep. 18, 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4823.

[53] D. Singh and B. Singh, « Investigating the impact of data normalization on classification performance », *Applied Soft Computing*, vol. 97, p. 105 524, Dec. 2020, ISSN: 1568-4946. DOI: 10.1016/j.asoc.2019.105524. Accessed: Mar. 4, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494619302947.

[54] N. Singh and P. Singh, « Exploring the effect of normalization on medical data classification », *in 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, Sep. 2021, pp. 1–5. DOI: 10.1109/AIMV53313.2021.9670938. Accessed: Mar. 5, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9670938.

[55] C. Shorten and T. M. Khoshgoftaar, « A survey on Image Data Augmentation for Deep Learning », *Journal of Big Data*, vol. 6, *1*, p. 60, Jul. 2019, ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0. Accessed: Sep. 18, 2023. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0.

[56] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, « Image-To-Image Translation With Conditional Adversarial Networks », *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134. Accessed: May 11, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html.

[57] K. Lum, « Limitations of mitigating judicial bias with machine learning | Nature Human Behaviour », *Nature Human Behaviour*, vol. 1, *7*, p. 0141, Jun. 2017, ISSN: 2397-3374. DOI: 10.1038/s41562-017-0141. Accessed: Sep. 18, 2023. [Online]. Available: https://www.nature.com/articles/s41562-017-0141.

[58] M.-H. Huang and R. T. Rust, « A strategic framework for artificial intelligence in marketing », en, *Journal of the Academy of Marketing Science*, vol. 49, *1*, pp. 30–50, Jan. 2021, ISSN: 1552-7824. DOI: 10.1007/s11747-020-00749-9. Accessed: Sep. 18, 2023. [Online]. Available: https://doi.org/10.1007/s11747-020-00749-9.

[59] D. V. Ps, « How can we manage biases in artificial intelligence systems – A systematic literature review », *International Journal of Information Management Data Insights*, vol. 3, *1*, p. 100 165, Apr. 2023, ISSN: 2667-0968. DOI: 10.1016/j.jjimei.2023.100165. Accessed: Sep. 18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096823000125.

[60] A. Khademi and V. Honavar, « Algorithmic Bias in Recidivism Prediction: A Causal Perspective (Student Abstract) », en, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, *10*, pp. 13 839–13 840, Apr. 2020, ISSN: 2374-3468.

DOI: 10.1609/aaai.v34i10.7192. Accessed: Sep. 18, 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/7192.

[61] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, « A Survey on Bias and Fairness in Machine Learning », *ACM Computing Surveys*, vol. 54, *6*, 115:1–115:35, Jul. 2021, ISSN: 0360-0300. DOI: 10.1145/3457607. Accessed: Sep. 18, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3457607.

[62] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, « Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations », en, *Nature Medicine*, vol. 27, *12*, pp. 2176–2182, Dec. 2021, Publisher: Nature Publishing Group, ISSN: 1546-170X. DOI: 10.1038/s41591-021-01595-0. Accessed: Jul. 12, 2024. [Online]. Available: https://www.nature.com/articles/s41591-021-01595-0.

[63] M. A. Ricci Lara, R. Echeveste, and E. Ferrante, « Addressing fairness in artificial intelligence for medical imaging », en, *Nature Communications*, vol. 13, *1*, p. 4581, Aug. 2022, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-022-32186-3. Accessed: Jul. 11, 2024. [Online]. Available: https://www.nature.com/articles/s41467-022-32186-3.

[64] G. Varoquaux and V. Cheplygina, « Machine learning for medical imaging: methodological failures and recommendations for the future », en, *npj Digital Medicine*, vol. 5, *1*, pp. 1–8, Apr. 2022, Publisher: Nature Publishing Group, ISSN: 2398-6352. DOI: 10.1038/s41746-022-00592-y. Accessed: Jul. 10, 2024. [Online]. Available: https://www.nature.com/articles/s41746-022-00592-y.

[65] F. Zhan et al., « Multimodal Image Synthesis and Editing: A Survey and Taxonomy », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2023, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2023.3305243.

[66] *IXI Dataset – Brain Development*. Accessed: Sep. 13, 2023. [Online]. Available: https://brain-development.org/ixi-dataset/.

[67] O. Dalmaz, B. Saglam, G. Elmas, M. Mirza, and T. Çukur, « Denoising Diffusion Adversarial Models for Unconditional Medical Image Generation », *in 2023 31st Signal Processing and Communications Applications Conference (SIU)*, ISSN: 2165-0608, Jul. 2023, pp. 1–5. DOI: 10.1109/SIU59756.2023.10223912.

[68] J. Ho, A. Jain, and P. Abbeel, « Denoising Diffusion Probabilistic Models », vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851. Accessed: Sep. 13, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1a
Abstract.html.

[69] S. Hong et al., « 3D-StyleGAN: A Style-Based Generative Adversarial Network for Generative Modeling of Three-Dimensional Medical Images », en, S. Engelhardt et al., Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 24–34, ISBN: 978-3-030-88210-5. DOI: 10.1007/978-3-030-88210-5_3.

[70] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, « Adversarial text-to-image synthesis: A review », *Neural Networks*, vol. 144, pp. 187–209, Dec. 2021, ISSN: 0893-6080. DOI: 10.1016/j.neunet.2021.07.019. Accessed: Sep. 13, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608021002823.

[71] H. Zhang et al., « StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks », 2017, pp. 5907–5915. Accessed: Sep. 13, 2023. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Zhang_StackGAN_Text_to_ICCV_2017_paper.html.

[72] M. Kang et al., « Scaling Up GANs for Text-to-Image Synthesis », 2023, pp. 10 124–10 134. Accessed: Sep. 13, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Kang_Scaling_Up_GANs_for_Text-to-Image_Synthesis_CVPR_2023_paper.html.

[73] S. S. Baraheem, T.-N. Le, and T. V. Nguyen, « Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook », en, *Artificial Intelligence Review*, vol. 56, *10*, pp. 10 813–10 865, Oct. 2023, ISSN: 1573-7462. DOI: 10.1007/s10462-023-10434-2. Accessed: Sep. 13, 2023. [Online]. Available: https://doi.org/10.1007/s10462-023-10434-2.

[74] C. Saharia et al., « Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding », *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, Dec. 2022. Accessed: Sep. 13, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf
Abstract-Conference.html.

[75] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, *Hierarchical Text-Conditional Image Generation with CLIP Latents*, arXiv:2204.06125 [cs], Apr. 2022. DOI: 10.48550/arXiv.2204.06125. Accessed: Sep. 13, 2023. [Online]. Available: http://arxiv.org/abs/2204.06125.

[76] H. Chang et al., *Muse: Text-To-Image Generation via Masked Generative Transformers*, arXiv:2301.00704 [cs], Jan. 2023. DOI: 10.48550/arXiv.2301.00704. Accessed: Sep. 13, 2023. [Online]. Available: http://arxiv.org/abs/2301.00704.

[77] A. Ramesh et al., « Zero-Shot Text-to-Image Generation », en, *in Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8821–8831. Accessed: Sep. 13, 2023. [Online]. Available: https://proceedings.mlr.press/v139/ramesh21a.html.

[78] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, *DreamFusion: Text-to-3D using 2D Diffusion*, arXiv:2209.14988 [cs, stat], Sep. 2022. DOI: 10.48550/arXiv.2209.14988. Accessed: Sep. 13, 2023. [Online]. Available: http://arxiv.org/abs/2209.14988.

[79] C.-H. Lin et al., « Magic3D: High-Resolution Text-to-3D Content Creation », 2023, pp. 300–309. Accessed: Sep. 13, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Lin_Magic3D_High-Resolution_Text-to-3D_Content_Creation_CVPR_2023_paper.html.

[80] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, « Distributed Representations of Words and Phrases and their Compositionality », vol. 26, Curran Associates, Inc., 2013. Accessed: Sep. 13, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

[81] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, « Generative Adversarial Text to Image Synthesis », en, *in Proceedings of The 33rd International Conference on Machine Learning*, PMLR, Jun. 2016, pp. 1060–1069. Accessed: Sep. 13, 2023. [Online]. Available: https://proceedings.mlr.press/v48/reed16.html.

[82] T. Xu et al., « AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks », 2018, pp. 1316–1324. Accessed: Sep. 13, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Xu_AttnGAN_Fine-Grained_Text_CVPR_2018_paper.html.

[83] L. Chen, S. Srivastava, Z. Duan, and C. Xu, « Deep Cross-Modal Audio-Visual Generation », ser. Thematic Workshops '17, New York, NY, USA: Association for Computing Machinery, Oct. 2017, pp. 349–357, ISBN: 978-1-4503-5416-5. DOI: 10.1145/3126686.3126723. Accessed: Sep. 13, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3126686.3126723.

[84] J. Li et al., « Direct Speech-to-Image Translation », *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, *3*, pp. 517–529, Mar. 2020, ISSN: 1941-0484. DOI: 10.1109/JSTSP.2020.2987417.

[85] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, « A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild », ser. MM '20, New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 484–492, ISBN: 978-1-4503-7988-5. DOI: 10.1145/3394171.3413532. Accessed: Sep. 13, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413532.

[86] Y. Song, J. Zhu, D. Li, X. Wang, and H. Qi, *Talking Face Generation by Conditional Recurrent Adversarial Network*, arXiv:1804.04786 [cs], Jul. 2019. DOI: 10.48550/arXiv.1804.04786. Accessed: Sep. 13, 2023. [Online]. Available: http://arxiv.org/abs/1804.04786.

[87] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, « Sketch2Photo: internet image montage », *ACM Transactions on Graphics*, vol. 28, *5*, pp. 1–10, Dec. 2009, ISSN: 0730-0301. DOI: 10.1145/1618452.1618470. Accessed: Sep. 14, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/1618452.1618470.

[88] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, « SketchyCOCO: Image Generation From Freehand Scene Sketches », 2020, pp. 5174–5183. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Gao_SketchyCOCO_Image_Generation_From_Freehand_Scene_Sketches_CVPR_2020_paper.html.

[89] W. Chen and J. Hays, « SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis », 2018, pp. 9416–9425. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Chen_SketchyGAN_Towards_Diverse_CVPR_2018_paper.html.

[90] R. Zhang, P. Isola, and A. A. Efros, « Colorful Image Colorization », en, *in Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 649–666, ISBN: 978-3-319-46487-9. DOI: 10.1007/978-3-319-46487-9_40.

[91] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, « Infrared Image Colorization Based on a Triplet DCGAN Architecture », *in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, ISSN: 2160-7516, Jul. 2017, pp. 212–217. DOI: 10.1109/CVPRW.2017.32.

[92] C. Ledig et al., « Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network », *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690. Accessed: Sep. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.html.

[93] Y. Hu, X. Gao, J. Li, Y. Huang, and H. Wang, « Single image super-resolution with multi-scale information cross-fusion network », en, *Signal Processing*, vol. 179, p. 107 831, Feb. 2021, ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2020.107831. Accessed: Sep. 12, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168420303753.

[94] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, « TSIT: A Simple and Versatile Framework for Image-to-Image Translation », en, *in Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 206–222, ISBN: 978-3-030-58580-8. DOI: 10.1007/978-3-030-58580-8_13.

[95] L. A. Gatys, A. S. Ecker, and M. Bethge, « Image Style Transfer Using Convolutional Neural Networks », en, *in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2414–2423, ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.265. Accessed: Sep. 8, 2022. [Online]. Available: http://ieeexplore.ieee.org/document/7780634/.

[96] J. Long, E. Shelhamer, and T. Darrell, « Fully Convolutional Networks for Semantic Segmentation », 2015, pp. 3431–3440. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html.

[97] A. Chartsias et al., « Disentangled representation learning in cardiac image analysis », en, *Medical Image Analysis*, vol. 58, p. 101 535, Dec. 2019, ISSN: 1361-8415. DOI: 10.1016/j.media.2019.101535. Accessed: Sep. 12, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841519300684.

[98] S. Li, S. Gunel, M. Ostrek, P. Ramdya, P. Fua, and H. Rhodin, « Deformation-Aware Unpaired Image Translation for Pose Estimation on Laboratory Animals », 2020, pp. 13 158–13 168. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Deformation-Aware_Unpaired_Image_Translation_for_Pose_Estimation_on_Laboratory_Animals_CVPR_2020_paper.html.

[99] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, « Context Encoders: Feature Learning by Inpainting », 2016, pp. 2536–2544. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Pathak_Context_Encoders_Feature_CVPR_2016_paper.html.

[100] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, « Semantic Image Synthesis With Spatially-Adaptive Normalization », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346. Accessed: Sep. 8, 2022. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Semantic_Image_Synthesis_With_Spatially-Adaptive_Normalization_CVPR_2019_paper.html.

[101] H. Tang, D. Xu, Y. Yan, P. H. S. Torr, and N. Sebe, « Local Class-Specific and Global Image-Level Generative Adversarial Networks for Semantic-Guided Scene Generation », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7870–7879. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Tang_Local_Class-Specific_and_Global_Image-Level_Generative_Adversarial_Networks_for_Semantic-Guided_CVPR_2020_paper.html.

[102] A. Buades, B. Coll, and J.-M. Morel, « A non-local algorithm for image denoising », *in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, ISSN: 1063-6919, vol. 2, Jun. 2005, 60–65 vol. 2. DOI: 10.1109/CVPR.2005.38.

[103] X. Liu et al., « Open-Edit: Open-Domain Image Manipulation with Open-Vocabulary Instructions », en, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 89–106, ISBN: 9783030586218. DOI: 10.1007/978-3-030-58621-8_6.

[104] K. Crowson et al., « VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance », en, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., ser. Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2022, pp. 88–105, ISBN: 9783031198366. DOI: 10.1007/978-3-031-19836-6_6.

[105] Y. Shi, D. Deb, and A. K. Jain, « WarpGAN: Automatic Caricature Generation », 2019, pp. 10 762–10 771. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Shi_WarpGAN_Automatic_Caricature_Generation_CVPR_2019_paper.html.

[106] Y. Chen, Y.-K. Lai, and Y.-J. Liu, « CartoonGAN: Generative Adversarial Networks for Photo Cartoonization », 2018, pp. 9465–9474. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Chen_CartoonGAN_Generative_Adversarial_CVPR_2018_paper.html.

[107] M. Brown and S. Süsstrunk, « Multi-spectral SIFT for scene category recognition », *in CVPR 2011*, ISSN: 1063-6919, Jun. 2011, pp. 177–184. DOI: 10.1109/CVPR.2011.5995637. Accessed: Feb. 21, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/5995637.

[108] 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*, ISSN: 1063-6919, Jun. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848. Accessed: Feb. 21, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/5206848.

[109] M. Cordts et al., « The Cityscapes Dataset for Semantic Urban Scene Understanding », 2016, pp. 3213–3223. Accessed: Feb. 23, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html.

[110] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, « High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs », 2018, pp. 8798–8807. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_High-Resolution_Image_Synthesis_CVPR_2018_paper.html.

[111] Z. Yi, H. Zhang, P. Tan, and M. Gong, « DualGAN: Unsupervised Dual Learning for Image-To-Image Translation », 2017, pp. 2849–2857. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Yi_DualGAN_Unsupervised_Dual_ICCV_2017_paper.html.

[112] S. Ma, J. Fu, C. W. Chen, and T. Mei, « DA-GAN: Instance-Level Image Translation by Deep Attention Generative Adversarial Networks », 2018, pp. 5657–5666. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Ma_DA-GAN_Instance-Level_Image_CVPR_2018_paper.html.

[113] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, « Semi-supervised Learning with Deep Generative Models », *in Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014. Accessed: Sep. 14, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/d523773c6b194f37b938d340d5d02232-Abstract.html.

[114] A. Mustafa and R. K. Mantiuk, « Transformation Consistency Regularization – A Semi-supervised Paradigm for Image-to-Image Translation », en, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 599–615, ISBN: 9783030585235. DOI: 10.1007/978-3-030-58523-5_35.

[115] U. Ojha et al., « Few-Shot Image Generation via Cross-Domain Correspondence », 2021, pp. 10 743–10 752. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Ojha_Few-Shot_Image_Generation_via_Cross-Domain_Correspondence_CVPR_2021_paper.html.

[116] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, « Transferring GANs: generating images from limited data », 2018, pp. 218–234. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/yaxing_wang_Transferring_GANs_generating_ECCV_2018_paper.html.

[117] M. Tomei, M. Cornia, L. Baraldi, and R. Cucchiara, « Art2Real: Unfolding the Reality of Artworks via Semantically-Aware Image-To-Image Translation », 2019, pp. 5849–5859. Accessed: Sep. 15, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Tomei_Art2Real_Unfolding_the_Reality_of_Artworks_via_Semantically-Aware_Image-To-Image_Translation_CVPR_2019_paper.html.

[118] X. Huang and S. Belongie, « Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization », *in Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510. Accessed: Sep. 8, 2022. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.html.

[119] X. Luo, Z. Han, and L. Yang, « Progressive Attentional Manifold Alignment for Arbitrary Style Transfer », en, *in Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3206–3222. Accessed: May 11, 2023. [Online]. Available: https://openaccess.thecvf.com/content/ACCV2022/html/Luo_Progressive_Attentional_Manifold_Alignment_for_Arbitrary_Style_Transfer_ACCV_2022_paper.html.

[120] D. Y. Park and K. H. Lee, « Arbitrary Style Transfer With Style-Attentional Networks », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5880–5888. Accessed: May 11, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Arbitrary_Style_Transfer_With_Style-Attentional_Networks_CVPR_2019_paper.html.

[121] X. Liu, P. Sanchez, S. Thermos, A. Q. O'Neil, and S. A. Tsaftaris, « Learning disentangled representations in the imaging domain », en, *Medical Image Analysis*, vol. 80, p. 102 516, Aug. 2022, ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102516. Accessed: Sep. 8, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522001633.

[122] J. Chen, S. Chen, L. Wee, A. Dekker, and I. Bermejo, « Deep learning based unpaired image-to-image translation applications for medical physics: a systematic review », en, *Physics in Medicine & Biology*, vol. 68, *5*, 05TR01, Feb. 2023, Publisher: IOP Publishing, ISSN: 0031-9155. DOI: 10.1088/1361-6560/acba74. Ac-

cessed: Nov. 29, 2023. [Online]. Available: https://dx.doi.org/10.1088/1361-6560/acba74.

[123] I. Higgins et al., *Towards a Definition of Disentangled Representations*, arXiv:1812.02230 [cs, stat], Dec. 2018. DOI: 10.48550/arXiv.1812.02230. Accessed: Dec. 4, 2023. [Online]. Available: http://arxiv.org/abs/1812.02230.

[124] V. Thomas et al., *Independently Controllable Factors*, en, arXiv:1708.01289 [cs, stat], Aug. 2017. Accessed: Sep. 22, 2022. [Online]. Available: http://arxiv.org/abs/1708.01289.

[125] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, *Disentangled Representation Learning*, arXiv:2211.11695 [cs], Aug. 2023. DOI: 10.48550/arXiv.2211.11695. Accessed: Feb. 14, 2024. [Online]. Available: http://arxiv.org/abs/2211.11695.

[126] H.-Y. Lee et al., « DRIT++: Diverse Image-to-Image Translation via Disentangled Representations », en, *International Journal of Computer Vision*, vol. 128, *10*, pp. 2402–2417, Nov. 2020, ISSN: 1573-1405. DOI: 10.1007/s11263-019-01284-z. Accessed: Sep. 8, 2022. [Online]. Available: https://doi.org/10.1007/s11263-019-01284-z.

[127] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, « Multimodal Unsupervised Image-to-image Translation », *in Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189. Accessed: Sep. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/Xun_Huang_Multimodal_Unsupervised_Image-to-image_ECCV_2018_paper.html.

[128] J. Lin, Y. Xia, T. Qin, Z. Chen, and T.-Y. Liu, « Conditional Image-to-Image Translation », 2018, pp. 5524–5532. Accessed: Sep. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Lin_Conditional_Image-to-Image_Translation_CVPR_2018_paper.html.

[129] L. Hui, X. Li, J. Chen, H. He, and J. Yang, « Unsupervised Multi-Domain Image Translation with Domain-Specific Encoders/Decoders », *in 2018 24th International Conference on Pattern Recognition (ICPR)*, ISSN: 1051-4651, Aug. 2018, pp. 2044–2049. DOI: 10.1109/ICPR.2018.8545169.

[130] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, « StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation », 2018, pp. 8789–8797. Accessed: Sep. 15, 2023. [Online]. Available: https://

openaccess.thecvf.com/content_cvpr_2018/html/Choi_StarGAN_Unified_Generative_CVPR_2018_paper.html.

[131] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, « Deep learning in medical image registration: a review », en, *Physics in Medicine &amp$\mathsemicolon$ Biology*, vol. 65, *20*, 20TR01, Oct. 2020, Publisher: IOP Publishing, ISSN: 0031-9155. DOI: 10.1088/1361-6560/ab843e. Accessed: Sep. 12, 2022. [Online]. Available: https://doi.org/10.1088/1361-6560/ab843e.

[132] A. Tiwari, S. Srivastava, and M. Pant, « Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019 », *Pattern Recognition Letters*, vol. 131, pp. 244–260, Mar. 2020, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2019.11.020. Accessed: Sep. 18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786551930340X.

[133] N. Decaux et al., « Semi-automatic muscle segmentation in MR images using deep registration-based label propagation », *Pattern Recognition*, vol. 140, p. 109 529, Aug. 2023, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2023.109529. Accessed: Sep. 18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323002297.

[134] J. Almotiri, K. Elleithy, and A. Elleithy, « Retinal Vessels Segmentation Techniques and Algorithms: A Survey », en, *Applied Sciences*, vol. 8, *2*, p. 155, Feb. 2018, ISSN: 2076-3417. DOI: 10.3390/app8020155. Accessed: Sep. 18, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/8/2/155.

[135] D. Nie, X. Cao, Y. Gao, L. Wang, and D. Shen, « Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks », en, G. Carneiro et al., Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2016, pp. 170–178, ISBN: 9783319469768. DOI: 10.1007/978-3-319-46976-8_18.

[136] X. Han, « MR-based synthetic CT generation using a deep convolutional neural network method », en, *Medical Physics*, vol. 44, *4*, pp. 1408–1419, 2017, ISSN: 2473-4209. DOI: 10.1002/mp.12155. Accessed: Sep. 18, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.12155.

[137] Y. Hiasa et al., « Cross-Modality Image Synthesis from Unpaired Data Using CycleGAN », en, *in Simulation and Synthesis in Medical Imaging*, A. Gooya, O. Goksel, I. Oguz, and N. Burgos, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 31–41, ISBN: 978-3-030-00536-8. DOI: 10.1007/978-3-030-00536-8_4.

[138] R. Gupta, A. Sharma, and A. Kumar, « Super-Resolution using GANs for Medical Imaging », *Procedia Computer Science*, International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020, vol. 173, pp. 28–35, Jan. 2020, ISSN: 1877-0509. DOI: 10.1016/j.procs.2020.06.005. Accessed: Sep. 18, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920315076.

[139] H.-M. Zhang and B. Dong, « A Review on Deep Learning in Medical Image Reconstruction », en, *Journal of the Operations Research Society of China*, vol. 8, *2*, pp. 311–340, Jun. 2020, ISSN: 2194-6698. DOI: 10.1007/s40305-019-00287-4. Accessed: Sep. 12, 2022. [Online]. Available: https://doi.org/10.1007/s40305-019-00287-4.

[140] I. Sanchez and V. Vilaplana, « Brain MRI super-resolution using 3D generative adversarial networks », en, *in Medical Imaging with Deep Learning (MIDL)*, 2018. Accessed: Apr. 4, 2024. [Online]. Available: https://openreview.net/forum?id=rJevSbniM.

[141] J. L. Prince, A. Carass, C. Zhao, B. E. Dewey, S. Roy, and D. L. Pham, « Chapter 1 - Image synthesis and superresolution in medical imaging », en, *in Handbook of Medical Image Computing and Computer Assisted Intervention*, ser. The Elsevier and MICCAI Society Book Series, S. K. Zhou, D. Rueckert, and G. Fichtinger, Eds., Academic Press, Jan. 2020, pp. 1–24, ISBN: 978-0-12-816176-0. DOI: 10.1016/B978-0-12-816176-0.00006-5. Accessed: Sep. 12, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128161760000065.

[142] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, « Compressed Sensing MRI Reconstruction Using a Generative Adversarial Network With a Cyclic Loss », *IEEE Transactions on Medical Imaging*, vol. 37, *6*, pp. 1488–1497, Jun. 2018, Conference Name: IEEE Transactions on Medical Imaging, ISSN: 1558-254X. DOI: 10.1109/TMI.2018.2820120.

[143] G. Yang et al., « DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction », *IEEE Transactions on Medical Imaging*, vol. 37, *6*, pp. 1310–1321, Jun. 2018, Conference Name: IEEE Transactions on Medical Imaging, ISSN: 1558-254X. DOI: 10.1109/TMI.2017.2785879.

[144] K. de Haan, Y. Rivenson, Y. Wu, and A. Ozcan, « Deep-Learning-Based Image Reconstruction and Enhancement in Optical Microscopy », *Proceedings of the IEEE*, vol. 108, *1*, pp. 30–50, Jan. 2020, ISSN: 1558-2256. DOI: 10.1109/JPROC.2019.2949575.

[145] A. Makropoulos et al., « The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction », eng, *NeuroImage*, vol. 173, pp. 88–112, Jun. 2018, ISSN: 1095-9572. DOI: 10.1016/j.neuroimage.2018.01.054.

[146] J. S. Isaac and R. Kulkarni, « Super resolution techniques for medical image processing », *in 2015 International Conference on Technologies for Sustainable Development (ICTSD)*, Feb. 2015, pp. 1–6. DOI: 10.1109/ICTSD.2015.7095900. Accessed: Mar. 3, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7095900/figures#figures.

[147] M. H. Ishak, N. N. Sofia Mohd Marzuki, M. F. Abdullah, Z. H. Che Soh, I. S. Isa, and S. N. Sulaiman, « Image Quality Assessment for Image Filtering Algorithm: Qualitative and Quantitative Analyses », *in 2019 9th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, Nov. 2019, pp. 162–167. DOI: 10.1109/ICCSCE47578.2019.9068565.

[148] M. Khosravy, N. Patel, N. Gupta, and I. K. Sethi, « Image Quality Assessment: A Review to Full Reference Indexes », en, *in Recent Trends in Communication, Computing, and Electronics*, A. Khare, U. S. Tiwary, I. K. Sethi, and N. Singh, Eds., ser. Lecture Notes in Electrical Engineering, Singapore: Springer, 2019, pp. 279–288, ISBN: 9789811326851. DOI: 10.1007/978-981-13-2685-1_27.

[149] S. Dost, F. Saud, M. Shabbir, M. G. Khan, M. Shahid, and B. Lovstrom, « Reduced reference image and video quality assessments: review of methods », *EURASIP Journal on Image and Video Processing*, vol. 2022, *1*, p. 1, Jan. 2022, ISSN: 1687-5281. DOI: 10.1186/s13640-021-00578-y. Accessed: Feb. 13, 2023. [Online]. Available: https://doi.org/10.1186/s13640-021-00578-y.

[150] L. S. Chow and R. Paramesran, « Review of medical image quality assessment », en, *Biomedical Signal Processing and Control*, vol. 27, pp. 145–154, May 2016, ISSN: 1746-8094. DOI: 10.1016/j.bspc.2016.02.006. Accessed: Jun. 27, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809416300180.

[151] Z. Wang and A. C. Bovik, « Reduced- and No-Reference Image Quality Assessment », *IEEE Signal Processing Magazine*, vol. 28, *6*, pp. 29–40, Nov. 2011, Conference Name: IEEE Signal Processing Magazine, ISSN: 1558-0792. DOI: 10.1109/MSP.2011.942471. Accessed: Mar. 4, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6021882/authors#authors.

[152] Z. Wang, A. C. Bovik, and E. P. Simoncelli, « 8.3 - Structural Approaches to Image Quality Assessment », en, *in Handbook of Image and Video Processing (Second Edition)*, ser. Communications, Networking and Multimedia, A. Bovik, Ed., Burlington: Academic Press, Jan. 2005, pp. 961–974, ISBN: 978-0-12-119792-6. DOI: 10.1016/B978-012119792-6/50119-4. Accessed: Sep. 12, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780121197926501194.

[153] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, « GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium », *in Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. Accessed: Oct. 3, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

[154] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, « Demystifying MMD GANs », en, *in International Conference on Learning Representations*, Feb. 2018. Accessed: Apr. 4, 2024. [Online]. Available: https://openreview.net/forum?id=r1lUOzWCW.

[155] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, « The Unreasonable Effectiveness of Deep Features as a Perceptual Metric », 2018, pp. 586–595. Accessed: Jul. 25, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

[156] T. Salimans et al., « Improved Techniques for Training GANs », *in Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016. Accessed: Sep. 12, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

[157] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, « Rethinking the Inception Architecture for Computer Vision », 2016, pp. 2818–2826. Accessed: Sep. 12, 2022. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.

[158] L. N. Vaserstein, « Markovian processes on countable space product describing large systems of automata », Russian, *Problemy Peredachi Informatsii*, vol. 5, *3*, pp. 64–72, 1969, ISSN: 0555-2923.

[159] M. Woodland et al., « Evaluating the Performance of StyleGAN2-ADA on Medical Images », en, *in Simulation and Synthesis in Medical Imaging*, C. Zhao, D. Svoboda, J. M. Wolterink, and M. Escobar, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2022, pp. 142–153, ISBN: 978-3-031-16980-9. DOI: 10.1007/978-3-031-16980-9_14.

[160] I. Stępień and M. Oszust, « A Brief Survey on No-Reference Image Quality Assessment Methods for Magnetic Resonance Images », en, *Journal of Imaging*, vol. 8, *6*, p. 160, Jun. 2022, Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2313-433X. DOI: 10.3390/jimaging8060160. Accessed: Feb. 13, 2023. [Online]. Available: https://www.mdpi.com/2313-433X/8/6/160.

[161] J. Jang, K. Bang, H. Jang, D. Hwang, and f. t. A. D. N. Initiative, « Quality evaluation of no-reference MR images using multidirectional filters and image statistics », en, *Magnetic Resonance in Medicine*, vol. 80, *3*, pp. 914–924, 2018, ISSN: 1522-2594. DOI: 10.1002/mrm.27084. Accessed: Jul. 28, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.27084.

[162] S. Nabavi, H. Simchi, M. E. Moghaddam, A. F. Frangi, and A. A. Abin, *Automatic Multi-Class Cardiovascular Magnetic Resonance Image Quality Assessment using Unsupervised Domain Adaptation in Spatial and Frequency Domains*, arXiv:2112.06806 [eess], Dec. 2021. DOI: 10.48550/arXiv.2112.06806. Accessed: Jul. 28, 2023. [Online]. Available: http://arxiv.org/abs/2112.06806.

[163] R. Rodrigues et al., *Objective quality assessment of medical images and videos: Review and challenges*, arXiv:2212.07396 [eess], Dec. 2022. DOI: 10.48550/arXiv.2212.07396. Accessed: Jul. 28, 2023. [Online]. Available: http://arxiv.org/abs/2212.07396.

[164] L. S. Chow and H. Rajagopal, « Modified-BRISQUE as no reference image quality assessment for structural MR images », en, *Magnetic Resonance Imaging*, vol. 43, pp. 74–87, Nov. 2017, ISSN: 0730-725X. DOI: 10.1016/j.mri.2017.07.016. Accessed: Jul. 28, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0730725X17301340.

[165] M. Oszust, A. Piórkowski, and R. Obuchowicz, « No-reference image quality assessment of magnetic resonance images with high-boost filtering and local features », en, *Magnetic Resonance in Medicine*, vol. 84, *3*, pp. 1648–1660, 2020, ISSN: 1522-2594. DOI: 10.1002/mrm.28201. Accessed: Jul. 29, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.28201.

[166] I. Stępień and M. Oszust, « No-Reference Image Quality Assessment of Magnetic Resonance images with multi-level and multi-model representations based on fusion of deep architectures », en, *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106 283, Aug. 2023, ISSN: 0952-1976. DOI: 10.1016/j.engappai.2023.106283. Accessed: Jul. 29, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623004670.

[167] A. Liebgott, T. Küstner, S. Gatidis, F. Schick, and B. Yang, « Active learning for magnetic resonance image quality assessment », *in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISSN: 2379-190X, Mar. 2016, pp. 922–926. DOI: 10.1109/ICASSP.2016.7471810.

[168] R. Obuchowicz, M. Oszust, M. Bielecka, A. Bielecki, and A. Piórkowski, « Magnetic Resonance Image Quality Assessment by Using Non-Maximum Suppression and Entropy Analysis », en, *Entropy*, vol. 22, *2*, p. 220, Feb. 2020, ISSN: 1099-4300. DOI: 10.3390/e22020220. Accessed: Jul. 28, 2023. [Online]. Available: https://www.mdpi.com/1099-4300/22/2/220.

[169] M. Osadebey, M. Pedersen, D. Arnold, and K. Wendel-Mitoraj, « Bayesian framework inspired no-reference region-of-interest quality measure for brain MRI images », *Journal of Medical Imaging*, vol. 4, *2*, p. 025 504, Apr. 2017, ISSN: 2329-4302.

DOI: 10.1117/1.JMI.4.2.025504. Accessed: Jul. 29, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5469420/.

[170] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış, and N. S. Gezer, *CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data*, Apr. 2019. DOI: 10.5281/zenodo.3431873. Accessed: Mar. 8, 2024. [Online]. Available: https://zenodo.org/records/3431873.

[171] S. Chatterjee, C. Sarasaen, G. Rose, A. Nürnberger, and O. Speck, *DDoS-UNet: Incorporating temporal information using Dynamic Dual-channel UNet for enhancing super-resolution of dynamic MRI*, arXiv:2202.05355 [physics], Feb. 2022. DOI: 10.48550/arXiv.2202.05355. Accessed: Sep. 8, 2022. [Online]. Available: http://arxiv.org/abs/2202.05355.

[172] J. N. Freedman et al., « Rapid 4D-MRI reconstruction using a deep radial convolutional neural network: Dracula », *Radiotherapy and Oncology*, vol. 159, pp. 209–217, Jun. 2021, ISSN: 0167-8140. DOI: 10.1016/j.radonc.2021.03.034. Accessed: Mar. 8, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167814021061740.

[173] S. Zhi et al., « Coarse–Super-Resolution–Fine Network (CoSF-Net): A Unified End-to-End Neural Network for 4D-MRI With Simultaneous Motion Estimation and Super-Resolution », *IEEE Transactions on Medical Imaging*, vol. 43, *1*, pp. 162–174, Jan. 2024, Conference Name: IEEE Transactions on Medical Imaging, ISSN: 1558-254X. DOI: 10.1109/TMI.2023.3294245. Accessed: Mar. 8, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10177965?casa_token=u1WZA-eopS8AAAAA:DcbPwAcucZnB6iQ5cLUWqUeDmrDG_K0gwK5J_84qk2uYK8NNyOFWpjOqJLN2U60nboQd2CFIsg.

[174] C. Sarasaen, S. Chatterjee, M. Breitkopf, G. Rose, A. Nürnberger, and O. Speck, « Fine-tuning deep learning model parameters for improved super-resolution of dynamic MRI with prior-knowledge », *Artificial Intelligence in Medicine*, vol. 121, p. 102 196, Nov. 2021, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2021.102196. Accessed: Mar. 8, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365721001895.

[175] R. Sarah, M. Adrien, B. S. Douraied, and R. François, « EEG Source Imaging by Supervised Learning », *in 2023 31st European Signal Processing Conference (EUSIPCO)*, ISSN: 2076-1465, Sep. 2023, pp. 1170–1174. DOI: 10.23919/

EUSIPCO58844.2023.10290011. Accessed: Nov. 16, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10290011.

[176] C. Charles, « Introduction aux problèmes inverses », French, *Notes de Statistique et d'Informatique*, 2014, Publisher: Gembloux Agro-Bio Tech (GxABT). Unité de Statistique, Informatique et Mathématiques Appliquées (SIMa), Gembloux, Belgium. Accessed: Jul. 3, 2023. [Online]. Available: https://orbi.uliege.be/handle/2268/162659.

[177] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, « Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems », *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, *4*, pp. 586–597, Dec. 2007, ISSN: 1941-0484. DOI: 10.1109/JSTSP.2007.910281.

[178] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, « Deep Learning Techniques for Inverse Problems in Imaging », *IEEE Journal on Selected Areas in Information Theory*, vol. 1, *1*, pp. 39–56, May 2020, Conference Name: IEEE Journal on Selected Areas in Information Theory, ISSN: 2641-8770. DOI: 10.1109/JSAIT.2020.2991563.

[179] W. Crum, O. Camara, and D. Hill, « Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis », *IEEE Transactions on Medical Imaging*, vol. 25, *11*, pp. 1451–1461, Nov. 2006, Conference Name: IEEE Transactions on Medical Imaging, ISSN: 1558-254X. DOI: 10.1109/TMI.2006.880587. Accessed: Sep. 28, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1717643.

[180] V. M. Krasnopolsky and H. Schiller, « Some neural network applications in environmental sciences. Part I: forward and inverse problems in geophysical remote measurements », en, *Neural Networks*, Neural Network Analysis of Complex Scientific Data: Astronomy and Geosciences, vol. 16, *3*, pp. 321–334, Apr. 2003, ISSN: 0893-6080. DOI: 10.1016/S0893-6080(03)00027-3. Accessed: Jul. 3, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608003000273.

[181] M. Jenkinson and S. Smith, « A global optimisation method for robust affine registration of brain images », en, *Medical Image Analysis*, vol. 5, *2*, pp. 143–156, Jun. 2001, ISSN: 1361-8415. DOI: 10.1016/S1361-8415(01)00036-6. Accessed: Jun. 5,

2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841501000366.

[182] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, « Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images », en, *NeuroImage*, vol. 17, *2*, pp. 825–841, Oct. 2002, ISSN: 1053-8119. DOI: 10.1006/nimg.2002.1132. Accessed: Jun. 5, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811902911328.

[183] J. Jiang, P. Trundle, and J. Ren, « Medical image analysis with artificial neural networks », *Computerized Medical Imaging and Graphics*, vol. 34, *8*, pp. 617–631, Dec. 2010, ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2010.07.003. Accessed: Jun. 6, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611110000741.

[184] F. Pérez-García, R. Sparks, and S. Ourselin, « TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning », *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106 236, Sep. 2021, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2021.106236. Accessed: Jan. 30, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721003102.

[185] O. Ronneberger, P. Fischer, and T. Brox, « U-Net: Convolutional Networks for Biomedical Image Segmentation », en, *in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

[186] K. He, X. Zhang, S. Ren, and J. Sun, « Deep Residual Learning for Image Recognition », 2016, pp. 770–778. Accessed: Sep. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

[187] N. Otsu, « A threshold selection method from gray-level histograms », *IEEE transactions on systems, man, and cybernetics*, vol. 9, *1*, pp. 62–66, 1979.

[188] S. Jung and M. Keuper, « Internalized Biases in Fréchet Inception Distance », en, Dec. 2021. Accessed: Jul. 2, 2024. [Online]. Available: https://openreview.net/forum?id=mLG96UpmbYz.

[189] H. Sasaki, C. G. Willcocks, and T. P. Breckon, *UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models*, en, arXiv:2104.05358 [cs, eess], Apr. 2021. Accessed: Nov. 14, 2022. [Online]. Available: http://arxiv.org/abs/2104.05358.

[190] S. Sun, L. Wei, J. Xing, J. Jia, and Q. Tian, « SDDM: Score-Decomposed Diffusion Models on Manifolds for Unpaired Image-to-Image Translation », en, *in Proceedings of the 40th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Jul. 2023, pp. 33 115–33 134. Accessed: Sep. 27, 2024. [Online]. Available: https://proceedings.mlr.press/v202/sun23n.html.

[191] Y. Luo et al., « Target-Guided Diffusion Models for Unpaired Cross-Modality Medical Image Translation », *IEEE Journal of Biomedical and Health Informatics*, vol. 28, *7*, pp. 4062–4071, Jul. 2024, Conference Name: IEEE Journal of Biomedical and Health Informatics, ISSN: 2168-2208. DOI: 10.1109/JBHI.2024.3393870. Accessed: Sep. 27, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10508481/?arnumber=10508481.

[192] M. Özbey et al., « Unsupervised Medical Image Translation With Adversarial Diffusion Models », *IEEE Transactions on Medical Imaging*, vol. 42, *12*, pp. 3524–3539, Dec. 2023, Conference Name: IEEE Transactions on Medical Imaging, ISSN: 1558-254X. DOI: 10.1109/TMI.2023.3290149. Accessed: Sep. 27, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10167641/?arnumber=10167641.

[193] Y. Fan, H. Liao, S. Huang, Y. Luo, H. Fu, and H. Qi, « A survey of emerging applications of diffusion probabilistic models in MRI », *Meta-Radiology*, vol. 2, *2*, p. 100 082, Jun. 2024, ISSN: 2950-1628. DOI: 10.1016/j.metrad.2024.100082. Accessed: Sep. 27, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2950162824000353.

[194] V. M. H. Phan, Z. Liao, J. W. Verjans, and M.-S. To, « Structure-preserving synthesis: maskgan for unpaired mr-ct translation », *in International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 56–65.

[195] 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, Aug. 2013, Conference Name: IEEE Transactions

on Pattern Analysis and Machine Intelligence, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2013.50.

[196] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, « InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets », *in Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016. Accessed: Sep. 12, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html.

[197] J. Wu and S. Zhou, « A Disentangled Representations based Unsupervised Deformable Framework for Cross-modality Image Registration », *in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, ISSN: 2694-0604, Nov. 2021, pp. 3531–3534. DOI: 10.1109/EMBC46164.2021.9630778.

[198] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen, « Unsupervised Deformable Registration for Multi-modal Images via Disentangled Representations », en, *in Information Processing in Medical Imaging*, A. C. S. Chung, J. C. Gee, P. A. Yushkevich, and S. Bao, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 249–261, ISBN: 978-3-030-20351-1. DOI: 10.1007/978-3-030-20351-1_19.

[199] Z. Yu, Y. Zhai, X. Han, T. Peng, and X.-Y. Zhang, « MouseGAN: GAN-Based Multiple MRI Modalities Synthesis and Segmentation for Mouse Brain Structures », en, *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne et al., Eds., vol. 12901, Series Title: Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 442–450, ISBN: 978-3-030-87192-5 978-3-030-87193-2. DOI: 10.1007/978-3-030-87193-2_42. Accessed: Nov. 15, 2022. [Online]. Available: https://link.springer.com/10.1007/978-3-030-87193-2_42.

[200] K. Li, L. Yu, S. Wang, and P.-A. Heng, « Unsupervised Retina Image Synthesis via Disentangled Representation Learning », en, *in Simulation and Synthesis in Medical Imaging*, N. Burgos, A. Gooya, and D. Svoboda, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 32–41, ISBN: 978-3-030-32778-1. DOI: 10.1007/978-3-030-32778-1_4.

[201] N. Pawlowski, D. Coelho de Castro, and B. Glocker, « Deep Structural Causal Models for Tractable Counterfactual Inference », *in Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 857–869. Accessed: Dec. 5, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

[202] T. Karras, S. Laine, and T. Aila, « A Style-Based Generator Architecture for Generative Adversarial Networks », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410. Accessed: Sep. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

[203] H. Kim and A. Mnih, « Disentangling by Factorising », en, *in Proceedings of the 35th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Jul. 2018, pp. 2649–2658. Accessed: Sep. 8, 2022. [Online]. Available: https://proceedings.mlr.press/v80/kim18b.html.

[204] C.-S. Lai, Z. You, C.-C. Huang, Y.-H. Tsai, and W.-C. Chiu, « Colorization of Depth Map via Disentanglement », en, *in Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 450–466, ISBN: 978-3-030-58571-6. DOI: 10.1007/978-3-030-58571-6_27.

[205] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, « Image-to-image translation for cross-domain disentanglement », *in Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018. Accessed: Feb. 14, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/dc6a70712a252123c40d2adba6a11d84-Abstract.html.

[206] J. Kalkhof, C. González, and A. Mukhopadhyay, « Disentanglement Enables Cross-Domain Hippocampus Segmentation », *in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, ISSN: 1945-8452, Mar. 2022, pp. 1–5. DOI: 10.1109/ISBI52829.2022.9761560. Accessed: Dec. 5, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9761560.

[207] S. Chu, D. Kim, and B. Han, « Learning Debiased and Disentangled Representations for Semantic Segmentation », *in Advances in Neural Information Processing Systems*, vol. 34, Curran Associates, Inc., 2021, pp. 8355–8366. Accessed: Feb. 15,

2024. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/465636eb4a7ff4b267f3b765d07a02da-Abstract.html.

[208] J. Yang et al., « Domain-Agnostic Learning With Anatomy-Consistent Embedding for Cross-Modality Liver Segmentation », 2019, pp. 0–0. Accessed: Dec. 5, 2023. [Online]. Available: https://openaccess.thecvf.com/content_ICCVW_2019/html/VRMI/Yang_Domain-Agnostic_Learning_With_Anatomy-Consistent_Embedding_for_Cross-Modality_Liver_Segmentation_ICCVW_2019_paper.html.

[209] Y. Liu, S. J. Wagner, and T. Peng, « Multi-Modality Microscopy Image Style Augmentation for Nuclei Segmentation », en, *Journal of Imaging*, vol. 8, *3*, p. 71, Mar. 2022, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2313-433X. DOI: 10.3390/jimaging8030071. Accessed: Dec. 5, 2023. [Online]. Available: https://www.mdpi.com/2313-433X/8/3/71.

[210] Q. Meng et al., « Mutual Information-Based Disentangled Neural Networks for Classifying Unseen Categories in Different Domains: Application to Fetal Ultrasound Imaging », *IEEE Transactions on Medical Imaging*, vol. 40, *2*, pp. 722–734, Feb. 2021, Conference Name: IEEE Transactions on Medical Imaging, ISSN: 1558-254X. DOI: 10.1109/TMI.2020.3035424. Accessed: Dec. 5, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9247170.

[211] T. Van Steenkiste, D. Deschrijver, and T. Dhaene, « Interpretable ECG Beat Embedding using Disentangled Variational Auto-Encoders », *in 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, ISSN: 2372-9198, Jun. 2019, pp. 373–378. DOI: 10.1109/CBMS.2019.00081. Accessed: Dec. 5, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/8787396.

[212] A. Ben-Cohen, R. Mechrez, N. Yedidia, and H. Greenspan, « Improving CNN Training using Disentanglement for Liver Lesion Classification in CT », *in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, ISSN: 1558-4615, Jul. 2019, pp. 886–889. DOI: 10.1109/EMBC.2019.8857465. Accessed: Dec. 5, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8857465.

[213] E. Ferreira dos Santos and A. Mileo, « From Disentangled Representation to Concept Ranking: Interpreting Deep Representations in Image Classification Tasks », en, *in Machine Learning and Principles and Practice of Knowledge Discovery in*

Databases, I. Koprinska et al., Eds., ser. Communications in Computer and Information Science, Cham: Springer Nature Switzerland, 2023, pp. 322–335, ISBN: 978-3-031-23618-1. DOI: 10.1007/978-3-031-23618-1_22.

[214] A. Staffini, T. Svensson, U.-i. Chung, and A. K. Svensson, « A Disentangled VAE-BiLSTM Model for Heart Rate Anomaly Detection », en, *Bioengineering*, vol. 10, *6*, p. 683, Jun. 2023, Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2306-5354. DOI: 10.3390/bioengineering10060683. Accessed: Dec. 5, 2023. [Online]. Available: https://www.mdpi.com/2306-5354/10/6/683.

[215] X. Li, I. Kiringa, T. Yeap, X. Zhu, and Y. Li, « Anomaly Detection Based on Unsupervised Disentangled Representation Learning in Combination with Manifold Learning », *in 2020 International Joint Conference on Neural Networks (IJCNN)*, ISSN: 2161-4407, Jul. 2020, pp. 1–10. DOI: 10.1109/IJCNN48605.2020.9207046. Accessed: Dec. 5, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9207046.

[216] W.-Y. Lee and Y.-C. F. Wang, « Learning Disentangled Feature Representations For Anomaly Detection », *in 2020 IEEE International Conference on Image Processing (ICIP)*, ISSN: 2381-8549, Oct. 2020, pp. 2156–2160. DOI: 10.1109/ICIP40778.2020.9191201. Accessed: Dec. 5, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9191201.

[217] K. Do and T. Tran, « Theory and Evaluation Metrics for Learning Disentangled Representations », en, *in International Conference on Learning Representations*, Sep. 2019. Accessed: Mar. 27, 2024. [Online]. Available: https://openreview.net/forum?id=HJgK0h4Ywr.

[218] C. Eastwood and C. K. I. Williams, « A Framework for the Quantitative Evaluation of Disentangled Representations », en, *in International Conference on Learning Representations*, Feb. 2018. Accessed: Feb. 14, 2024. [Online]. Available: https://openreview.net/forum?id=By-7dz-AZ.

[219] A. Achille and S. Soatto, « Emergence of Invariance and Disentanglement in Deep Representations », en, *in 2018 Information Theory and Applications Workshop (ITA)*, San Diego, CA: IEEE, Feb. 2018, pp. 1–9, ISBN: 978-1-72810-124-8. DOI: 10.1109/ITA.2018.8503149. Accessed: Feb. 14, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/8503149/.

[220] D. P. Kingma and M. Welling, « Auto-Encoding Variational Bayes », en, *in 2014 International Conference on Learning Representations (ICLR)*, 2014. Accessed: Apr. 4, 2024. [Online]. Available: https://openreview.net/forum?id=33X9fd2-9FyZd.

[221] D. Rezende and S. Mohamed, « Variational Inference with Normalizing Flows », en, *in Proceedings of the 32nd International Conference on Machine Learning*, ISSN: 1938-7228, PMLR, Jun. 2015, pp. 1530–1538. Accessed: Feb. 15, 2024. [Online]. Available: https://proceedings.mlr.press/v37/rezende15.html.

[222] A. Tjandra, R. Pang, Y. Zhang, and S. Karita, *Unsupervised Learning of Disentangled Speech Content and Style Representation*, arXiv:2010.12973 [cs, eess], Jun. 2021. DOI: 10.48550/arXiv.2010.12973. Accessed: Mar. 13, 2024. [Online]. Available: http://arxiv.org/abs/2010.12973.

[223] S. Ioffe and C. Szegedy, « Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift », en, *in Proceedings of the 32nd International Conference on Machine Learning*, ISSN: 1938-7228, PMLR, Jun. 2015, pp. 448–456. Accessed: Sep. 8, 2022. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html.

[224] V. Dumoulin, J. Shlens, and M. Kudlur, « A Learned Representation For Artistic Style », *in International Conference on Learning Representations (ICLR)*, Apr. 2017, p. 9.

[225] E. Perez, F. Strub, H. d. Vries, V. Dumoulin, and A. Courville, « FiLM: Visual Reasoning with a General Conditioning Layer », en, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, *1*, Apr. 2018, Number: 1, ISSN: 2374-3468. DOI: 10.1609/aaai.v32i1.11671. Accessed: Sep. 12, 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11671.

[226] X. Jiang, C. Han, Y. A. Li, and N. Mesgarani, *Listen, Chat, and Edit: Text-Guided Soundscape Modification for Enhanced Auditory Experience*, arXiv:2402.03710 [cs, eess], Feb. 2024. DOI: 10.48550/arXiv.2402.03710. Accessed: Feb. 14, 2024. [Online]. Available: http://arxiv.org/abs/2402.03710.

[227] S. Saxena, M. Sharma, and O. Kroemer, « Multi-Resolution Sensing for Real-Time Control with Vision-Language Models », en, *in 7th Annual Conference on Robot Learning*, Aug. 2023. Accessed: Apr. 4, 2024. [Online]. Available: https://openreview.net/forum?id=WuBv9-IGDUA.

[228] Y. Tewel et al., *Training-Free Consistent Text-to-Image Generation*, arXiv:2402.03286 [cs], Feb. 2024. DOI: 10.48550/arXiv.2402.03286. Accessed: Feb. 8, 2024. [Online]. Available: http://arxiv.org/abs/2402.03286.

[229] Y.-H. Cao et al., « Patient-specific 4DCT respiratory motion synthesis using tumor-aware GANs », Milan, Italy, Nov. 2022. Accessed: Feb. 8, 2024. [Online]. Available: https://hal.science/hal-03811270.

[230] P. Zhao et al., *Audio-Infused Automatic Image Colorization by Exploiting Audio Scene Semantics*, arXiv:2401.13270 [cs], Jan. 2024. DOI: 10.48550/arXiv.2401.13270. Accessed: Feb. 8, 2024. [Online]. Available: http://arxiv.org/abs/2401.13270.

[231] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, « StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models », *in Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/3eaad2a0b62b5ed7a2e66c2188
Abstract-Conference.html.

[232] P. Esser, E. Sutter, and B. Ommer, « A Variational U-Net for Conditional Appearance and Shape Generation », *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8857–8866. Accessed: Feb. 15, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Esser_A_Variational_U-Net_CVPR_2018_paper.html.

[233] C. Li et al., « ALICE: Towards Understanding Adversarial Learning for Joint Distribution Matching », *in Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. Accessed: Mar. 15, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/ade55409d1224074754035a5a937d2e0-Abstract.html.

[234] R. D. Hjelm et al., « Learning deep representations by mutual information estimation and maximization », en, *in International Conference on Learning Representations*, Sep. 2018. Accessed: Mar. 15, 2024. [Online]. Available: https://openreview.net/forum?id=Bklr3j0cKX.

[235] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, « Unsupervised Feature Learning via Non-Parametric Instance Discrimination », *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742. Accessed: Mar. 15, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.html.

[236] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, « A Simple Framework for Contrastive Learning of Visual Representations », en, *in Proceedings of the 37th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Nov. 2020, pp. 1597–1607. Accessed: Mar. 15, 2024. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html.

[237] M.-A. Carbonneau, J. Zaïdi, J. Boilard, and G. Gagnon, « Measuring Disentanglement: A Review of Metrics », *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022, Conference Name: IEEE Transactions on Neural Networks and Learning Systems, ISSN: 2162-2388. DOI: 10.1109/TNNLS.2022.3218982. Accessed: Feb. 16, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9947342.

[238] R. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, « Isolating Sources of Disentanglement in Variational Autoencoders », *in Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018. Accessed: Sep. 27, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html.

[239] I. Higgins et al., « Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework », en, Jul. 2022. Accessed: Sep. 26, 2022. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl.

[240] S. Duan et al., « Unsupervised Model Selection for Variational Disentangled Representation Learning », en, Sep. 2019. Accessed: Mar. 27, 2024. [Online]. Available: https://openreview.net/forum?id=SyxL2TNtvr.

[241] P. Khosla et al., « Supervised Contrastive Learning », *in Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 18 661–18 673. Accessed: Mar. 31, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

[242] Y. Pang, J. Lin, T. Qin, and Z. Chen, « Image-to-Image Translation: Methods and Applications », *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2022, Conference Name: IEEE Transactions on Multimedia, ISSN: 1941-0077. DOI: 10.1109/TMM.2021.3109419.

[243] X. Liu, S. Thermos, G. Valvano, A. Chartsias, A. O'Neil, and S. A. Tsaftaris, « Measuring the Biases and Effectiveness of Content-Style Disentanglement », en, *in British Machine Vision Conference (BMVC)*, 2021.

[244] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, « Measuring and testing dependence by correlation of distances », *The Annals of Statistics*, vol. 35, *6*, pp. 2769–2794, Dec. 2007, Publisher: Institute of Mathematical Statistics, ISSN: 0090-5364, 2168-8966. DOI: 10.1214/009053607000000505. Accessed: Mar. 27, 2024. [Online]. Available: https://projecteuclid.org/journals/annals-of-statistics/volume-35/issue-6/Measuring-and-testing-dependence-by-correlation-of-distances/10.1214/009053607000000505.full.

[245] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv:1802.03426 [cs, stat], Sep. 2020. DOI: 10.48550/arXiv.1802.03426. Accessed: Apr. 28, 2024. [Online]. Available: http://arxiv.org/abs/1802.03426.

[246] C. Scavinner-Dorval, R. Bailly, B. Borotikar, S. Brochard, D. Ben Salem, and F. Rousseau, « Learning disentangled representations for unpaired synthesis of high-resolution dynamic mri », *Machine Learning for Biomedical Imaging*, vol. 3, pp. 16–37, February 2025 issue 2025, ISSN: 2766-905X. DOI: https://doi.org/10.59275/j.melba.2025-e8dc. [Online]. Available: https://melba-journal.org/2025:002.

[247] C. Scavinner-Dorval, R. Bailly, B. Borotikar, S. Brochard, D. B. Salem, and F. Rousseau, « Analysis of disentangled representation learning for high-resolution dynamic MRI synthesis », *in Medical Imaging 2024: Image Processing*, vol. 12926, SPIE, Apr. 2024, pp. 694–699. DOI: 10.1117/12.3006829. Accessed: Apr. 4, 2024. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12926/129262U/Analysis-of-disentangled-representation-learning-for-high-resolution-dynamic-MRI/10.1117/12.3006829.full.

[248] E. L. Denton and v. Birodkar, « Unsupervised Learning of Disentangled Representations from Video », *in Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. Accessed: Sep. 12, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/2d2ca7eedf739ef4c3800713ec482e1a-Abstract.html.

COLLEGES SCIENCES
BRETAGNE POUR L'INGENIEUR
LOIRE ET LE NUMERIQUE

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

Titre : Synthèse d'image par apprentissage pour l'IRM dynamique

Mot clés : Synthèse d'image, IRM dynamique, Paralysie cérébrale, Équin, Apprentissage Profond

Résumé : La paralysie cérébrale (PC) est la déficience motrice la plus fréquente chez les enfants. Elle résulte de lésions cérébrales irréversibles survenues au cours de la période périnatale. L'équin est la déformation la plus fréquente chez les enfants atteints de PC. Il se définit comme une incapacité de flexion dorsale du pied au-dessus du plantigrade, entraînant des déformations de la démarche et des os au cours de la croissance. Pour approfondir la compréhension de cette pathologie, le fonctionnement de l'articulation de la cheville et des muscles peut être observé via IRM dynamique. Cependant, l'acquisition d'une IRM dynamique à haute résolution requiert de longs temps d'imagerie et des mouvements répétés à vitesse constante, ce qui la rend difficile pour les patients souffrant de troubles musculo-squelettiques. Les temps d'acquisition peuvent être réduits en dégradant la qualité de l'image, réduisant ainsi l'interprétabilité. Aujourd'hui, les méthodes de synthèse d'image basées sur l'apprentissage profond constituent l'état de l'art en matière de super-résolution et d'amélioration de la qualité de l'imagerie médicale. Cette thèse explore le potentiel des méthodes d'apprentissage pour la synthèse de séquences RM dynamiques à haute résolution à partir de séquences à basse résolution. Nous proposons un modèle qui s'appuie sur des données d'IRM statiques pour améliorer la qualité des séquences d'IRM dynamiques à faible résolution dans un contexte non apparié. Cette approche permet une meilleure interprétation des séquences IRM dynamiques.

Title: Learning-based Image Synthesis for Dynamic MRI

Keywords: Image Synthesis, Dynamic MRI, Cerebral Palsy, Equinus, Deep Learning

Abstract: Cerebral palsy (CP) is the most prevalent motor impairment in children, resulting from irreversible brain injuries during the perinatal period. Equinus is the most common deformity among children with CP. It is defined as the inability to dorsiflex the foot above plantigrade, leading to gait and bone deformities with growth. To enhance understanding of the equinus pathology, ankle joint and muscle mechanics can be observed through dynamic MRI. However, acquiring high-resolution dynamic MRI requires long imaging times and repeated motions at constant speed, making it challenging for patients with musculoskeletal disorders. Acquisitions times can be reduced by lowering image quality, thereby reducing interpretability. Today, deep learning-based image synthesis methods have become the state-of-the-art in super-resolution and quality enhancement in medical imaging. This thesis examines the potential of learning-based methods for synthesizing high-resolution dynamic MR sequences from low-resolution sequences. We propose a model leveraging static MRI data to enhance the quality of low-resolution dynamic MR sequences in an unpaired setting. This method will offer a greater interpretability of the dynamic MR sequences.