Contraintes géométriques et topologiques pour la segmentation d'images médicales : approches hybrides variationnelles et par apprentissage profond

Zoé Lambert

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité Mathématiques appliquées et informatique

Préparée au sein de l'Institut National des Sciences Appliquées de Rouen Normandie

Contraintes géométriques et topologiques pour la segmentation d'images médicales: approches hybrides variationnelles et par apprentissage profond.

Présentée et soutenue par
Zoé LAMBERT

Thèse soutenue publiquement le 09/12/2022 devant le jury composé de		
M. Jean-François AUJOL	Professeur, Université de Bordeaux	Rapporteur
M. Olivier BERNARD	Professeur, INSA Lyon	Rapporteur
Mme Noémie DEBROUX	Maître de Conférences, Université de Clermont Auvergne	Examinatrice
M. Vincent DUVAL	Chargé de Recherche, INRIA Paris	Examinateur
M. Gilles GASSO	Professeur, INSA Rouen Normandie	Examinateur
Mme Carole LE GUYADER	Professeur, INSA Rouen Normandie	Directrice de thèse
Mme Caroline PETITJEAN	Professeur, Université Rouen Normandie	Directrice de thèse
M. Fabien PIERRE	Maître de Conférences, Université de Lorraine	Examinateur

Thèse dirigée par Mme Carole LE GUYADER, Laboratoire de Mathématiques de l'INSA de Rouen Normandie – LMI (EA 3226) et Mme Caroline PETITJEAN, Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes – LITIS (EA 4108).

INSA ROUEN NORMANDIE

ed 590 MIIS Mathématiques, Information, Ingénierie des Systèmes

LMI - EA 3226

Remerciements

Le doctorat est une sacrée aventure pour laquelle il faut être bien entouré. De nombreuses personnes ont contribué à l'aboutissement du mien et je tiens à leur adresser ces quelques mots.

En premier lieu, je souhaite exprimer mon infinie gratitude envers mes deux directrices de thèse, Carole et Caroline. Je me sens extrêmement chanceuse et honorée d'avoir travaillé à vos côtés pendant ces trois années. Non seulement, merci pour votre grande expertise scientifique chacune dans votre domaine, mais aussi pour votre disponibilité, vos encouragements, votre bienveillance, les discussions scientifiques tout comme celles plus informelles. Je n'aurais pas rêvé meilleur encadrement pour ma thèse.

Je remercie l'ensemble des membres du jury pour l'intérêt qu'ils ont porté à mes travaux, leur gentillesse à mon égard et les échanges scientifiques très enrichissants qui alimentent mes réflexions pour la suite. En effet, un immense merci à Jean-François Aujol et Olivier Bernard d'avoir accepté de rapporter cette thèse, ainsi qu'à Noémie Debroux, Gilles Gasso, Fabien Pierre et Vincent Duval pour avoir examiné ce travail de thèse.

Je tiens à remercier l'ensemble des membres du Laboratoire de Mathématiques de l'INSA de Rouen Normandie que j'ai pris plaisir à côtoyer tout au long de ces années. Notamment, un grand merci à Christian Gout (malgré ses goûts sportifs très discutables !) qui a toujours cru en moi et m'a ouvert les portes du monde de la recherche et de l'enseignement. Je n'oublie pas Nicolas Forcadel pour sa bonne humeur et son efficacité, ainsi que Brigitte Diarra pour sa prévenance et son aide inestimable concernant le parcours administratif complexe et la logistique. De plus, je témoigne toute mon amitié aux doctorants et post-doctorants du laboratoire, particulièrement à Rym, Timothée, Théau, Augustin, Nathan, Averil et Piero.

J'adresse maintenant des remerciements chaleureux aux enseignants de l'INSA Rouen Normandie qui m'ont confié la charge d'assurer leur TD. Je pense tout particulièrement à Jean-Marc Cabanial qui a d'abord été un professeur marquant, jouant un rôle prépondérant dans mon orientation universitaire, puis m'a accompagnée avec bienveillance lors de mes débuts dans l'enseignement.

Par ailleurs, ce travail n'aurait pu être mené à bien sans le Centre Régional Informa-

tique et d'Applications Numériques de Normandie (CRIANN) qui m'a permis de faire des simulations numériques sur de grands volumes de données. En particulier, je remercie Benoist Gaston pour l'aide précieuse quant à l'optimisation de mes codes de calculs. Merci également à Ghislain Lartigue qui m'a donné les clés afin d'améliorer l'implémentation de mes travaux. En outre, je remercie la Région Normandie qui a financé ce travail ainsi que le Centre Becquerel pour les échanges fructueux et la mise à disposition de données.

Si j'ai été particulièrement bien encadrée professionnellement, je n'aurais pas pu réussir cette thèse sans le soutien de mes proches.
Tout d'abord, celui de mes deux amies, Hélène et Claire, devenues mes collègues le temps d'une année après avoir elles-mêmes brillamment obtenu leur thèse. Vous compreniez mieux que personne les étapes à franchir, vous m'avez rassurée et motivée, vos partages d'expériences (de templates ou dossiers administratifs aussi) m'ont sans cesse mise sur les bons rails. Merci Clémence, d'abord collègue puis amie fidèle, tu as veillé à toujours remettre de l'essence dans mon moteur (sous forme de chocolat, ravioles, fondants et autres douceurs en tout genre), tu m'as écoutée (râler principalement) sans répit et aidée plus que tu ne l'imagines à ne jamais chavirer (excepté en canoë). Nos déjeuners et pauses-café/thé toutes ensemble ont illuminé mes journées.
Je pense également à tous mes amis proches qui ont largement œuvré à la réussite de cette thèse de par leur présence depuis tant d'années et jusqu'à la soutenance de ce doctorat ainsi que par leurs encouragements incessants. Merci Camille pour tes mots toujours bien choisis et ta confiance en moi, Mathilde et Caroline pour nos aventures chaque été qui m'ont redonné l'énergie de repartir année après année, Arthur pour ta bonne humeur communicative et ta légèreté qui m'ont souvent permis de prendre du recul mais aussi ta sollicitude quand il le fallait, Alice, Anaïs et Appoline pour les petites soirées indispensables pour décompresser. Merci Corentin, Ydriss, Victor, Simon, Audrey, Mathieu, PH, Romane, Juliette, Raphaël, Alexis, Isa, Flo, Belaf et tous les autres que je n'oublie pas. La vie est une fête à vos côtés.

Ces remerciements ne peuvent s'achever sans une pensée émue pour ma famille. Merci JE pour ta patience, ton écoute, tes questions et tes conseils qui auront déverrouillé nombreuses situations, et surtout pour ta gentillesse infinie qui rend plus doux les moments difficiles (ainsi que le frigo toujours rempli ces dernières semaines). Je n'oublie pas mes mamies, toutes deux incroyables et précieuses, votre amour m'a rendue plus forte. Je remercie mes deux petits frères, Martin et Arthur, votre humour a été la meilleure arme pour relever ce défi et tous les autres (surtout parce que je suis bon public, donc on se calme !), passer du temps avec vous s'apparente à une véritable thérapie. Enfin, ma reconnaissance incommensurable va à mes parents. Vous avez su éveiller ma curiosité scientifique dès le plus jeune âge et me guider tout en me laissant (presque toujours) la liberté de mes choix. En grands philosophes vous me disiez : *«Lorsqu'on commence un sport on va jusqu'au bout !»*, votre soutien et vos encouragements sans faille m'ont permis de ne jamais baisser les bras pour gagner ce match. Un immense merci à toi, papa, pour la relecture (nocturne) de ce manuscrit. Je vous dois beaucoup.

Résumé

La segmentation d'images constitue un traitement central de la vision par ordinateur, et particulièrement pour l'analyse d'images médicales. Lors de la planification d'un traitement par radiothérapie, il est nécessaire de segmenter la tumeur cible ainsi que les organes sains adjacents (appelés organes à risque). Si les réseaux de neurones convolutifs exhibent des segmentations précises, certains artefacts subsistent néanmoins (pixels isolés, trous etc.). Ainsi, l'inclusion d'informations *a priori* dans une tâche de segmentation, qu'il s'agisse de contraintes topologiques telles que le nombre de composantes connexes, la convexité partielle de la frontière d'un objet, ou de prescriptions géométriques via par exemple la pénalisation du volume par des contraintes, s'avère critique. Notamment, lorsqu'on souhaite préserver les relations contextuelles entre les objets et obtenir une segmentation homéomorphe à un *a priori* connu. Motivé par cette observation, ce travail de thèse vise à fournir un cadre hybride variationnel/apprentissage profond incluant des contraintes géométriques et topologiques dans l'apprentissage des réseaux de neurones convolutifs, sous la forme d'une pénalisation dans la fonction de perte. L'objectif réside dans l'amélioration de la qualité des segmentations d'images médicales, pour lesquelles les contours des objets à segmenter ne sont pas bien définis. Ainsi, un premier modèle inclut des contraintes géométriques par le biais d'une régularisation bâtie sur la variation totale pondérée, d'une pénalisation du volume/de l'aire, et d'un terme d'attache aux données de Mumford-Shah. Dans un second modèle, nous interprétons le processus de segmentation comme une tâche de recalage appariant la vérité terrain et l'image à étiqueter, fondée sur des principes d'élasticité non linéaire. L'application de conditions d'incompressibilité sur le déterminant de la matrice jacobienne de la déformation garantit la préservation du volume et de la topologie, sans auto-intersection de la matière. Des résultats théoriques soulignant la solidité mathématique des modèles sont fournis, ainsi qu'une analyse des algorithmes numériques appropriés basés sur une stratégie de séparation de variables et donnant des sous-problèmes qui admettent, pour la plupart, des solutions *closed form*. Les expériences sont principalement menées sur la base de données SegTHOR qui contient des scanners thoraciques de patients soignés par radiothérapie et dont on cherche à segmenter 4 organes à risque à préserver des rayons. Elles démontrent que nos méthodes apportent des améliorations significatives par rapport aux approches non contraintes existantes, à la fois en termes de critères quantitatifs tels que la mesure du chevauchement des régions et d'évaluation qualitative, en particulier lorsque les classes sont déséquilibrées.

Abstract

Image segmentation is a central process in computer vision, especially for medical image analysis. When planning a radiotherapy treatment, it is necessary to segment the target tumour as well as adjacent healthy organs (so-called organs at risk). Although convolutional neural networks exhibit accurate segmentations, some artefacts remain (isolated pixels, holes etc.). Thus, incorporating prior knowledge into a segmentation process, whether it be topological prescriptions such as the number of related components, the (partial) convexity of the boundary of an object, or geometrical constraints via, for example, the penalisation of the volume by constraints, is critical. In particular, when one wishes to preserve contextual relationships between objects and obtain a segmentation that is homeomorphic to a known a priori. Inspired by this observation, this thesis aims to provide a hybrid variational/deep learning framework including geometric and topological constraints in the training of convolutional neural networks, in the form of a penalty in the loss function. The objective is to improve the quality of medical image segmentations, for which the contours of the objects to be segmented are not well defined. Thus, a first model includes geometric constraints through a regularisation based on the weighted total variation, a volume/area penalty and a Mumford-Shah term. In a second model, we interpret the segmentation process as a registration task pairing the ground truth and the image to be labelled, based on non-linear elasticity principles. The application of incompressibility conditions on the determinant of the Jacobian matrix of the deformation guarantees preservation of volume and topology, without self-intersection of the material. Theoretical results highlighting the mathematical soundness of the models are provided, as well as an analysis of appropriate numerical algorithms based on a splitting strategy and yielding subproblems that admit, for the most part, closed form solutions. The experiments are mainly conducted on the SegTHOR database which contains thoracic CT scans of patients treated by radiotherapy and aim to segment 4 organs at risk to be preserved from radiation. They demonstrate that our methods provide significant improvements over existing unconstrained approaches, both in terms of quantitative criteria such as the measurement of region overlap and qualitative assessment, especially when the classes are unbalanced.

Table des matières

Introduction générale

La segmentation d'images consiste à identifier des constituants ou des régions sémantiquement significatifs d'une image donnée, généralement en vue d'une analyse quantitative ultérieure. Elle revêt un caractère critique dans l'analyse d'images médicales, essentielle pour le diagnostic assisté par ordinateur, la chirurgie guidée par l'image ou encore la planification de la radiothérapie. C'est ainsi qu'elle peut viser, par exemple, à détecter et localiser une tumeur pour en évaluer le volume ou le caractère cancéreux, ou encore des tissus mous et des organes à préserver des rayonnements.
Plus formellement, si on considère le domaine Ω de l'image, le problème de segmentation revient alors à partitionner l'image en sous-ensembles homogènes $S_k \subset \Omega$, dont l'union est le domaine entier Ω. Ainsi, les régions qui forment une segmentation doivent satisfaire

$$\Omega = \bigcup_{k=1}^{K} S_k$$

où $S_k \cap S_j = \emptyset$ pour $k \neq j$ et K désignant le nombre de classes ([109]).
Ce processus simple en apparence représente un véritable défi. Si l'être humain est aisément capable d'identifier les différentes structures d'une image en observant celle-ci dans son ensemble et en s'appuyant sur ses propres connaissances, cette aptitude visuelle apparaît beaucoup plus compliquée quand il s'agit de la formaliser mathématiquement, puis sur le plan algorithmique. En effet, l'objectif réside dans la reproduction du système de vision humain pour attribuer une étiquette à chaque pixel en se fondant uniquement sur des caractéristiques de bas-niveau de l'image telles que le niveau de gris, la couleur, la texture, le gradient, etc.

Depuis plusieurs décennies, la segmentation d'images constitue un traitement central de la vision par ordinateur, et de nombreux chercheurs se sont consacrés à l'étude de ce problème appréhendé dans le cadre d'un large spectre d'applications telles que l'astrophysique, le développement des véhicules autonomes qui nécessite la reconnaissance des passants et des autres voitures, ou encore la détection d'objet dans des scènes naturelles. Aujourd'hui, il n'existe pas de solution générale au problème de segmentation mais un ensemble de solutions adaptées à des problèmes et des images particuliers. Par conséquent, de nombreuses méthodes spécifiques à chaque application ont été développées. D'abord

des méthodes dites traditionnelles comme le seuillage, le filtrage, la croissance de région, les modèles déformables ou les modèles variationnels. Puis récemment, les méthodes d'apprentissage automatique, et notamment les réseaux de neurones convolutifs, font figure d'état de l'art pour toutes les tâches de segmentation d'images. Plus spécifiquement, en imagerie médicale, l'architecture U-Net et ses variantes apparaissent comme une référence en la matière. Toutefois, si ces méthodes offrent les meilleurs résultats pour des applications populaires, les segmentations obtenues manquent tout de même de précision, notamment au voisinage des frontières, et exhibent des erreurs aberrantes qui peuvent impacter négativement les décisions médicales.

Il est vrai que la détection et localisation automatique d'objets anatomiques est un sujet vaste et très complexe de par plusieurs aspects : (i) en premier lieu, la dépendance à la modalité des images. En effet, il existe une grande variété de modes d'acquisition des images, choisis en fonction de l'objectif de l'examen médical. Les plus communs sont, entre autres, l'imagerie par résonance magnétique (IRM), la tomodensitométrie (TDM ou CT-scan), la tomographie par émission de positons (TEP-scan), les rayons X, ou bien l'imagerie par ultrasons (IU) qui capturent des images de natures et de caractéristiques très différentes ; (ii) ensuite, la définition du mot "objet" est ambiguë et soumise à interprétation. Il peut s'agir d'une structure générale comme des tissus mous ou des organes, d'un ensemble de petites structures telles que les vaisseaux sanguins ou les cellules, ou encore d'une sous-partie d'un objet donné, par exemple une tumeur dans les poumons ou uniquement l'atrium dans le coeur ; (iii) et enfin, la forte variabilité intra- et inter-patients due à la singularité de chaque corps humain et du moment d'acquisition.
Pour résumer, les objets à segmenter sont pluriels, de natures et de structures complexes et variées mais aussi dépendants du mode et du moment d'acquisition ainsi que du patient lui-même. L'ensemble de ces éléments couplé au manque de contraste dans les images rendent particulièrement difficile l'opération.

Par conséquent, selon le type d'images et les régions d'intérêt concernées, les frontières de certaines structures anatomiques sont floues voire complètement inexistantes, au point d'aboutir à des erreurs de segmentation grossières et non conformes à la réalité anatomique (pixels isolés, trous, etc.). Pour tenter de corriger ces artefacts, à l'image des méthodes traditionnelles et pour retranscrire les connaissances des spécialistes, l'intégration de connaissances *a priori* dans un processus de segmentation par réseau de neurones profond constitue un champ actif de recherche. L'incorporation de ces informations peut prendre la forme de contraintes géométriques via une pénalisation de volume ou de forme, de prescriptions topologiques pour préserver les relations contextuelles entre les objets, ou par exemple de convexité (partielle) afin d'obtenir des segmentations plus homogènes/plausibles.

Ce travail de thèse s'inscrit dans cette problématique d'inclusion de connaissances préalables pour la segmentation automatique d'images médicales et nos contributions revêtent différentes formes :

1. de nature méthodologique en abordant cette inclusion dans un cadre hybride variationnel - apprentissage profond, nous permettant de tirer profit à la fois du caractère global et continu des méthodes variationnelles et des bonnes capacités de généralisation des méthodes Deep-Learning. Nous proposons ainsi deux modélisations portant sur l'inclusion de contraintes géométriques d'une part, et topologiques d'autre part, dans l'apprentissage des réseaux de neurones convolutifs sous la forme d'une pénalisation dans la fonction de perte ;
2. de nature théorique en explicitant notamment des résultats d'existence de solutions et de convergence garantissant le caractère bien posé de nos modèles ;
3. de nature plus appliquée, avec des évaluations approfondies sur des images de scanner thoracique, grâce à des algorithmes optimisés. En outre, les deux méthodes sont généralisables à d'autres applications. En particulier, nous expérimentons le second modèle sur un dataset composé d'images IRM cardiaques. Les résultats obtenus sont analysés à la fois quantitativement et qualitativement.

Le plan de ce manuscrit de thèse est énoncé ci-dessous avec un bref résumé des quatre chapitres qui le composent. Les deux premiers chapitres servent de préambule à nos contributions détaillées dans les deux derniers chapitres.

Chapitre 1 : État de l'art pour la segmentation d'images médicales
Ce chapitre passe d'abord en revue certaines méthodes phares pour la segmentation d'images médicales. Une attention particulière est portée aux réseaux de neurones profonds. En particulier, nous en décrivons les principaux éléments dont les couches de convolution, les couches de sous-échantillonnage et celles entièrement connectées, les fonctions d'activation, etc., ainsi que les architectures renommées et les fonctions de perte. Une section est dédiée à l'inclusion de connaissances *a priori*. Dans une seconde partie, nous nous focalisons sur le jeu de données SegTHOR (Segmentation of THoracic ORgans) rendu public au travers d'une compétition que nous avons organisée en 2019 afin d'encourager et d'inspirer la recherche sur cette application clinique. Nous introduisons sU-Net, une version simplifiée de l'architecture U-Net, qui nous sert de référence pour la segmentation automatique ce dataset et les expériences menées par la suite.

Chapitre 2 : Outils et rappels mathématiques
Dans cette section, des outils mathématiques nécessaires à l'élaboration et à l'analyse des modèles introduits sont présentés. Les propriétés de certains espaces fonctionnels, la théorie du calcul des variations, la notion d'élasticité non linéaire ou encore certains principes d'optimisation permettent de justifier la conception de ces modèles et leur caractère bien posé (existence de solution, résolution algorithmique réalisable, et résultat de conver-

gence).

Chapitre 3 : Inclusion de contraintes géométriques dans les réseaux de neurones convolutifs
Dans ce chapitre, nous présentons un cadre unifiant approches variationnelles et approches fondées sur le Deep-Learning, afin d'inclure des contraintes géométriques dans l'apprentissage des réseaux de neurones convolutifs, sous la forme d'une pénalisation dans la fonction de perte. Ces contraintes géométriques prennent plusieurs formes et incluent (i) l'alignement des contours entre la segmentation prédite et la vérité terrain via une régularisation bâtie sur la variation totale pondérée ; (ii) une pénalisation d'aire exprimée dans la modélisation par une contrainte d'égalité ; (iii) et enfin un critère d'homogénéité de l'intensité lumineuse des pixels fondé sur une combinaison du terme de Dice avec le terme d'attache aux données de Mumford-Shah. Le problème d'optimisation qui en résulte est résolu par l'introduction d'une variable auxiliaire, donnant lieu à deux sous-problèmes, et par un schéma alternatif. Deux méthodes sont étudiées, analysées et comparées pour résoudre le sous-problème lié aux *a priori* et sont complétées par des résultats théoriques. De plus, cette modélisation permet la mise en œuvre d'une implémentation optimisée qui s'appuie sur un procédé de parallélisation et un interfaçage GPU/CPU.

Chapitre 4 : Inclusion de contraintes topologiques dans les réseaux de neurones convolutifs
Ici, dans la même veine que le chapitre précédent mais pour pallier le problème de non-respect du nombre de composantes connexes, un cadre hybride variationnel/apprentissage profond incluant des contraintes géométriques et topologiques dans l'apprentissage des réseaux de neurones convolutifs est présenté. Pour ce faire, nous interprétons le processus de segmentation comme une tâche de recalage appariant la vérité terrain et l'image à étiqueter, fondée sur des principes d'élasticité non linéaire. L'application de conditions d'incompressibilité sur le déterminant de la matrice jacobienne de la déformation garantit la préservation du volume et de la topologie, sans auto-intersection de la matière. Des résultats théoriques soulignant la solidité mathématique du modèle sont fournis, parmi lesquels l'existence de minimiseurs du problème de minimisation introduit, montrant en particulier que la déformation recherchée est un homéomorphisme, ainsi qu'une analyse d'un algorithme numérique approprié qui s'appuie sur une stratégie de séparation et donnant des sous-problèmes qui admettent, pour la plupart, des solutions *closed form*. Une nouvelle fois, des efforts d'optimisation concernant l'implémentation sont produits ici.

Communications scientifiques

Journaux internationaux

1. Z. Lambert, C. Le Guyader, C. Petitjean. Enforcing geometrical priors in deep networks for semantic segmentation applied to radiotherapy planning. Journal of Mathematical Imaging and Vision (JMIV), pages 1-24., 2022, ([75]).

Conférences internationales

2. Z. Lambert, C. Le Guyader, C. Petitjean. On the inclusion of geometric and topological constraints in CNN. SIAM Conference on Imaging Science (IS22), Mars 2022, Berlin, Allemagne.
3. Z. Lambert, C. Le Guyader, C. Petitjean. A geometrically-gonstrained deep network for CT image segmentation. Dans 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pages 29-33. Avril 2021, Nice, France, ([74]).
4. Z. Lambert, C. Le Guyader, C. Petitjean. Analysis of the weighted Van der Waals-Cahn-Hilliard model for image segmentation. Dans 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1-6. Novembre 2020, Paris, France, ([73]).
5. Z. Lambert, C. Petitjean, B. Dubray, S. Ruan. SegTHOR : Segmentation of Thoracic Organs at Risk in CT images. Dans 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1-6. Novembre 2020, Paris, France, ([76]).

Conférences nationales

6. Z. Lambert, C. Le Guyader, C. Petitjean. Inclusion de contraintes géométriques et topologiques dans un CNN pour la segmentation d'images médicales. Dans Mini-symposium : Mathematical models and numerics for image processing au 45ème Congrès National d'Analyse Numérique (CANUM). Juin 2022, Evian-lès-Bains, France.

Acronymes

Nous listons ici les acronymes, en anglais, fréquemment utilisés tout au long de ce manuscrit.

DL	Deep-Learning
MLP	MultiLayer Perceptron
FCN	Fully Convolutional Network(s)
CNN	Convolutional Neural Network(s)
SGD	Stochastic Gradient Descent
ReLU	Rectified Linear Units
GT	Ground Truth
OAR	Organ(s) At Risk
CT	Computed Tomography
HD	Haudorff Distance
MHD	Mean Hausdorff Distance
MAE	Mean Absolute Error
ADMM	Alternating Direction Method of Multipliers
DR	Douglas-Rachford
PD	Primal-Dual
MS	Mumford-Shah
TV	Total Variation
WTV	Weighted Total Variation
CRF	Conditional Random Fields
DST	Discrete Sine Transform

CHAPITRE 1

La segmentation d'images médicales

1.1 Introduction

La segmentation d'images médicales a fait l'objet d'innombrables études ces dernières décennies. De nombreuses revues ont recensé et classifié, selon divers critères, les différentes méthodes dédiées à la segmentation d'images ([49, 149]), et plus spécifiquement à l'imagerie médicale ([109]). Dans [141] les méthodes se divisent en trois générations, où chacune ajoute un niveau supplémentaire de complexité algorithmique. La première génération regroupe les méthodes les plus simples qui s'appuyent sur des caractéristiques bas-niveau telles que le seuillage [106], la croissance de régions [103] et la détection de contours par des filtres (Robert, Prewitt, Sobel et Canny). Les modèles d'incertitude et d'optimisation constituent les ressorts de la deuxième. Enfin la dernière génération incorpore des connaissances préalables dans le processus de segmentation.
Nos travaux s'articulant sur des concepts de la deuxième et de la troisième génération de cette classification, nous n'abordons pas les méthodes de la première génération. Par conséquent, dans une première partie de ce chapitre, nous explicitons quelques méthodes répertoriées dans la deuxième. D'une part, les méthodes variationnelles dont les modèles de contours actifs et ceux construits sur la fonctionnelle de Mumford-Shah. D'autre part, les méthodes d'apprentissage automatique, et notamment les réseaux de neurones convolutifs. Nous rappelons le fonctionnement de ces réseaux et les éléments fondamentaux : les couches connectées, les fonctions d'activation, les fonctions de perte. Nous définissons des métriques d'évaluation pour les résultats de segmentation qui permettent d'en faire leur analyse quantitative. Suite à cela, nous donnons un aperçu non exhaustif des méthodes incluant des informations *a priori*.
Dans une seconde partie, nous présentons quelques jeux de données disponibles en imagerie médicale, et en particulier un nouveau dataset, SegTHOR, dédié à la segmentation d'organes à risque (OAR) dans des images de scanner du thorax en prévision d'un traitement par radiothérapie. Pour terminer, nous introduisons un réseau de neurones convolutif simple mais fonctionnel pour cette application [76].

1.2 État de l'art pour la segmentation d'images médicales

1.2.1 Méthodes variationnelles

Une approche variationnelle consiste d'une part en la modélisation d'une réalité physique sous la forme d'un problème d'optimisation, et d'autre part en la mise en oeuvre d'une méthode d'optimisation, autrement dit, la méthode numérique nécessaire à la résolution du problème introduit. Le formalisme variationnel se voit largement utilisé dans un objectif de segmentation, mais également pour divers autres types de traitement d'image. Pour ce faire, l'image I étudiée est appréhendée comme un continuum $I : \Omega \mapsto \mathbb{R}$, où Ω est un sous-ensemble de $\mathbb{R}^d$ (avec $d = 2$ pour une image plane et $d = 3$ pour une image volumique), et non comme un ensemble discret de pixels ou de voxels [110]. La fonction de segmentation recherchée $u : \Omega \mapsto \mathbb{R}$ correspond au(x) minimiseur(s) d'une fonctionnelle construite E telle que

$$u^* = \arg\min_{u \in \mathcal{C}} E(u), \tag{1.1}$$

avec $\mathcal{C}$ l'espace fonctionnel des segmentations admissibles. Une étape clé réside dans la conception de cette fonctionnelle. Elle combine communément un terme dit d'attache aux données ou de fidélité $\mathcal{F}_{id}$ avec un terme de régularisation $\mathcal{R}$ portant sur la segmentation recherchée, de sorte que

$$E(u) = \mathcal{F}_{id}(u, I) + \lambda \mathcal{R}(u), \tag{1.2}$$

avec $\lambda > 0$ le facteur de pondération entre les termes. Le terme de fidélité aux données mesure la proximité entre l'image et la segmentation. Celui des moindres carrés correspond à un choix assez classique. La régularisation agit comme une connaissance *a priori* sur la solution à reconstruire. Elle peut porter sur la fonction de segmentation u, sur ses dérivées et même associer ces éléments. Nous cherchons alors à construire une fonctionnelle qui soit évidemment adaptée au modèle observé, mais pour laquelle la différentiabilité et la convexité apparaissent comme des critères fortement désirables, afin de formuler un problème d'optimisation bien posé. Dans ce cadre, la recherche du minimum repose sur le calcul des variations. Les équations d'Euler-Lagrange constituent un outil de base pour ce type de problème et se résolvent généralement par une descente de gradient. Nous distinguons deux types d'approches : d'une part celles s'aidant des informations encodées par le gradient de l'image pour en détecter les contours, et celles s'appuyant sur des informations d'intensité de l'image pour en dégager des régions homogènes d'autre part.

Les méthodes basées sur les contours

Dans ce type de modèle, il s'agit de faire évoluer un contour vers les frontières de l'objet d'intérêt contenu dans l'image. En ce sens, Kass, Witkin et Terzopoulos [68] proposent le premier modèle de contour actif paramétrique, aussi appelé *snake*, où la courbe est représentée explicitement via des splines. Dans un contexte 2D, notant l'image I, on considère

$C = (C_1, C_2) : [0,1] \mapsto \mathbb{R}^2$ et $s \mapsto C(s)$ une courbe paramétrée. Pour faire progresser le contour, les auteurs formulent un problème de minimisation de la fonctionnelle E suivante

$$E(C) = \alpha \int_0^1 |C'(s)|^2 \, ds + \beta \int_0^1 |C''(s)|^2 \, ds + \lambda \int_0^1 g(|\nabla I(C(s))|) \, ds, \tag{1.3}$$

avec α, β et λ des termes de pondération positifs.
L'idée consiste à déformer une courbe initiale soumise à des forces internes de régularisation (deux premiers termes) et des forces externes induites par les données (troisième terme). Les forces internes préconisent un contour assez lisse par pénalisation de la longueur et de la courbure. Celles externes attirent le contour vers les frontières de l'objet, où se situent les zones de large gradient. De fait, la fonction g joue le rôle de détecteur de contours et dans [68], Kass *et al.* choisissent $g(|\nabla I|) = -|\nabla I|^2$.
Ce modèle pionnier et très populaire a ouvert la voie à de nombreux modèles déformables paramétriques puis géométriques, qui ont suscité un engouement particulier en imagerie médicale [92]. En effet, deux limitations émergent de ces premiers modèles : (i) la représentation explicite du contour ne permet pas de changement automatique de topologie du contour initial, et (ii) une sensibilité à l'initialisation et au bruit dû au fait de la non-convexité de la fonctionnelle et au caractère local lié à l'utilisation du gradient de l'image, comme souligné dans [137, Chapitre 9]. Dans ses travaux, Cohen [33] insère une force d'inflation au modèle *snake* pour le rendre moins sensible aux artefacts, et par conséquent à l'initialisation. Cette méthode est expérimentée pour extraire le ventricule d'images ultrasons du coeur.

La représentation implicite des courbes comme ensemble de niveau d'une fonction [105], dite représentation *level-set*, dans les modèles de contours actifs géométriques, apporte une solution à la difficulté relative au changement de topologie. Ainsi, le modèle géodésique introduit par Caselles *et al.* [17] cherche la courbe minimisant la fonctionnelle

$$E(C) = \int_0^1 |C'(s)| \, g(|\nabla I(C(s))|) ds, \tag{1.4}$$

et apparaît comme une variante intrinsèque de la fonctionnelle (1.3) (dans le sens où un changement de paramétrisation du contour ne modifie pas l'expression de l'énergie), avec $\beta = 0$ et en rassemblant les forces internes et externes. La minimisation de cette fonctionnelle revient à rechercher une courbe de longueur minimale dans un espace riemannien et dans lequel la métrique dépend de l'image I.
Cette variété de méthodes se montre intéressante pour la segmentation d'images médicales. Par exemple, Yezzi *et al.* [144] utilisent un modèle déformable géométrique pour détecter les contours du myocarde au sein d'images IRM cardiaques, d'un kyste à partir d'une image échographique, du sein ou encore d'os dans des images de scanner. Dans [90], Malladi *et al.* comparent des approches 2D et 3D de modèles déformables géométriques pour la segmentation de tissus mous dans des CT-scans de deux cuisses, ainsi que des

cavités cardiaques dans des images IRM, lors du cycle diastolique et systolique du cœur.

L'évolution de la courbe vers un minimum global dépend toujours fortement de l'initialisation et l'algorithme peut rester coincé dans un minimum local. En effet, les images montrant des artefacts et des objets dont les contrastes apparaissent émoussés, présentent une difficulté majeure pour ces méthodes. Il convient alors de s'appuyer sur des propriétés plus globales de l'image telles que l'intensité des régions qui la composent.

Les méthodes basées sur les régions

Cette classe de méthodes s'appuie donc sur la similarité entre pixels/voxels pour partitionner une image en régions homogènes. Dans la littérature, l'un des problèmes de minimisation les plus populaires, et parmi les plus étudiés, est celui de la fonctionnelle conçue par Mumford et Shah [98]

$$E(u, K) = \mu \int_{\Omega} (u - I)^2 \, dx + \int_{\Omega \setminus K} |\nabla u|^2 \, dx + \mathcal{H}(K), \tag{1.5}$$

où $\mathcal{H}$ est la mesure de Hausdorff, u est une approximation lisse par morceaux de l'image initiale I et K constitue l'ensemble des discontinuités. Le processus de segmentation repose sur deux principes : (i) l'image I peut être partitionnée en régions au sein desquelles l'intensité lumineuse varie peu, et (ii) l'image I varie de manière discontinue ou abrupte à travers les frontières K des régions. La minimisation de cette fonctionnelle se révèle particulièrement complexe de par sa non-convexité et de par les natures distinctes des inconnues, l'une étant fonctionnelle, l'autre constituée d'un ensemble de courbes. Dans [3], Ambrosio et Tortorelli résolvent le problème par le biais d'une approximation elliptique et fournissent un résultat de Γ-convergence.

Par ailleurs, un cas particulier de la formulation (1.5) apparaît lorsque l'énergie E se restreint aux fonctions u constantes par morceaux, c'est-à-dire, $u = c_i$ une constante dans chaque sous-ensemble ouvert Ω_i du domaine Ω. Ce problème de partition minimale revient à minimiser

$$E(u, K) = \mu \sum_i \int_{\Omega_i} (u - I)^2 \, dx + \mathcal{H}(K). \tag{1.6}$$

Pléthore de travaux se livre à l'analyse de cette fonctionnelle. Notamment, Chan et Vese [25] en proposent une approximation fondée sur les ensembles de niveaux, d'abord pour partitionner l'image en deux phases (le fond et l'objet), puis en la généralisant à plusieurs phases [136]. Chambolle *et al.* [19] opèrent une convexification pour résoudre ce problème de minimisation. Dans [126], Storath et Weinmann présentent une approche régularisée par une norme L^0 de la fonctionnelle, pour pénaliser les sauts, et proposent une résolution alternée pour solutionner ce problème de minimisation . Comme évoqué précédemment, la difficulté liée à la non-unicité d'une solution persiste et les méthodes risquent de converger

vers un minimum local. Dans le cas binaire, Chan *et al.* [24] proposent une approche pour surmonter ce problème et obtenir un résultat de minimum global de la formulation implicite introduite dans [25]. Dans [14], Bresson *et al.* s'attellent à fournir un cadre de minimisation globale pour des modèles basés d'une part, sur celui des contours actifs sans bord [25] et d'autre part, sur l'approximation lisse par morceaux de la fonctionnelle de Mumford-Shah. Des expériences, en particulier sur des images médicales, montrent des segmentations plus précises.
En outre, dans [26], les auteurs appliquent leur modèle de contours actifs fondé sur la fonctionnelle de Mumford-Shah constante par morceaux à la segmentation de tissus osseux ou encore d'une tumeur cérébrale. Tsai *et al.* [132] abordent le paradigme de Mumford-Shah du point de vue de l'évolution de courbes pour réaliser la segmentation d'une image pathologique du cerveau humain et d'une image échographique doppler du cœur.

1.2.2 Méthodes d'apprentissage profond/automatique

Dans cette partie, nous étudions les méthodes d'apprentissage profond qui ont largement été explorées dans une perspective de segmentation ([54, 94]), en particulier pour l'imagerie médicale ([5, 86]). En effet, ces dix dernières années, les méthodes traditionnelles de segmentation d'images ont été dépassées par celles fondées sur l'apprentissage profond (*Deep-Learning*), et notamment par les réseaux de neurones convolutifs (CNN, *Convolutional Neural Network*). Initialement développés à des fins de classification, les CNN se sont largement étendus aux autres tâches de vision par ordinateur et la segmentation d'images ne fait pas exception.

Ces méthodes d'apprentissage se divisent couramment en trois catégories : supervisées, semi/faiblement supervisées et non supervisées. Notre attention se porte principalement sur celles supervisées. Dans ce type d'approches, on considère un dataset $\mathcal{D} = \left\{(x^k, y^k)_{k=1}^K\right\}$, avec x une donnée associée à y son étiquette, également appelée vérité terrain (souvent notée GT, *Ground Truth*), et K le nombre d'observations. L'objectif vise à optimiser les paramètres θ d'un modèle f pour le traitement souhaité, de telle sorte que $f(x;\theta)$ donne la prédiction la plus proche possible de y pour l'entrée observée x. Dans le cadre de la segmentation sémantique, x est une image et l'on souhaite affecter une étiquette à chaque pixel ou voxel. La vérité terrain se traduit par une grille de segmentation discrète $\mathcal{G}$ où chaque noeud est classifié. En comparaison, dans un problème de classification d'images classique, l'étiquette est un scalaire.

Nous détaillons maintenant l'ensemble des innovations et des astuces qui ont conduit aux réseaux de neurones convolutifs, très efficaces pour la segmentation sémantique. Dans un premier temps, nous revenons brièvement sur les premiers réseaux de neurones, et spécifiquement le Perceptron multicouche, point de départ de l'apprentissage profond. Cela nous permet d'aborder ensuite les réseaux de neurones convolutifs, d'abord en vue d'une

classification d'images, puis étendus à la segmentation. En particulier, nous explicitons les concepts fondamentaux qui forment ces méthodes : les différents types de couches connectées, les fonctions d'activation et de perte. Les métriques d'évaluation qui permettent de quantifier les résultats de segmentation obtenus sont également décrites.

Réseaux de neurones profonds

Plusieurs avancées majeures ont permis le développement des réseaux de neurones profonds tels que nous les connaissons aujourd'hui. Initialement, l'idée consiste à essayer de modéliser le système nerveux pour concevoir des réseaux de neurones artificiels. Historiquement, le Perceptron formulé par Rosenblatt [113], qui simule l'activation d'un neurone biologique, apparaît comme une première révolution. Formellement, le Perceptron est un modèle f de classification binaire paramétré par $\theta = (w, b)$, où w joue le rôle des poids synaptiques et b est un biais (ou seuil). Ce modèle prend en entrée les différentes valeurs d'une donnée notée x, de dimension n, et donne en sortie la valeur 1 si le neurone est activé, 0 sinon. Par conséquent, ce modèle décrivant le fonctionnement d'un neurone est simplement défini par :

$$f(x; \theta) = \sigma(w^T x + b),$$

avec $x \in \mathbb{R}^n$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ et σ la fonction d'activation non linéaire appliquée localement. Cette fonction de décision dessine une droite pour séparer les données en deux classes et les paramètres θ en donnent les caractéristiques. On recherche donc ces paramètres qui caractérisent au mieux le problème par le biais d'un algorithme itératif *forward-backward*. Cela se traduit par une passe-avant *(forward)* qui estime la réponse attendue y à partir de l'échantillon x, telle que $\hat{y} = f(x; \theta)$. La construction d'une fonctionnelle d'erreur à minimiser $\mathcal{L}(\theta)$ par rapport aux paramètres est alors nécessaire. Elle doit permettre de mesurer la dissimilarité entre l'estimation obtenue $\hat{y}$ et la vérité terrain y. La passe-arrière (*backward*) permet finalement d'actualiser les paramètres par une méthode de descente de gradient définie par

$$\begin{aligned} w^{t+1} &\leftarrow w^t - \eta \nabla_w \mathcal{L}(\theta) \\ b^{t+1} &\leftarrow b^t - \eta \nabla_b \mathcal{L}(\theta), \end{aligned}$$

avec η le taux d'apprentissage et t l'itération considérée.

Le Perceptron multicouche (MLP, *MultiLayer Perceptron*) se révèle être une évolution de ce modèle. Un signal d'entrée x se trouve maintenant connecté à plusieurs neurones qui forment ce qu'on appelle une couche. Chaque neurone de cette couche transmet alors une certaine activation relativement à l'importance de l'information qu'il véhicule. Plusieurs couches intermédiaires, nommées couches cachées, peuvent s'empiler jusqu'à la couche de sortie finale. Chaque neurone d'une couche $l - 1$ se connecte à tous les neurones de la couche l. Cette innovation en connectant plusieurs couches successives de neurones les

unes aux autres forme l'un des premiers réseaux de neurones profonds que l'on modélise, en suivant les notations de [86], par

$$f(x;\theta) = \sigma(W^L\sigma(W^{L-1}...\sigma(W^0x + b^0) + b^{L-1}) + b^L),$$

avec W^l une matrice composée des poids w^n et L la profondeur du réseau. La force de ce réseau réside dans sa capacité à représenter une structure complexe des données d'entrée. L'entraînement du MLP bénéficie de la popularisation par [116] de la rétropropagation du gradient de l'erreur. Cet algorithme, comme son nom l'indique, permet de propager l'erreur du gradient à travers l'ensemble des couches du réseau grâce à la dérivation des fonctions composées (*chain rule*). Le fonctionnement efficace d'un réseau profond repose sur un choix judicieux de sa profondeur, du nombre de neurones composant chaque couche cachée ainsi que des fonctions d'activation. Ces choix s'effectuent généralement empiriquement afin d'obtenir un réseau performant tout en veillant à ce qu'il conserve son caractère optimisable. En effet, un premier souci inhérent aux réseaux de neurones profonds, en lien avec la quantité très importante de paramètres entraînables, se trouve dans l'explosion ou l'annulation du gradient. Par ailleurs, plus cette quantité augmente, plus l'apprentissage d'un tel modèle requiert une puissance de calcul et une capacité de mémoire conséquentes. Bien que des techniques de régularisation existent pour contrer ces effets, le traitement d'une image fait exploser le nombre de paramètres. Les réseaux de neurones convolutifs, ainsi que l'utilisation de cartes graphiques, font figure de solutions à la fois algorithmiques et matérielles dans ce contexte.

Réseaux de neurones convolutifs

De façon concomitante à l'invention de Rosenblatt, les études du cortex visuel d'Hubel et Wiesel [64] démontrent que les neurones de cette région du cerveau ne réagissent qu'à un stimulus visuel issu d'une région limitée du champ visuel, appelée champ récepteur. Les champs récepteurs se chevauchent pour paver l'ensemble du champ visuel. Ces observations, combinées aux prémices de la méthode de rétropropagation, donnent naissance au Neocognitron [50], l'ancêtre des réseaux de neurones convolutifs. La publication de [79] scelle l'avènement des CNN avec la présentation de LeNet-5 (Figure 1.1), un réseau destiné à la reconnaissance de documents manuscrits qui conjugue l'extraction de caractéristiques à leur classification, via des couches entièrement connectées de type MLP, dans un seul processus. De fait, comme illustré en partie par la Figure 1.1, le CNN empile différents types de couches : des couches de convolution et de sous-échantillonnage, qui constituent une nouveauté, et plus classiquement des couches d'activation non linéaires et entièrement connectées. Nous explicitons maintenant ces quelques éléments.

Couche de convolution Cette opération, élément central du réseau, simule les champs récepteurs du cortex visuel. Les connexions d'un neurone se limitent dorénavant aux élé-

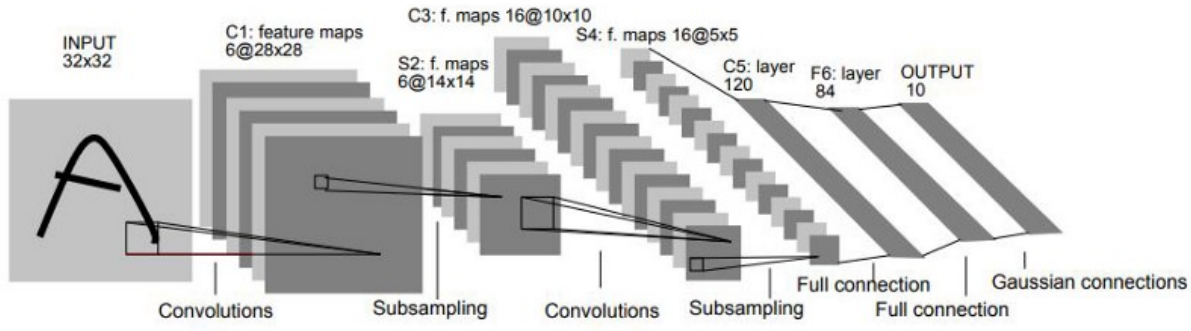


FIGURE 1.1 – Architecture LeNet-5 proposée par [79].

ments dans son champ récepteur, et plus à tous les éléments de l'image ou de la couche précédente. L'idée consiste à faire glisser un ensemble de filtres (aussi appelé noyaux de convolution), de taille pré-définie, afin d'effectuer une convolution sur une information en entrée. Chaque filtre se compose de plusieurs coefficients et correspond à des caractéristiques spécifiques, plus ou moins abstraites (lignes, contours, formes) dont on recherche la présence. Le terme de convolution est en réalité un abus de langage puisqu'en pratique on utilise la corrélation croisée. Dans le cadre discret, on considère une entrée I avec (i, j) ses pixels, soit F un filtre de dimension (m, n), la couche de convolution se définit par l'opération :

$$(F * I)(i, j) = \sum_m \sum_n F(m, n) I(i + m, j + n) \tag{1.7}$$

Le résultat prend la forme d'une carte de caractéristiques (*feature map*), associée à F, dont les éléments quantifient la présence des caractéristiques en question, et les localisent. À chaque couche de convolution, k filtres interviennent indépendamment pour former autant de cartes de caractéristiques en sortie. Au cours de l'apprentissage, les coefficients des différents filtres, ainsi que les biais, sont ajustés afin d'extraire automatiquement les caractéristiques les plus discriminantes pour les données observées.
Les filtres sont de taille impaire, communément 3×3, 5×5 voire 7×7. Dans [84], les auteurs utilisent des noyaux de convolution 1×1 pour fusionner l'information concentrée dans les différents canaux. Cette astuce permet de réduire la dimensionnalité et donc la quantité de paramètres, en projetant simplement les informations contenues dans les cartes caractéristiques via une combinaison linéaire.
Le partage des paramètres incarne la grande force des couches de convolution. En effet, si un filtre apparaît efficace pour la reconnaissance d'une certaine caractéristique localisée à un endroit particulier dans l'image reçue, alors il l'est également dans l'ensemble de l'image. Par conséquent, les coefficients d'un filtre restent identiques et ceci a pour effet de réduire drastiquement la quantité de paramètres.

Fonction d'activation Dans les réseaux de neurones, la fonction d'activation interprète le potentiel d'activation d'un neurone biologique, c'est-à-dire qu'elle décide de transmettre

ou non une information si le seuil d'activation est atteint, entraînant alors une réponse du neurone. Généralement, cette fonction mathématique est non linéaire et appliquée élément par élément directement après une opération de convolution. La non-linéarité permet au réseau de représenter une structure de données plus complexe, c'est-à-dire qui ne sont pas simplement linéairement séparables. Par conséquent, la fonction d'activation confère un grand pouvoir discriminant au réseau. Au commencement de l'apprentissage profond, un choix traditionnel se porte sur une fonction sigmoïdale telle que la fonction logistique (ou sigmoïde) définie par $\sigma(z) = \frac{1}{1+e^{-z}}$, pour chaque entrée z, avec $\sigma : \mathbb{R} \mapsto [0; 1]$, ou bien, comme dans LeNet-5 [79], sur la fonction tangente hyperbolique définie par $\sigma(z) = \tanh(z)$, avec $\sigma : \mathbb{R} \mapsto [-1; 1]$. Cependant, dans [55], les auteurs explicitent que le caractère saturant de ces fonctions pousse à la disparition et l'explosion du gradient lors de l'optimisation par rétropropagation. Pour parer à cet effet, [99] amène une solution aujourd'hui très populaire, la fonction ReLU (*Rectified Linear Units*) linéaire par morceaux et définie par $\sigma(z) = max(0, z)$, avec $\sigma : \mathbb{R} \mapsto [0; +\infty[$. Malgré tout, le risque de saturation des valeurs négatives peut persister et des variantes, dont PReLU [60] et ELU [31], sont proposées.

Couche de sous-échantillonnage L'opération déterministe de sous-échantillonnage (*subsampling*) sert à réduire progressivement la dimension spatiale de l'image d'entrée, et intervient à la suite d'un bloc de convolution. Cette couche agit localement, et indépendamment sur chaque canal, via une petite fenêtre glissante de taille fixe, qui balaye l'entrée et agrège les valeurs comprises dans la fenêtre par une statistique de type moyenne ou maximum. Ainsi, aucun apprentissage de paramètres ne s'opère ici.
Une fenêtre de taille 2×2 conjointement à un pas de 2 et le maximum comme fonction d'agrégation (*maxpooling*) représentent le choix le plus commun, divisant par 4 la surface des cartes de caractéristiques. Cette étape permet également d'extraire les caractéristiques les plus discriminantes tout en réduisant la quantité de paramètres dans les couches suivantes, et diminue ainsi le risque de sur-ajustement des données. La contrepartie de ce gain se manifeste par une perte importante de contexte global.

Couche entièrement connectée Une fois les caractéristiques extraites, le réseau doit les analyser en bout de chaîne : c'est le rôle des couches entièrement connectées. Toutes les cartes de caractéristiques finales sont redimensionnées en un seul vecteur dont chaque valeur devient l'entrée d'un réseau de type MLP. S'il faut déterminer la profondeur et le nombre de neurones par couche, en revanche l'espace de sortie de l'ultime couche se fixe par rapport à la tâche considérée. Par exemple, pour LeNet-5 schématisé sur la Figure 1.1, la classification s'effectue entre 10 classes et par conséquent le MLP compte autant de neurones de sortie.

Couche softmax Cette couche au sommet du réseau agit comme la fonction d'activation finale et permet de générer des probabilités d'appartenance à chaque classe considérée.

Dans le cadre d'un problème à C classes, la dernière couche entièrement connectée détient donc autant de neurones de sortie et la fonction softmax, $s : \mathbb{R} \mapsto [0; 1]$, se définit par :

$$s(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}},$$

où x est le vecteur de sortie de dimension C et $s(x)_i$ désigne la probabilité d'appartenance à la $i^{\text{ième}}$ classe.

Un axe de recherche consiste à trouver le meilleur agencement de ces éléments pour obtenir l'architecture la plus performante, aussi bien d'un point de vue de la précision des résultats que de la faisabilité et de l'optimisation de la méthode. Le réseau AlexNet [71] démocratise l'utilisation des réseaux de neurones convolutifs pour la classification d'images en gagnant largement la compétition ImageNet ILSVRC en 2012. Cette architecture réinvestit celle de LeNet-5 [79] en augmentant un peu la profondeur par l'ajout de couches de convolution successives, sans intervertir entre elles de couches de sous-échantillonnage. De plus, ils adoptent ReLU pour fonction d'activation. D'autres innovations contribuent à l'expansion, l'efficience, et la popularité des CNN en vision par ordinateur. En particulier, l'architecture ResNet [61] qui introduit les connexions résiduelles pour entraîner efficacement un réseau très profond et remporte à son tour le défi ILSVRC 2015. Un bloc résiduel se compose de deux couches de convolution qui produisent un signal en sortie, auquel s'ajoute celui reçu en entrée. Ceci a pour effet d'améliorer la rétropropagation du gradient.
Maintenant que les notions élémentaires de l'apprentissage profond par CNN pour des problèmes de classification d'images sont élucidées, nous nous intéressons à l'extension de ces travaux au cas de la segmentation d'images, spécifiquement médicales et biomédicales.

Architectures pour la segmentation

Une des premières approches permettant la segmentation automatique de bout-en-bout est le réseau entièrement convolutif (FCN, *Fully Convolutional Network*) schématisé sur la Figure 1.2 et introduit par [88]. Ce type de réseau reprend tout à fait les principes des CNN destinés à la classification mais en remplaçant le MLP en bout de réseau par des opérations de déconvolution, aussi appelées convolution transposée, afin de sur-échantillonner les caractéristiques apprises et projetées dans un sous-espace latent, de façon à retrouver les dimensions spatiales de l'image initiale. Le FCN a tendance à produire des segmentations assez grossières du fait de la perte d'informations et de contexte liée aux différentes couches de sous-échantillonnage. Dans [88], les auteurs proposent de fusionner plusieurs prédictions de segmentation, issues d'un sur-échantillonnage de cartes de caractéristiques à différentes profondeurs, dans le but de réintroduire l'information perdue.

Les FCN ouvrent la voie à une nouvelle structure de réseau nommée encodeur-décodeur. La plus populaire est l'architecture U-Net [112] initialement conçue pour des applications

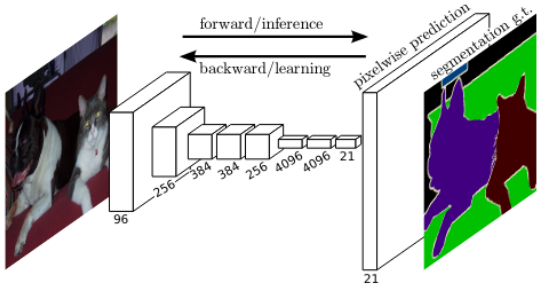


FIGURE 1.2 – Architecture FCN proposée par [88]

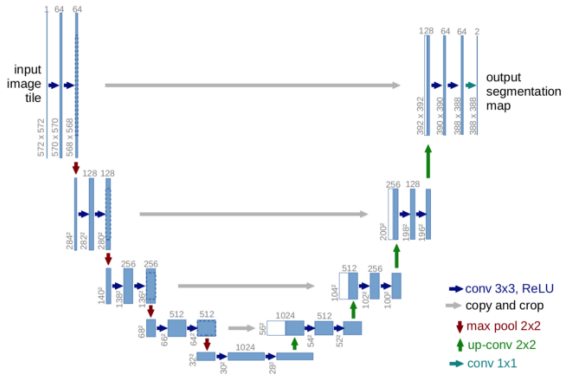


FIGURE 1.3 – Architecture U-Net proposée par [112]

médicales en 2D. Elle présente une structure encodeur-décodeur symétrique en forme de U (cf. Figure 1.3), avec une partie contractante qui encode et sous-échantillonne l'image, puis une phase d'expansion composée des mêmes blocs convolutifs et de couches de sur-échantillonnage pour retrouver progressivement les dimensions spatiales initiales. Des connexions résiduelles *(skip connections)* relient les blocs symétriques entre les deux phases pour concaténer des informations de haut et bas niveau, toujours dans cette idée de retrouver le contexte spatial, à l'image de ce que font Long *et al.* dans [88]. Cette architecture a largement fait ses preuves non seulement pour la segmentation d'images médicales, mais se voit également plébiscitée pour d'autres applications. Des variantes de U-Net consistent à en modifier le squelette. De plus, des extensions en 3D sont élaborées dont les modèles 3D U-Net [30] ou V-Net [93]. Ce dernier incorpore en plus des blocs résiduels inspirés de [61].

Par exemple, dans [114], les auteurs entraînent un FCN 3D multi-classes pour la segmentation de sept structures abdominales dans des CT-scans. Dans [102], 21 OAR de la tête et du cou sont segmentés en utilisant une architecture 3D U-Net. La segmentation du foie dans des images de scanner s'obtient grâce à un réseau 3D profondément supervisé dans [41], ou à une architecture U-Net hybride densément connectée dans [83].

Fonction de perte

Le choix d'une fonction de perte, aussi appelée fonction de coût ou objectif, se révèle déterminant pour l'apprentissage des poids et biais d'un réseau de neurones. Elle doit rendre compte des erreurs commises par le modèle et être différentiable par rapport aux paramètres θ pour les optimiser par rétropropagation de l'erreur du gradient à travers l'ensemble des couches. En matière de segmentation, le choix se porte communément vers l'entropie croisée ou le Dice généralisé, voire une combinaison des deux. Pour les définir, on considère une image x et on note y^l la version binaire *onehot* de la vérité terrain associée, avec $l \in \{1, ..., L\}$ et L le nombre de classes à segmenter. On note $s(\theta) = f(x; \theta)$ la fonction segmentation prédite par le CNN, paramétrée par les poids θ du réseau, dont la $l^{ième}$ composante donne la probabilité $s^l(\theta)_{i,j}$ de chaque pixel (i, j) d'appartenir à la classe l.

L'entropie croisée comme fonction de perte se formule comme

$$\mathcal{L}_{EC}(\theta) = -\sum_{l=1}^{L} \sum_{i,j} y_{i,j}^l \log s_{i,j}^l(\theta). \tag{1.8}$$

Cette fonction pénalise de manière équivalente chaque classe pour chaque pixel, ce qui s'avère problématique lorsqu'on observe un déséquilibre des classes à segmenter. En imagerie médicale, cet aspect se retrouve fréquemment puisque certaines structures anatomiques sont beaucoup plus petites que certains organes, et surtout, elles occupent moins d'espace que le fond de l'image. Des variantes de l'entropie croisée tendent à corriger ce problème. Par exemple, l'entropie croisée pondérée affecte un poids plus important aux classes sous-représentées. De son coté, la focal loss [85] cherche à diminuer la contribution des exemples les plus faciles à prédire via une faible pondération.
Milletari *et al.* [93] apportent une autre solution pour pallier ce problème de classes déséquilibrées. Ils introduisent une approximation du score de Dice dans le cadre de la segmentation binaire, notamment en le rendant différentiable et donc efficace pour l'optimisation par descente de gradient. Cette fonction pénalise la disparité de superposition entre la prédiction et la vérité terrain. Sudre *et al.* [127] proposent une généralisation adaptée à la segmentation multi-classes et la fonction objectif se formule alors comme

$$\mathcal{L}_{Dice}(\theta) = -2 \sum_{l=1}^{L} \frac{\sum_{i,j} s_{i,j}^l(\theta) y_{i,j}^l}{\sum_{i,j} s_{i,j}^l(\theta)^2 + \sum_{i,j} {y_{i,j}^l}^2}. \tag{1.9}$$

Ici aussi, des variantes existent pour moduler l'importance, par exemple, de certaines classes ou bien des faux-positifs et faux-négatifs [118].

1.2.3 Métriques d'évaluation

Pour quantifier la qualité des résultats de segmentation, il existe deux grands types de métriques. D'une part, les métriques de superposition dont le score de Dice fait partie. Il

mesure la similarité globale entre deux ensembles X et Y et se calcule comme

$$\text{Dice(X,Y)} = \frac{2|X \cap Y|}{|X| + |Y|}.$$

Cet indice donne un score compris entre 0, pour deux ensembles disjoints, et 1 lorsqu'au contraire les deux ensembles apparaissent complètement superposés. D'autre part, les métriques de distances surfaciques complètent ces premiers indices. La distance de Hausdorff permet de rendre compte de valeurs aberrantes et s'exprime en millimètres comme

$$\text{HD(X,Y)} = \max\{\sup_{y \in Y} \inf_{x \in X} d(x, y), \sup_{x \in X} \inf_{y \in Y} d(x, y)\},$$

avec d la distance entre deux points. La distance de Hausdorff moyenne peut également s'utiliser et revêt de fait un caractère plus global, et par conséquent moins pénalisant, puisqu'elle tient compte de tous les éléments des ensembles et non plus d'un seul. Elle se formule par

$$\text{MHD(X,Y)} = \frac{1}{2}\left(\frac{1}{|X|}\sum_{x \in X} \inf_{y \in Y} d(x, y) + \frac{1}{|Y|}\sum_{y \in Y} \inf_{x \in X} d(x, y)\right).$$

L'association et l'analyse de ces métriques, conjointement à une analyse qualitative, donne la possibilité d'évaluer la précision du modèle et aussi de mettre en lumière ses faiblesses. En effet, si les méthodes d'apprentissage profond exhibent des segmentations de bonnes qualités, certains artefacts subsistent malgré tout, tels qu'un groupe de pixels isolés et mal segmentés, des objets non constants par morceaux qui présentent des trous, etc.. Par conséquent les segmentations obtenues manquent de réalisme, ce qui devient très problématique dans un contexte médical. Le manque de plausibilité anatomique rend caduque la prise de décision sous-jacente à la tâche de segmentation.

1.2.4 Intégration de connaissances préalables

Afin d'obtenir des résultats plus précis et plus plausibles, de nombreux efforts ont été déployés ces dernières années, notamment pour incorporer des connaissances préalables dans les algorithmes de segmentation d'images. Ceci est particulièrement vrai pour la segmentation en imagerie médicale, où la cohérence anatomique se révèle essentielle. Par exemple, en raison du manque de contraste entre les organes et les tissus environnants, le problème de la segmentation des organes à risque nécessite de s'appuyer sur des connaissances externes, telles que des paires d'images de scanner et leur vérité terrain correspondante. L'utilisation de connaissances préalables et d'images étiquetées s'opère depuis longtemps dans la segmentation d'images médicales, pour guider le processus de segmentation en cas de bruit et d'occlusion, et pour traiter la variabilité des objets. En effet, comme précisé dans [141], les méthodes d'optimisation, dites de deuxième génération,

bien que très utiles, ne sont pas suffisantes en elles-mêmes pour produire des segmentations automatiques précises. D'où l'intérêt des méthodes qualifiées de troisième génération qui intègrent une petite quantité d'informations de haut niveau dans le processus de segmentation, telles que des informations a priori, des règles définies par des experts ou des modèles de l'objet souhaité.

Comme Nosrati et Hamarneh [104] le soulignent, les approches variationnelles ont recours depuis plusieurs années déjà à une variété d'informations *a priori*, telles que des connaissances préalables sur la forme [1], la garantie de préservation de la topologie [78], la prescription du nombre de composantes connexes/trous [120], ou la convexité (partielle) [123], et prouvent qu'il est possible d'obtenir des résultats plus précis [45]. En outre, comme spécifié dans la Section 1.2.1, les termes de régularisation constituent en eux-même un *a priori* sur les segmentations attendues (lisses, constantes par morceaux, etc.). En segmentation, les techniques de recalage d'atlas se révèlent très efficaces pour insérer des *a priori*. Cette méthode consiste à chercher la déformation optimale entre une image à segmenter et une image atlas, à laquelle est associée sa vérité terrain. Une fois la transformation faisant correspondre ces deux images obtenue, elle s'applique à l'atlas étiqueté pour obtenir la carte de segmentation finale de l'image initialement considérée. Si la transformation est un homéomorphisme, elle garantit en particulier la préservation de la topologie et de l'orientation de l'atlas, et la segmentation obtenue bénéficie de bonnes propriétés. Par conséquent, un modèle de recalage bien défini et un atlas correctement choisi s'avèrent primordiaux pour parvenir à un résultat précis. Dans [57], une méthode qui s'appuie sur un atlas, en plus d'autres techniques, est utilisée pour segmenter 17 OAR dans l'ensemble du corps. La segmentation d'organes thoraciques à risque est obtenue dans [119] en combinant un recalage déformable multi-atlas avec une recherche locale s'appuyant sur une méthode *level-set*.

Bien que l'intégration de connaissances *a priori* soit problématique avec les CNN, notamment en raison du critère de différentiabilité requis pour la rétropropagation, cela devient de plus en plus souhaitable. Ainsi, plusieurs méthodes ont vu le jour pour résoudre ce problème, souvent inspirées par des techniques issues d'un cadre variationnel.

Une première façon d'incorporer des informations préalables dans le processus consiste à utiliser un CNN avec des méthodes traditionnelles de modèles de formes et de modèles déformables dans un schéma séquentiel de type prédiction/correction. Ainsi, le CNN fournit en premier lieu une prédiction de segmentation qui est ensuite post-traitée et corrigée en respectant les contraintes prescrites. De façon inverse, le CNN peut opérer un prétraitement utile à la méthode de segmentation appliquée dans un second temps. Bohlender *et al.* [12] passent en revue de nombreux articles qui utilisent les réseaux de neurones convolutifs comme étape de pré- ou post-traitement. Par exemple, Rupprecht *et al.* [117]

utilisent un modèle de contours actifs [68] s'appuyant sur la détection de contours, pour fournir des segmentations grossières du ventricule gauche qui sont ensuite affinées par un CNN. Au contraire, Kamnitsas *et al.* [67] améliorent les résultats de la segmentation des lésions cérébrales obtenus avec un CNN en utilisant des champs aléatoires conditionnels (CRF, *Conditional Random Fields*) [72]. Les CRF sont capables de prendre en compte le contexte de l'image grâce à l'utilisation de connaissances globales (contours, intensités, etc.).

Une autre solution consiste à modifier directement l'architecture du réseau de neurones pour prendre en compte une information supplémentaire directement dans le processus d'apprentissage. Par exemple, dans [52], les auteurs ajoutent des branches connectées à la structure d'un CNN dans le but d'incorporer des contraintes spatiales via d'une part une carte de distances par rapport à des points de repère et d'autre part, l'intégration d'un atlas probabiliste. El Jurdi *et al.* [44] intègrent un *a priori* de localisation et de forme dans le processus d'apprentissage, en introduisant des filtres limitants au niveau des connexions résiduelles dans un modèle U-Net. Dans [131], une carte de distances fournit la localisation de chaque organe ainsi que la relation spatiale entre eux afin de guider la tâche de segmentation dans un cadre entièrement convolutif.

Une troisième alternative pour incorporer des connaissances préalables consiste à concevoir une fonction de perte qui peut les intégrer. À titre d'exemple dans [28], Chen *et al.* proposent une fonction de perte supervisée inspirée de l'énergie globale de Chan-Vese, pour intégrer les informations d'aire et de taille, dans le but de segmenter les différentes parties du cœur. Kim et Ye [70] définissent une nouvelle fonction de perte fondée sur la fonctionnelle de Mumford-Shah pour garantir l'homogénéité de l'intensité lumineuse des pixels et régulariser la longueur du contour. Ils l'appliquent à la segmentation de tumeurs du cerveau et du foie. De plus, il existe d'autres types de fonctions de perte qui permettent l'incorporation d'informations spatiales ou de connaissances *a priori*. El Jurdi et al [45] présentent nombre d'entre elles et les classifient selon la nature de l'*a priori* (contraintes de forme, de taille, de topologie et d'inter-régions des objets). Par exemple, Ganaye *et al.* [53] introduisent une fonction coût qui pénalise les relations d'adjacence anatomiquement incorrectes dans un contexte d'IRM du cerveau et d'images de scanner du corps entier. Dans les travaux de Clough *et al.* [32], une connaissance préalable de la topologie est incorporée dans la fonction objectif et expérimentée, entre autres, en imagerie cardiaque. Kervadec *et al.* [69] établissent une fonction objectif pour appliquer des contraintes d'inégalité de taille dans le contexte des images cardiaques. Ces méthodes basées sur des pénalisations sont très simples à utiliser mais ne garantissent pas la satisfaction de la contrainte. Ainsi, Dolz *et al.* [39] proposent de contraindre la fonction de perte avec un *a priori* de compacité de forme et résolvent le problème d'optimisation via un algorithme de la méthode des multiplicateurs à direction alternée (ADMM) dans le but de segmenter l'aorte, l'œsophage ou le ventricule droit. Peng *et al.* [107] utilisent également cet algo-

rithme pour appliquer des contraintes de taille et une régularisation de la longueur des contours dans l'apprentissage d'un CNN.
Dans [129], les auteurs suggèrent plutôt d'entrelacer les cadres variationnel et d'apprentissage profond. Plus précisément, les algorithmes ADMM sont utilisés comme substitut à la descente de gradient dans l'étape d'entraînement. Finalement, Liu *et al.* [87] arguent que les connaissances préalables n'interviennent pas au moment de l'inférence. Pour remédier à cela, ils proposent d'interpréter la fonction d'activation de softmax comme une variable duale d'un problème variationnel. De cette façon, ils peuvent réaliser une régularisation spatiale grâce à la variation totale avec une contrainte de volume ou un *a priori* de forme (ouvert étoilé) dans des CNN, en considérant chaque étape de leur algorithme itératif comme des couches du réseau. Ce dispositif est testé en particulier pour la segmentation de lésions cutanées.

1.2.5 Discussion et conclusion

Dans cette première partie, nous avons présenté les méthodes variationnelles et les méthodes d'apprentissage profond supervisé, dédiées à la segmentation d'images médicales. Les premières reposent sur une longue histoire permettant une évolution vers des méthodes de plus en plus robustes et efficaces. De plus, elles bénéficient du cadre continu et s'appuient sur des concepts mathématiques clairs et solides. Les secondes, et en particulier les réseaux de neurones convolutifs, sont en revanche beaucoup plus récentes. Malgré un manque de compréhension formel et d'explicabilité des réseaux profonds, souvent qualifiés de boîtes noires, leur grande efficacité à produire des segmentations précises et leur capacité de généralisation les propulsent au rang d'état de l'art. Néanmoins, quelle que soit l'approche envisagée, les résultats de segmentation montrent souvent des erreurs aberrantes et rédhibitoires pour l'analyse médicale telles que le diagnostic d'une pathologie ou la mise en place d'un traitement par radiothérapie. En résultent des méthodes qui incorporent des connaissances préalables pour guider le processus de segmentation et qui ont fait leurs preuves dans le cadre variationnel, inspirant la communauté Deep-Learning. Par ailleurs, les challenges s'avèrent indispensables pour faire progresser à la fois les applications ainsi que les technologies, autrement dit les architectures des CNN mais aussi chacun des éléments qui les composent. C'est pourquoi nous en avons mis un en place pour la segmentation de l'ensemble de données SegTHOR que nous présentons dans la section suivante. Cette application constitue une source de motivation pour nos travaux.

1.3 sU-Net : un CNN de base pour le dataset SegTHOR

1.3.1 Motivation et contexte

Une des grandes difficultés de la segmentation par apprentissage profond entièrement supervisé réside dans la nécessité de disposer d'une quantité massive de données annotées. Cette annotation requiert une interaction humaine manuelle, et dans le cadre de données médicales, celle d'un expert dont le temps est précieux et souvent limité. Comme précisé dans [62], la collecte de cet énorme ensemble de cas labellisés dans les images médicales est particulièrement délicate. D'une part, [111] explique que le caractère confidentiel de ce type d'informations rend le partage de ces images plus difficile que pour d'autres images. D'autre part, cette tâche fastidieuse et chronophage oblige le clinicien à s'appuyer sur son expérience et des directives médicales. C'est d'ailleurs pour cette raison qu'une approche automatique semble essentielle pour améliorer et simplifier la segmentation de structures anatomiques.

À l'ère de la science ouverte, les datasets médicaux publics, ainsi qu'un protocole expérimental commun, simplifient le processus de conception et de validation des algorithmes de science des données ; ils contribuent également à faciliter la reproductibilité et la comparaison équitable entre les méthodes. Dans cette idée, diverses bases de données et challenges pour la segmentation d'images médicales sont disponibles dont, entre autres :

1. LiTS [11] (Liver Tumor Segmentation Challenge [1]) qui soumet le défi de segmenter les lésions hépatiques dans les images de scanner des abdominaux avec contraste. Le dataset dispose de 200 CT-scans de patients. Ce challenge a été organisé conjointement pour ISBI 2017 et MICCAI 2017 ;
2. ACDC [9] (Automated Cardiac Diagnosis Challenge [2]), où le dataset possède 150 images IRM cardiaques et dont l'objectif est de segmenter le ventricule gauche, le myocarde et le ventricule droit chez plusieurs types de patients répartis en cinq groupes. Un premier groupe de patients est en bonne santé tandis que les autres présentent quatre pathologies différentes. Ce challenge a également été élaboré pour MICCAI 2017 ;
3. MSD [4] (Medical Segmentation Decathlon Challenge) [3] qui propose de réaliser les segmentations de diverses structures anatomiques telles que des organes, des tumeurs ou des vaisseaux sanguins. En tout, 10 jeux de données différents sont disponibles avec des modalités et un nombre d'observations variés.

1. `https://competitions.codalab.org/competitions/17094`
2. `https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html`
3. `https://decathlon-10.grand-challenge.org/`

Chaque challenge soulève ses propres défis. Cependant, il n'en existe que très peu pour la planification de la radiothérapie. Ce traitement demeure standard pour soigner les cancers du poumon et de l'œsophage. Il consiste à irradier la tumeur avec des faisceaux ionisants pour empêcher la prolifération des cellules cancéreuses. Le but est alors de détruire la tumeur ciblée tout en épargnant les organes qui la jouxtent, appelés Organes À Risque (OAR), des radiations. Ainsi, délimiter la tumeur et les OAR dans les images de scanner représente la première étape dans la planification du traitement. Les approches automatiques apparaissent très désirables pour améliorer et simplifier cette étape en amont de la thérapie, et par conséquent réduire les effets indésirables de la radiothérapie.

Dans un esprit de partage, pour favoriser la recherche dans ce domaine et rendre automatique et plus large la segmentation des OAR, nous avons récemment établi un jeu de données avec des images acquises au Centre Henri Becquerel (CHB), un centre régional anti-cancer à Rouen, France. Ce jeu de données, nommé SegTHOR pour *Segmentation of THoracic Organs at Risk*, contient 60 CT-scans de patients atteints d'un cancer du poumon ou d'un lymphome de Hodgkin. Avec ce dataset, nous nous intéressons aux organes thoraciques dont l'aorte, le cœur, la trachée et l'œsophage.

À notre connaissance, il n'existe pas beaucoup de jeux de données destinés à la segmentation des organes à risque. Le challenge [4] proposé par l'AAPM (American Association of Physicists in Medicine) a un but similaire : il aspire à segmenter l'œsophage, le cœur, la moelle épinière, les poumons droit et gauche dans des images de scanner. L'ensemble d'entraînement est composé de 30 patients tandis que l'ensemble de test inclut les scanners de 12 patients. Les organes diffèrent de ceux de SegTHOR puisque leur dataset n'inclut ni la trachée, ni l'aorte. Plus récemment, le challenge StructSeg 2019 [5] propose deux tâches de segmentation d'OAR. L'objectif de la première est d'en segmenter 22 dans les images de scanner de la tête et du cou de patients atteints de cancer du nasopharynx. La seconde ambitionne de segmenter 6 OAR dans les scanners thoraciques de patients atteints d'un cancer du poumon. Les organes sont les mêmes que ceux du challenge de l'AAPM, avec la trachée en supplément. Dans les deux cas, 50 CT-scans composent l'ensemble d'entraînement et 10 autres constituent celui de test.
Nous présentons maintenant plus en détail le jeu de données SegTHOR ainsi que le challenge dont il a fait l'objet. Puis, nous donnons quelques résultats de base en utilisant les réseaux de segmentation les plus modernes.

1.3.2 Le dataset et le challenge SegTHOR

La base de données comprend 60 scanners thoraciques, acquis avec ou sans contraste intraveineux, de 60 patients diagnostiqués d'un cancer du poumon ou d'un lymphome de

4. `http://aapmchallenges.cloudapp.net/competitions/3`
5. `https://structseg2019.grand-challenge.org/`

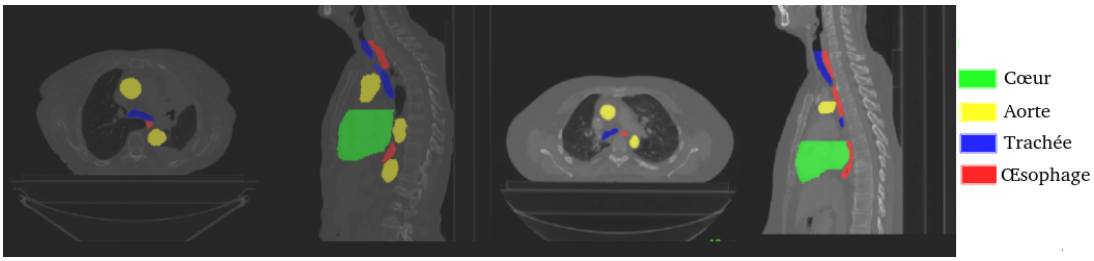


FIGURE 1.4 – Exemple de scanners thoraciques de 2 patients, dont les vues axiale (gauche) et sagittale (droite), avec la superposition des 4 OAR segmentés manuellement.

Hodgkin. Ces patients ont reçu un traitement par radiothérapie curative-intensive, entre Février 2016 et Juin 2017, au Centre Henri Becquerel de Rouen (CHB, centre anti-cancer régional). Roger Trullo [130] est à l'origine de ce projet et a pris en charge la collecte de ces données avec un radiothérapeute du CHB.

Concernant les caractéristiques techniques, tous les CT-scans présentent une taille de $512\times512\times(135 \sim 284)$ voxels. En effet, le nombre de coupes change selon le patient. La résolution dans le plan varie entre 0.90 mm et 1.37 mm par pixel et la résolution en z fluctue entre 2 mm et 3.7 mm par pixel. Finalement, la résolution la plus commune est $0.98\times0.98\times2.5$ mm^3.
Nous associons à chaque CT-scan une segmentation manuelle, réalisée par un radiothérapeute expérimenté du CHB, en utilisant la plateforme SomaVision (Varian Medical Systems, Inc, Palo Alto, USA). Cette segmentation manuelle nécessite approximativement 30 minutes pour chaque patient. D'abord, les contours du corps et les poumons sont détectés grâce aux outils automatiques disponibles sur la plateforme. L'œsophage est délimité manuellement à partir de la quatrième vertèbre cervicale jusqu'à la jonction œsophago-gastrique. Le cœur est segmenté en suivant les recommandations du Groupe d'Oncologie de la Radiothérapie. La trachée est profilée depuis la limite inférieure du larynx jusqu'à 2 cm en-dessous de la carène, à l'exclusion des bronches lobaires. Enfin, l'aorte est délinéée depuis son origine au-dessus du cœur jusqu'en dessous des piliers du diaphragme.

La segmentation de ces quatre OAR soulève les défis suivants. En premier lieu, les tissus qui englobent le cœur, et particulièrement l'œsophage, montrent des niveaux de gris similaires à ceux des autres organes ; le manque de contraste force le radiothérapeute à s'appuyer sur ses connaissances anatomiques aboutissant à une segmentation qui ne repose pas uniquement sur l'image. L'œsophage est l'organe le plus difficile à contourer : non seulement il est pratiquement invisible, mais en plus, sa forme et sa position varient grandement d'un patient à un autre, mais aussi au sein du même patient comme en témoigne la Figure 1.5. Au contraire la trachée est facilement identifiable, car elle est

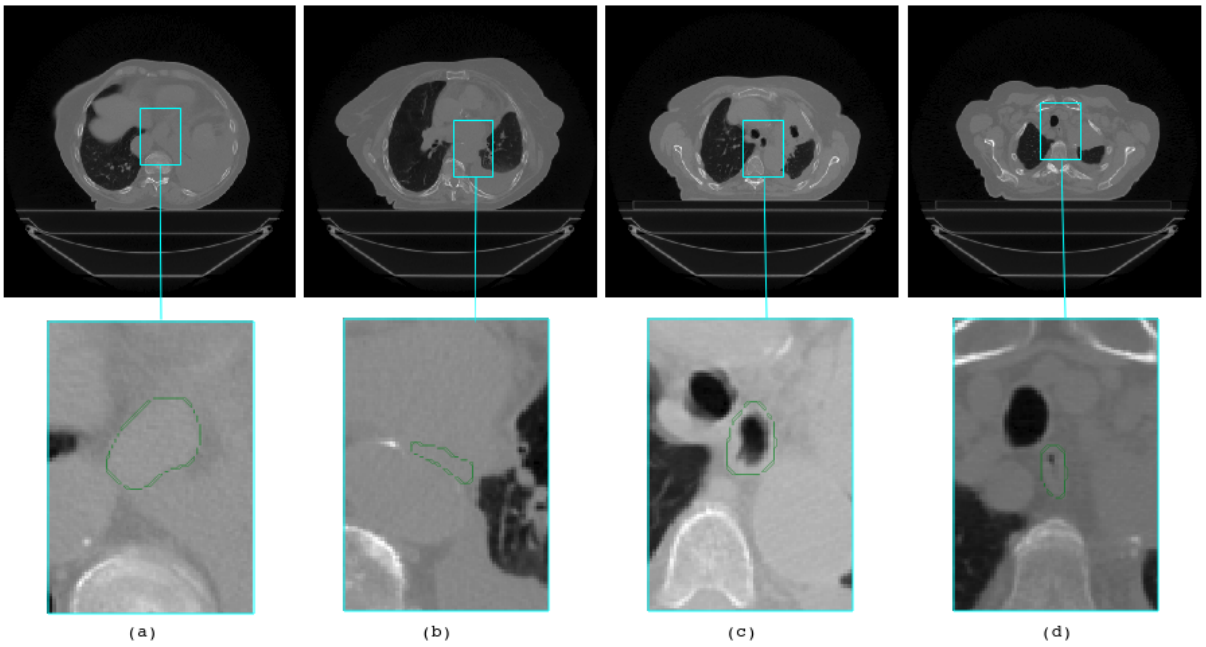


FIGURE 1.5 – Exemple de 4 coupes axiales 2D d'un patient avec les contours de la vérité terrain de l'œsophage superposés en vert.

remplie d'air et de fait apparaît en noir sur les scans. De plus, un autre challenge est la relation tridimensionnelle de ces OAR : ils sont étroitement imbriqués comme montré sur la Figure 1.4. Pour finir, les quatre OAR affichent des formes et des tailles variables : l'œsophage et la trachée, les plus petits organes, ont une structure tubulaire ; l'aorte revêt une forme de canne et le cœur, l'organe le plus grand, a une forme de blob.

Il en résulte que les quatre classes et le fond présentent un fort déséquilibre. En effet, le fond constitue environ 99% des voxels en moyenne. Le pourcentage restant se divise en 70.7% pour le cœur, 23% pour l'aorte, 3.7% pour l'œsophage et 2.6% la trachée. Par ailleurs, plusieurs disparités se retrouvent en deux dimensions (2D), puisque sur les 11084 coupes qui constituent le dataset, seulement 6891 présentent au moins un des quatre organes, soit 62.2%. Plus spécifiquement, et comme la Table 1.1 le précise, l'aorte et l'œsophage sont présents sur davantage de coupes en comparaison de la trachée et du cœur. En revanche, l'aire moyenne occupée par le cœur sur chaque coupe où il apparaît est environ 4.5, 17.6 et 33.6 fois supérieure à, respectivement, celle de l'aorte, l'œsophage et la trachée. Enfin, la Table 1.1 indique le nombre de composantes connexes (CC) de chaque organe qui varie encore une fois selon l'organe et la coupe étudiés.

TABLE 1.1 – Quelques caractéristiques des OAR d'un point de vue 2D.

	Œsophage	Cœur	Trachée	Aorte
Nombre de coupes	5782	2325	2940	5578
Aire moyenne par coupe (en px)	220	3877	171	844
Nombre de composantes connexes par coupe	1	1	1-2	1-2

Pour cadrer le challenge [6], nous avons divisé les données en un ensemble d'entraînement de 40 patients et un ensemble de test constitué des 20 autres patients, ce qui représente, respectivement, 7390 et 3694 coupes 2D. L'accès au challenge, et donc au dataset, s'effectue en ligne sur la plateforme Codalab [7], où une évaluation automatique en ligne est disponible. Deux métriques sont utilisées pour quantifier les résultats de segmentation sur l'ensemble de test. D'abord le score de Dice qui mesure le taux de superposition entre la segmentation automatique et manuelle. En complément de cette métrique, la distance de Hausdorff moyenne (MHD) est calculée en millimètres (mm) comme la moyenne des distances moyennes des contours manuels à ceux automatiques les plus proches, et des distances moyennes des contours automatiques à ceux manuels les plus proches. Les deux scores sont obtenus pour les 4 organes. Les participants à la compétition obtiennent alors un rang pour les huit scores, et la moyenne de ces huit rangs sert à définir le classement final. Le challenge s'est déroulé de Janvier à Avril 2019 et a été présenté lors du Symposium International IEEE sur l'Imagerie Biomédicale (ISBI) en Avril 2019 à Venise, en Italie. Bien que la compétition soit terminée, le challenge reste accessible pour permettre à la communauté scientifique de continuer à évaluer et comparer leurs méthodes de segmentation sur le dataset SegTHOR.

1.3.3 Un modèle de segmentation simplifié basé sur U-Net

L'architecture U-Net étant le modèle à état de l'art pour la segmentation d'images médicales, notre première intention est d'évaluer cette architecture avec chaque image 2D du dataset de test de SegTHOR. Étant donné que les contours des OAR présentent une grande variabilité inter- et intra-patient, nous estimons qu'ils sont sujets à un sur-apprentissage. Notre stratégie consiste à adapter U-Net à notre problème par quelques étapes très simples. La première étape pour lutter contre le sur-apprentissage réside dans l'ajout d'une régularisation appelée *dropout* [125] au réseau, un élément commun dans les CNN modernes. Le *dropout* consiste à ignorer aléatoirement chaque neurone du réseau avec une probabilité p, et par conséquent leurs connexions, à chaque étape de l'entraînement. Cela empêche les neurones de trop s'adapter les uns aux autres et donc de conserver leur capacité de généralisation. La réduction du nombre de couches et de cartes caractéristiques, afin de diminuer le nombre de paramètres entraînables dans le réseau, représente

6. L'organisation du challenge a eu lieu en amont de ce travail de thèse.
7. `https://competitions.codalab.org/competitions/21145`

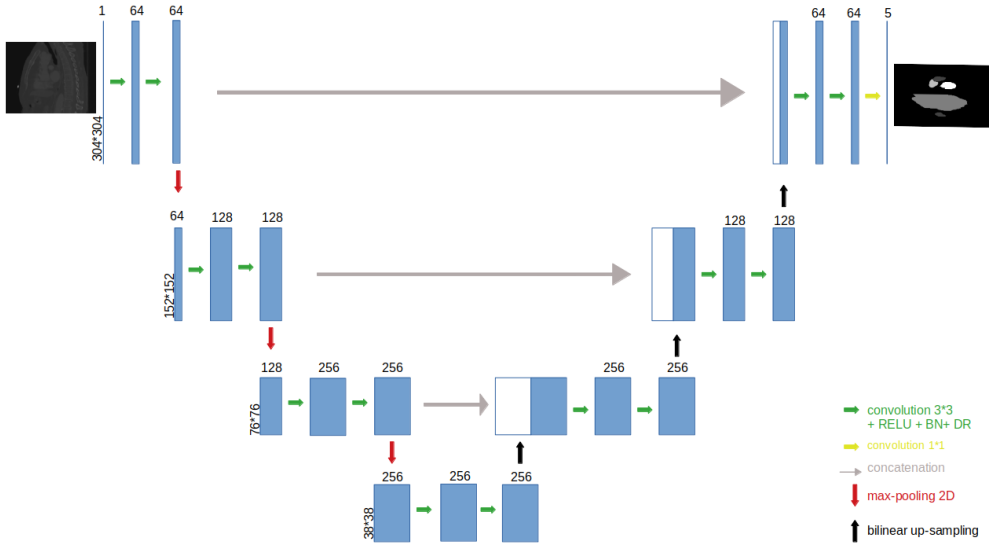


FIGURE 1.6 – Architecture de notre proposition simplifiée de U-Net, notée sU-Net. Chaque rectangle bleu indique une couche, pour laquelle le numéro au-dessus est le nombre de cartes caractéristiques. La résolution de l'image correspondante est spécifiée verticalement.

une seconde façon de tempérer le sur-apprentissage. Il en résulte une architecture simplifiée avec une couche cachée en moins et seulement jusqu'à 256 cartes caractéristiques calculées par couche. Enfin, nous choisissons de remplacer la convolution transposée (également appelée déconvolution) par une interpolation bilinéaire pour l'opération de sur-échantillonnage, dans la phase d'expansion. La première requiert l'apprentissage des poids des filtres, tandis que la seconde utilise les pixels voisins pour calculer la valeur du nouveau pixel par le biais d'interpolations linéaires, ce qui réduit encore le nombre de paramètres à apprendre.

Comme illustré sur la Figure 1.6, notre réseau simplifié, nommé sU-Net, exhibe un schéma encodeur-décodeur composé de sept blocs de convolution, et certains sont reliés par des connexions résiduelles. Chaque bloc de convolution consiste en deux opérations de convolution avec un noyau de taille 3×3. Une fonction d'activation ReLU (Rectified Linear Units), puis une normalisation par lots, sont appliquées aux sorties de chaque convolution. La régularisation *dropout* est intégrée à chacun des blocs. Dans la partie encodeur, les deux opérations de convolution sont suivies par une opération de *max-pooling* qui réduit de moitié la dimension spatiale de l'entrée, tandis que dans la partie décodeur, les deux opérations de convolution sont précédées par un sur-échantillonnage bilinéaire pour doubler la résolution et finalement retrouver la dimension initiale de l'image. Les trois connexions résiduelles servent à concaténer les caractéristiques des premières couches avec celles des couches plus profondes pour compenser la perte de résolution. Dans le

dernier bloc, une opération de convolution finale avec un noyau 1×1 sert à obtenir les cartes caractéristiques de chaque classe de segmentation. Finalement, cette architecture détient seulement 4.8 millions de paramètres entraînables en comparaison des 7.2 millions pour la même architecture avec les opérations de convolution transposée tandis que le U-Net original, basé sur un squelette VGG, possède environ 65 millions de paramètres entraînables.

1.3.4 Expériences et résultats

Pré-traitement

Toutes les images sont normalisées en soustrayant leur moyenne et en divisant par l'écart-type. Nous triplons artificiellement la taille de la base de données d'entraînement en utilisant des techniques d'augmentation de données. De la même manière que décrit dans [93], chaque image est modifiée par une transformation affine aléatoire d'une part, et est déformée en utilisant un champ de déformation dense obtenu par le biais d'une grille de points de contrôle $2 \times 2 \times 2$ puis une interpolation B-Spline d'autre part. Pour des raisons de calcul, nous rognons les images à partir de leur centre, afin d'en réduire la taille à 304×304 pixels. De plus, seules les coupes détenant au moins un organe sur quatre passent à travers le réseau pendant l'apprentissage.

Implémentation

Pour contrer le problème de déséquilibre des classes évoqué précédemment, nous utilisons comme fonction de perte le Dice multi-classes, qui est la généralisation de la version binaire [93, 127]. Les paramètres du réseau sont aléatoirement initialisés grâce à la technique d'initialisation de Glorot [55] et mis à jour par une descente stochastique de gradient (SGD, *Stochastic Gradient Descent*). Les hyperparamètres du réseau sont optimisés en suivant une technique de quadrillage, en particulier nous fixons le taux d'apprentissage initial à 10^{-3} et la taille du lot à 5. Ce taux est divisé par un facteur dix lorsque l'entraînement ne progresse plus durant 10 époques successives (une époque correspond à la visualisation par le réseau de tous les échantillons de la base de données). L'ensemble du code est développé en Python avec la librairie Pytorch. Nous présentons maintenant les résultats obtenus. Comme pour le challenge, nous quantifions nos résultats de segmentation avec le score de Dice et la distance de Hausdorff moyenne (MHD) en millimètres. Ces métriques sont données pour les quatre OAR.

Résultats

Dans une première expérience, nous comparons la performance de U-Net avec la version simplifiée sU-Net. Nous mesurons également la différence sans et avec la régularisation *dropout* (DR), avec une probabilité d'oubli p de 0.2. Ensuite, nous évaluons les deux configurations de sur-échantillonnage dans la phase décodeur : (i) avec une opération de

TABLE 1.2 – Résultats de segmentation (moyenne ± écart-type) obtenus avec U-Net et notre architecture sU-Net, avec et sans DRopout (DR). Métriques utilisées : score de Dice et distance de Hausdorff moyenne (MHD). Cases en jaune : valeurs sU-Net qui diffèrent significativement des valeurs de U-Net, avec DR (colonnes (4) vs (2)). En bleu et gras : valeurs avec DR qui diffèrent significativement de celles sans DR (colonnes (4) vs (3)).

		U-Net		U-Net simplifié (sU-Net)	
OAR	Métriques	sans DR (1)	avec DR (2)	sans DR **(3)**	avec DR **(4)**
Œsophage	Dice %	76 ± 10	79 ± 8	75 ± 11	**82 ± 5**
	MHD (mm)	1.74 ± 2.77	0.94 ± 0.63	1.69 ± 2.02	**0.70 ± 0.39**
Trachée	Dice %	85 ± 5	85 ± 4	**86 ± 4**	85 ± 4
	MHD (mm)	1.32 ± 1.20	1.30 ± 1.12	1.06 ± 0.83	1.21 ± 1.13
Aorte	Dice %	92 ± 5	91 ± 4	91 ± 2	91 ± 3
	MHD (mm)	0.50 ± 0.64	0.77 ± 0.93	0.57 ± 0.65	0.58 ± 0.67
Cœur	Dice %	93 ± 3	93 ± 3	92 ± 3	**93 ± 3**
	MHD (mm)	0.23 ± 0.21	0.25 ± 0.28	0.31 ± 0.22	**0.27 ± 0.20**

convolution transposée 2D (notée conv2Dtransposee dans la table résultat), et (ii) avec une opération de sur-échantillonnage bilinéaire 2D (notée interp2Dbilineaire), qui sont utilisées pour retrouver la dimension initiale de l'image. Chaque fois que nécessaire, nous calculons la signification statistique des résultats, en réalisant un test des rangs signés de Wilcoxon sur les valeurs de Dice et de MHD, entre les deux méthodes d'intérêt, et en utilisant un intervalle de confiance de 95%.

Comparaison de sU-Net vs U-Net et influence du *dropout* - Les résultats sont indiqués dans la Table 1.2. En comparant sU-Net à U-Net sans *dropout* (colonnes (1) et (3)), nous remarquons que les résultats sont clairement similaires. Maintenant, en incluant le *dropout* dans les deux réseaux, sU-Net montre des performances légèrement améliorées comparé à U-Net (colonnes (2) vs (4)) pour tous les organes exceptée la trachée. Cette observation est confirmée par la p-value du test de Wilcoxon, qui est inférieure au seuil de 0.05, pour l'œsophage, le cœur et l'aorte. Les résultats qualitatifs de la Figure 1.7 illustrent la différence entre les deux architectures pour l'œsophage. La contribution du *dropout* à l'architecture sU-Net peut être mesurée en analysant les colonnes (3) et (4), et nous remarquons que pour deux OAR, le *dropout* apporte une amélioration substantielle, notamment pour l'œsophage.

Influence de la méthode de sur-échantillonnage dans le décodeur - La méthode de convolution transposée est comparée à celle de l'interpolation bilinéaire dans la Table 1.3. Globalement, les valeurs de Dice et MHD n'apparaissent pas significativement différentes ($p \geq 0.05$), hormis pour l'aorte pour laquelle la p-value vaut $1.2e^{-4}$ en faveur du sur-

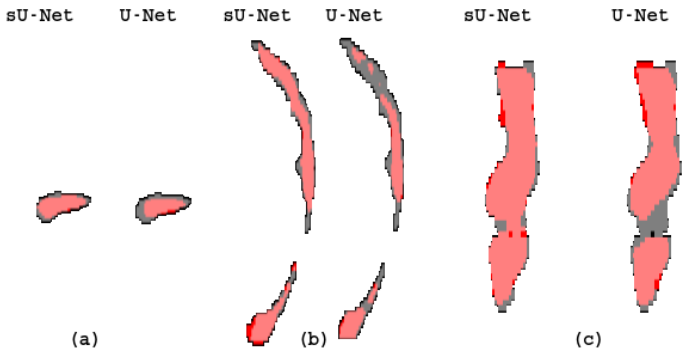


FIGURE 1.7 – Comparaison des résultats de segmentation de l'œsophage avec la version simplifiée sU-Net (à gauche) et l'original U-Net (à droite), pour chaque vue (a) axiale, (b) sagittale et (c) coronale. Les aires prédites apparaissent en rouge et la vérité terrain en gris.

échantillonnage bilinéaire. Par conséquent cette méthode s'avère plus que suffisante pour cette application. De plus, choisir l'interpolation bilinéaire peut aider à réduire le temps de calcul.

Limitations du jeu de données

La segmentation manuelle de l'ensemble de données SegTHOR est adaptée aux besoins de la radiothérapie et n'a pas été réalisée pour une évaluation systématique de la segmentation. Ainsi, en raison des recommandations formulées pour la segmentation manuelle, certaines coupes localisées en bas ou en haut des images 3D de scanner du patient n'ont pas été segmentées. Bien que cette absence de labellisation manuelle n'entrave pas l'évaluation de la segmentation du cœur, cela peut être un problème pour les organes tubulaires qui sont perpendiculaires au plan axial, comme l'œsophage, la trachée, et dans une moindre mesure, l'aorte. Pour une majorité des 20 patients testés, la segmentation automatique de l'œsophage, de la trachée et de l'aorte dépasse les limites supérieures et inférieures de la segmentation manuelle comme l'illustre la Figure 1.8 et produit un étiquetage considéré comme une mauvaise segmentation, puisque la vérité terrain (GT) correspondante n'existe pas. Nous réalisons de nouvelles expériences pour évaluer le gain lors de l'évaluation sur la plage restreinte de coupes où la GT est présente. Comme attendu et à partir de la Table 1.4, nous constatons pour l'œsophage, la trachée et l'aorte, une amélioration des scores de Dice, surtout pour la trachée, mais plus significative encore pour les distances moyennes de Hausdorff.

1.3.5 Discussion et conclusion

Dans cette section, nous avons présenté SegTHOR, un jeu de données pour la segmentation d'organes à risque dans les images de scanner, disponible sur la plateforme

TABLE 1.3 – Résultats de segmentation (moyenne ± écart-type) produits par U-Net et notre architecture sU-Net, dans les deux cas avec DR. Métriques utilisées : score de Dice et distance de Hausdorff moyenne (MHD). En bleu et gras : valeurs de sU-Net avec sur-échantillonnage bilinéaire qui différent significativement de celui avec la convolution transposée (colonnes (2) vs (3)).

		U-Net	U-Net simplifié (sU-Net)	
OAR	Métriques	conv2Dtransposee (1)	conv2Dtransposee **(2)**	interp2Dbilineaire **(3)**
Œsophage	Dice %	79 ± 8	82 ± 5	81 ± 6
	MHD (mm)	0.94 ± 0.63	0.70 ± 0.39	0.68 ± 0.35
Trachée	Dice %	85 ± 4	85 ± 4	86 ± 4
	MHD (mm)	1.30 ± 1.12	1.21 ± 1.13	1.08 ± 0.85
Aorte	Dice %	91 ± 4	91 ± 3	**92 ± 2**
	MHD (mm)	0.77 ± 0.93	0.58 ± 0.67	**0.52 ± 0.66**
Cœur	Dice %	93 ± 3	93 ± 3	93 ± 3
	MHD (mm)	0.25 ± 0.28	0.27 ± 0.20	0.26 ± 0.22

Codalab. Le challenge SegTHOR aspire à encourager la recherche sur cette application clinique, mais aussi à promouvoir le domaine de la segmentation multi-classes pour les images anatomiques (volumétriques). Nous avons proposé plusieurs variantes d'une architecture basée sur U-Net qui peut être utilisée comme approche initiale lors du traitement d'un nouveau problème de segmentation d'images médicales. Compte-tenu de la quantité limitée de données disponibles, une architecture trop profonde et comprenant un grand nombre de cartes caractéristiques ne semble pas le plus adapté à notre problème de segmentation sémantique, en particulier pour la segmentation de l'œsophage. Nous avons introduit un CNN simplifié plus approprié au problème posé. Les résultats montrent que l'ajout du *dropout* a une influence majeure sur la précision, et permet, pour la plupart des organes, d'améliorer la métrique de Dice ainsi que la distance de Hausdorff moyenne. Dans la phase d'expansion, la convolution transposée n'a pas donné de meilleurs résultats que l'opération de sur-échantillonnage bilinéaire ; dans ce cas, l'interpolation bilinéaire devrait être privilégiée pour réduire le nombre de paramètres et par conséquent le temps de calcul.

Une des limites de notre approche réside dans le fait que nous n'utilisons qu'une seule segmentation de référence. Il est connu que la variabilité de la segmentation manuelle, qu'elle soit intra- ou inter-expert, n'est pas négligeable. Plus important encore, la segmentation de l'OAR a une influence considérable sur les mesures dosimétriques [138]. Il serait donc intéressant d'évaluer quantitativement l'influence de la segmentation des OAR sur la dose dosimétrique. Dans une étude portant sur un patient atteint d'un cancer de l'oropharynx [101], les auteurs ont constaté des différences de dose substantielles résul-

FIGURE 1.8 – De gauche à droite : CT-scan d'un patient, vérité terrain labellisée et segmentation automatique produite par sU-Net en configuration sur-échantillonnage bilinéaire. Blanc : aorte, gris clair : trachée, gris foncé : cœur, gris très foncé : œsophage. Les lignes horizontales indiquent les niveaux de coupe inférieurs et supérieurs où la vérité terrain s'arrête.

tant strictement de la variation du contour, en fonction de la taille, de la forme et de l'emplacement de l'OAR. Cela souligne la nécessité de contourer avec précision l'OAR, en plus de la tumeur cible, lors de la planification d'une radiothérapie. Une étude dosimétrique permettrait également d'éviter le problème d'étiquetage présent dans l'ensemble de données.

L'apprentissage faiblement supervisé pour la segmentation d'images [128] ou la gestion des annotations manquantes [108] représentent un autre cas d'utilisation de ce jeu de données. L'apprentissage faiblement supervisé permet d'obtenir une segmentation complète à partir de données partiellement annotées, réduisant ainsi le coût de l'annotation complète. Ce paradigme, identifié comme un sujet majeur pour les années à venir [128], soulève de nouveaux défis : comment exploiter les étiquettes faibles ? Comment utiliser et modéliser les connaissances externes pour aider le processus ?

1.4 Conclusion

Pour conclure, dans ce chapitre, nous avons mis en avant les approches variationnelles et les approches Deep-Learning dans le contexte particulier de la segmentation d'images médicales, qui sont au cœur de nos travaux. Nous avons souligné les avantages et les limites de chacune de ces approches. Le manque de plausibilité anatomique, du fait des artefacts qu'exhibent les résultats de segmentation, est un fléau commun à ces méthodes. Cela empêche d'exploiter directement les résultats dans un cadre clinique. Par conséquent, l'intégration de connaissances préalables dans les différents modèles devient fortement souhaitable pour annihiler ces effets indésirables. Des solutions pour y parvenir ont été évoquées dans ce chapitre.
Par ailleurs, nous avons également introduit le jeu de données SegTHOR, dédié à la seg-

TABLE 1.4 – Résultats de segmentation (moyenne ± écart-type) obtenus pour le dataset original et le dataset restreint aux coupes labellisées, avec sU-Net + DR + sur-échantillonnage bilinéaire. Métriques utilisées : score de Dice et distance de Hausdorff moyenne (MHD). En gras, les meilleurs résultats.

OAR	Métriques	sU-Net dataset original	sU-Net dataset restreint
Œsophage	Dice %	81 ± 6	**83 ± 6**
	MHD (mm)	0.68 ± 0.35	**0.32 ± 0.20**
Trachée	Dice %	86 ± 4	**92 ± 2**
	MHD (mm)	1.08 ± 0.85	**0.15 ± 0.09**
Aorte	Dice %	92 ± 2	**93 ± 2**
	MHD (mm)	0.52 ± 0.66	**0.19 ± 0.31**
Cœur	Dice %	93 ± 3	93 ± 3
	MHD (mm)	0.26 ± 0.22	**0.16 ± 0.15**

mentation d'organes à risque dans des images de scanner en prévision d'un traitement par radiothérapie. Ce dataset a fait l'objet d'un challenge afin de contribuer à l'amélioration des méthodes pour cette application. Nous avons mis en place un réseau de neurones convolutifs appelé sU-Net, une version simplifiée du populaire U-Net, dans le but d'obtenir une première salve de résultats qui sert de repère pour la suite de nos expériences.
Cet état de l'art a permis de nous convaincre qu'un modèle créant une synergie entre les approches variationnelles et d'apprentissage profond permet d'intégrer efficacement des informations *a priori*, et donc d'améliorer les segmentations de SegTHOR, mais pas uniquement. Nous passons maintenant en revue les outils mathématiques nécessaires à l'élaboration des résultats théoriques de nos modèles.

CHAPITRE 2

Outils et rappels mathématiques

Dans ce chapitre, nous introduisons les outils mathématiques principaux nécessaires à l'élaboration et à l'analyse de nos modèles. Nous rappelons d'abord les définitions et propriétés des espaces fonctionnels qui apparaîtront naturellement dans nos modèles. Ensuite, nous résumons les étapes relatives à la méthode directe du calcul des variations qui sera invoquée de façon sous-jacente à plusieurs reprises. Enfin, nous abordons quelques notions utiles en élasticité non linéaire, théorie sur laquelle est bâtie le Chapitre 4 en particulier, puis certains éléments d'optimisation convexe qui irriguent les Chapitres 3 et 4.

2.1 Les espaces fonctionnels

Nous commençons donc par un rappel sur les différents espaces fonctionnels : L^p, Sobolev et BV.

2.1.1 Les espaces L^p

Les définitions et théorèmes énoncés dans cette partie et concernant les espaces L^p proviennent de [15]. Dans ce qui suit, Ω désigne un ouvert de $\mathbb{R}^N$ muni de la mesure de Lebesgue dx.

Définition 2.1.1. (Espace L^p)

Soit $p \in \mathbb{R}$ avec $1 \leq p < \infty$. On pose

$$L^p(\Omega) = \{f : \Omega \mapsto \mathbb{R}, f \text{ mesurable telle que } |f|^p \in L^1(\Omega)\}.$$

On note

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p dx\right)^{1/p}.$$

Définition 2.1.2. (Espace L^∞)

On pose

$L^\infty(\Omega) = \{f : \Omega \mapsto \mathbb{R}, f \text{ mesurable telle qu' } \exists \text{ une constante } C \text{ telle que } |f(x)| \leq C \text{ } p.p. \text{ sur } \Omega\}$.

On note

$$\|f\|_{L^\infty(\Omega)} = \inf\{C, |f(x)| \leq C \text{ } p.p. \text{ sur } \Omega\}.$$

Théorème 2.1.3.

$L^p(\Omega)$ est un espace de Banach et $\|.\|_{L^p(\Omega)}$ est une norme pour tout $1 \leq p \leq \infty$.

Nous donnons maintenant des résultats importants d'intégration.

Théorème 2.1.4. *(Théorème de convergence monotone de Beppo-Levi)*

Soit (f_n) une suite croissante de fonctions de $L^1(\Omega)$ telle que $\sup_n \int_\Omega f_n \, dx < +\infty$. Alors $f_n(x)$ converge presque partout sur Ω vers une limite finie notée $f(x)$. De plus, $f \in L^1(\Omega)$ et $\|f_n - f\|_{L^1(\Omega)} \longrightarrow 0$.

Théorème 2.1.5. *(Théorème de convergence dominée de Lebesgue)*

Soit (f_n) une suite de fonctions de $L^1(\Omega)$. On suppose que

1. *$f_n(x) \longrightarrow f(x)$ presque partout sur Ω,*
2. *il existe une fonction $g \in L^1(\Omega)$ telle que pour chaque $n \in \mathbb{N}$, $|f_n(x)| \leq g(x)$ presque partout sur Ω.*

Alors $f \in L^1(\Omega)$ et $\|f_n - f\|_{L^1(\Omega)} \longrightarrow 0$.

Lemme 2.1.6. *(Lemme de Fatou)*

Soit (f_n) une suite de fonctions de $L^1(\Omega)$ telle que

1. *pour chaque $n \in \mathbb{N}$, $f_n(x) \geq 0$ presque partout sur Ω,*
2. *$\sup_{n\in\mathbb{N}} \int_\Omega f_n < \infty$.*

Pour chaque $x \in \Omega$, on pose $f(x) = \liminf_{n\to+\infty} f_n(x)$. Alors $f \in L^1(\Omega)$ et

$$\int_\Omega f \, dx \leq \liminf_{n\to+\infty} \int_\Omega f_n \, dx.$$

Théorème 2.1.7. *(Inégalité de Hölder)*

Soient $f \in L^p(\Omega)$ et $g \in L^q(\Omega)$ avec $1 \leq p \leq \infty$ et $\frac{1}{p} + \frac{1}{q} = 1$. Alors $fg \in L^1(\Omega)$ et

$$\int_\Omega |fg| dx \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}.$$

Théorème 2.1.8. *(Inégalité de Young)*

On suppose que $1 < p < \infty$ *et* $\frac{1}{p} + \frac{1}{q} = 1$*, alors*

$$ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q \quad \forall a \geq 0, \quad \forall b \geq 0.$$

Théorème 2.1.9.

Soient (f_n) *une suite de* $L^p(\Omega)$ *et* $f \in L^p(\Omega)$*, telles que* $\|f_n - f\|_{L^p(\Omega)} \longrightarrow 0$*. Alors il existe une sous-suite* (f_{n_k}) *telle que*

1. $f_{n_k}(x) \to f(x)$ *presque partout sur* Ω*,*
2. *il existe* $h \in L^p(\Omega)$ *tel que* $|f_{n_k}(x)| \leq h(x)$ *presque partout sur* Ω *et* $\forall k \in \mathbb{N}$*.*

Théorème 2.1.10.

$L^p(\Omega)$ *est un espace réflexif pour* $1 < p < +\infty$ *et un espace séparable pour* $1 \leq p < +\infty$*. L'espace dual de* $L^p(\Omega)$ *pour* $1 < p < +\infty$ *est* $L^q(\Omega)$ *avec* q *tel que* $\frac{1}{p} + \frac{1}{q} = 1$*.*

Théorème 2.1.11. *(Propriétés générales de compacité)*

1. *Soit* X *un espace de Banach réflexif et soit* $C > 0$ *une constante réelle positive. Soit également* (u_n) *une suite de* X *telle que* $\|u_n\|_X \leq C$ *pour tout* $n \in \mathbb{N}$*. Alors il existe* $\bar{u} \in X$ *et une sous-suite* (u_{n_k}) *de* (u_n) *telles que* $u_{n_k} \underset{X}{\rightharpoonup} \bar{u}$ *quand* $k \to +\infty$*.*
2. *Soit* X *un espace de Banach séparable et soit* $C > 0$ *une constante réelle positive. Soit également* (l_n) *une suite de* X'*, l'espace dual de* X*, telle que* $\|l_n\|_{X'} \leq C$ *pour tout* $n \in \mathbb{N}$*. Alors il existe* $\bar{l} \in X'$ *et une sous-suite* (l_{n_k}) *de* (l_n) *telles que* $l_{n_k} \underset{X'}{\overset{*}{\rightharpoonup}} \bar{l}$ *quand* $k \to +\infty$*.*

2.1.2 Les espaces de Sobolev

À présent, nous rappelons les définitions et propriétés élémentaires relatives aux espaces de Sobolev, et principalement extraites de [15] et [35]. Dans ce qui suit, Ω désigne toujours un ouvert de $\mathbb{R}^N$ muni de la mesure de Lebesgue dx.

Définition 2.1.12.

On suppose que $u \in L^1_{loc}(\Omega)$. On dit que $v_i \in L^1_{loc}(\Omega)$ est la dérivée partielle faible de u par rapport à x_i dans Ω si

$$\int_\Omega u \frac{\partial \phi}{\partial x_i}\, dx = -\int_\Omega v_i \phi\, dx,$$

pour tout $\phi \in \mathcal{C}^\infty_c(\Omega)$, espace dont la définition suit (cf. Déf. 2.1.15, point 4).

Définition 2.1.13. (Espace de Sobolev)

Soit $1 \leq p \leq +\infty$. L'espace de Sobolev $W^{1,p}(\Omega)$ est défini par

$$W^{1,p}(\Omega) = \left\{ u \in L^p(\Omega) \;\middle|\; \begin{array}{c} \exists\,(g_1, \dots g_N) \in (L^p(\Omega))^N \text{ telles que} \\ \int_\Omega u \frac{\partial \phi}{\partial x_i}\, dx = -\int_\Omega g_i \phi\, dx,\ \forall \phi \in \mathcal{C}_c^\infty(\Omega), \forall i = 1, \dots, N \end{array} \right\},$$

$$= \left\{ u \in L^p(\Omega) \;\middle|\; \begin{array}{c} \forall i = 1, \dots, N, \frac{\partial u}{\partial x_i}, \text{ la dérivée faible de } u \text{ par rapport à } x_i \\ \text{existe et } \frac{\partial u}{\partial x_i} \in L^p(\Omega) \end{array} \right\}.$$

Si $u \in W^{1,p}(\Omega)$, on définit la norme associée par

$$\|u\|_{W^{1,p}(\Omega)} = \left(\|u\|_{L^p(\Omega)}^p + \|\nabla u\|_{L^p(\Omega)}^p \right)^{1/p} \text{ pour } 1 \leq p < \infty.$$

Pour une fonction $u : \Omega \mapsto \mathbb{R}^N$, $u = (u^1, \cdots, u^N)$, on dit que $u \in W^{1,p}(\Omega, \mathbb{R}^N)$ si $u^i \in W^{1,p}(\Omega)$ pour tout $i = 1, \dots, N$.

Proposition 2.1.14.

L'espace $W^{1,p}(\Omega)$ est un espace de Banach pour $1 \leq p \leq \infty$. C'est un espace réflexif pour $1 < p < \infty$ et un espace séparable pour $1 \leq p < \infty$.

Nous introduisons maintenant d'autres espaces fonctionnels intervenant naturellement dans certains théorèmes relatifs aux espaces de Sobolev (comme par exemple, les théorèmes d'injection).

Définition 2.1.15.

1. $\mathcal{C}^0(\Omega) = \mathcal{C}(\Omega)$ désigne l'ensemble des fonctions continues $u : \Omega \to \mathbb{R}$ muni de la norme $\|u\|_{\mathcal{C}^0(\Omega)} = \sup_{x \in \Omega} |u(x)|$.
2. $\mathcal{C}^0(\bar{\Omega})$ désigne l'ensemble des fonctions continues $u : \Omega \to \mathbb{R}$ qui peuvent être prolongées continûment à $\bar{\Omega}$. La norme associée est définie par $\|u\|_{\mathcal{C}^0(\bar{\Omega})} = \sup_{x \in \bar{\Omega}} |u(x)|$.
3. Le support d'une fonction $u : \Omega \to \mathbb{R}$ est défini comme $supp\ u := \overline{\{x \in \Omega : u(x) \neq 0\}}$.
4. $\mathcal{C}_c(\Omega) = \{u \in \mathcal{C}(\Omega) : supp\ u \in \Omega \text{ est compact}\}$.
5. Pour tout $k \in \mathbb{N}$, on note $\mathcal{C}^k(\Omega)$ l'espace des fonctions continues dont toutes les dérivées partielles jusqu'à l'ordre k sont également continues. La norme associée est définie par $\forall \alpha \in \mathbb{N}^N$, $\|u\|_{\mathcal{C}^k(\Omega)} = \max_{|\alpha| \leq k} \sup_{x \in \Omega} |D^\alpha u(x)|$.
6. Soit $0 \leq \alpha \leq 1$. On définit l'espace de Hölder $\mathcal{C}^{0,\alpha}(\bar{\Omega})$ comme l'ensemble des fonctions $u \in \mathcal{C}^0(\bar{\Omega})$ telles que

$$[u]_{\alpha,\bar{\Omega}} := \sup_{\substack{(x,y) \in \bar{\Omega} \times \bar{\Omega} \\ x \neq y}} \frac{|u(x) - u(y)|}{|x - y|^\alpha} < +\infty.$$

Il est muni de la norme

$$\|u\|_{\mathcal{C}^{0,\alpha}(\bar{\Omega})} := \|u\|_{\mathcal{C}^0(\bar{\Omega})} + [u]_{\alpha,\bar{\Omega}}\,.$$

Théorème 2.1.16.

On suppose que $1 \leq p \leq +\infty$.

1. ***(Dérivation d'un produit)*** *Soient* $u, v \in W^{1,p}(\Omega) \cap L^\infty(\Omega)$, *alors* $uv \in W^{1,p}(\Omega) \cap L^\infty(\Omega)$ *et*
$$\frac{\partial(uv)}{\partial x_i} = \frac{\partial u}{\partial x_i} v + u \frac{\partial v}{\partial x_i}, \quad i = 1, 2, \dots N.$$
2. ***(Règle de la chaîne)*** *Soit* $G \in C^1(\mathbb{R})$ *telle que* $G(0) = 0$ *et* $|G'(s)| \leq M, \forall s \in \mathbb{R}$. *Soit* $u \in W^{1,p}(\Omega)$, *alors*
$$G \circ u \in W^{1,p}(\Omega) \text{ et } \frac{\partial}{\partial x_i}(G \circ u) = (G' \circ u)\frac{\partial u}{\partial x_i}.$$

Théorème 2.1.17. *(Théorème de trace)*

On suppose que Ω *est un ensemble borné et de classe* C^1, *et soit* $1 \leq p < +\infty$. *Il existe un opérateur linéaire borné* $T : W^{1,p}(\Omega) \to L^p(\partial\Omega)$ *tel que* $Tu = u$ *sur* $\partial\Omega$ *pour tout* $u \in W^{1,p}(\Omega) \cap C(\bar{\Omega})$.
De plus, pour tout $\phi \in C_c^\infty(\mathbb{R}^N, \mathbb{R}^N)$ *et* $u \in W^{1,p}(\Omega)$,

$$\int_\Omega u\, div\, \phi\, dx = -\int_\Omega \nabla u \cdot \phi\, dx + \int_{\partial\Omega} (\phi \cdot \nu) Tu\, d\mathcal{H}^{N-1},$$

ν *désignant la normale extérieure unitaire à* $\partial\Omega$ *et* $\mathcal{H}^{N-1}$ *la mesure de Hausdorff de dimension* $N-1$.

Théorème 2.1.18. *(Inégalité généralisée de Poincaré, [38])*

Soit Ω *un ensemble ouvert borné Lipschitz de* $\mathbb{R}^N$. *Soit* $p \in [1, \infty)$ *et soit* $\mathcal{N}$ *une semi-norme continue de* $W^{1,p}(\Omega)$, *c'est-à-dire, une norme sur les fonctions constantes. Soit* $u \in W^{1,p}(\Omega)$. *Alors il existe une constante* $C > 0$ *qui dépend uniquement de* Ω, N *et* p, *telle que*

$$\|u\|_{W^{1,p}(\Omega)} \leq C \left(\left(\int_\Omega |\nabla u|^p\, dx \right)^{\frac{1}{p}} + \mathcal{N}(u) \right).$$

Nous appliquons ce résultat à $\mathcal{N}(u) = \int_{\partial\Omega} |u| d\sigma$.

Théorème 2.1.19. *(Théorème de Sobolev, [35, Théorème 12.11])*

Soit $\Omega \subset \mathbb{R}^N$ *un ensemble ouvert borné à frontière lipschitzienne. On a :*

— *si* $1 \leq p < N$, *alors* $W^{1,p}(\Omega) \subset L^q(\Omega)$ *pour tout* $q \in [1, p^*]$, *avec* p^* *l'exposant de Sobolev défini par* $\dfrac{1}{p^*} = \dfrac{1}{p} - \dfrac{1}{N}$. *Cela signifie que pour tout* $q \in [1, p^*]$ *il existe une constante* C *dépendant uniquement de* p, q *et* Ω *telle que* $\forall u \in W^{1,p}(\Omega)$,

$$\|u\|_{L^q(\Omega)} \leq C \|u\|_{W^{1,p}(\Omega)}.$$

— *si $p = N$, alors $W^{1,p}(\Omega) \subset L^q(\Omega)$, pour tout $q \in [1, \infty)$. De la même manière, cela signifie que pour tout $q \in [1, \infty)$, il existe une constante C dépendant uniquement de p, q et Ω telle que $\forall u \in W^{1,p}(\Omega)$,*

$$\|u\|_{L^q(\Omega)} \leq C\|u\|_{W^{1,p}(\Omega)}.$$

— *si $p > N$, alors $W^{1,p}(\Omega) \subset C^{0,\alpha}(\bar{\Omega})$ pour tout $\alpha \in [0, 1 - \frac{N}{p}]$. En particulier, il existe une constante C dépendant uniquement de p et Ω telle que $\forall u \in W^{1,p}(\Omega)$,*

$$\|u\|_{L^\infty(\Omega)} \leq C\|u\|_{W^{1,p}(\Omega)}.$$

Théorème 2.1.20. *(Théorème de Rellich-Kondrachov, [35, Théorème 12.12])*

Soit $\Omega \subset \mathbb{R}^N$ un ensemble ouvert borné à frontière lipschitzienne. On a :

— *si $1 \leq p < N$, alors l'injection de $W^{1,p}(\Omega)$ dans $L^q(\Omega)$ est compacte pour tout $q \in [1, p^*)$,*
— *si $p = N$, alors l'injection de $W^{1,p}(\Omega)$ dans $L^q(\Omega)$ est compacte pour tout $q \in [1, \infty)$,*
— *si $p > N$, alors l'injection de $W^{1,p}(\Omega)$ dans $C^{0,\alpha}(\bar{\Omega})$ est compacte pour tout $\alpha \in [0, 1 - \frac{N}{p}]$.*

2.1.3 Les espaces BV

Finalement, nous donnons des définitions et propriétés associées aux espaces des fonctions à variation bornée, issues de [38, 47]. Ces espaces fonctionnels sont particulièrement adaptés à la modélisation d'images exhibant des discontinuités le long de courbes. L'ensemble Ω désigne un ouvert de $\mathbb{R}^N$ muni de la mesure de Lebesgue dx.

Définition 2.1.21. (Espace $BV(\Omega)$)

Soit $u \in L^1(\Omega)$. u est une fonction à variation bornée dans Ω si

$$\|Du\|(\Omega) = |u|_{BV(\Omega)} := \sup \left\{ \int_\Omega u \, \mathrm{div}\phi \, dx \,|\, \phi \in C_c^1(\Omega, \mathbb{R}^N), |\phi(x)| \leq 1 \right\} < \infty.$$

$\|Du\|(\Omega)$ est appelée la variation totale de u dans Ω.

Proposition 2.1.22.

$BV(\Omega)$ est un espace de Banach muni de la norme $\|u\|_{BV(\Omega)} = \|u\|_{L^1(\Omega)} + |u|_{BV(\Omega)}$.

Définition 2.1.23. (Convergence faible-$*$ dans $BV(\Omega)$)

Soit $u \in BV(\Omega)$, $(u_n)_{n\in\mathbb{N}}$ une suite de fonctions à variation bornée dans Ω. On dit que la suite (u_n) converge faiblement-$*$ vers $u \in BV(\Omega)$ si (u_n) converge vers u dans $L^1(\Omega)$ et (Du_n) converge faiblement vers Du dans $\mathcal{M}(\Omega)$ (espace des mesures de Radon), c'est-à-dire,

$$\lim_{n\to+\infty} \|u_n - u\|_{L^1(\Omega)} = 0 \text{ et } \lim_{n\to+\infty} \int_\Omega v Du_n = \int_\Omega v Du, \forall v \in \mathcal{C}_c(\Omega, \mathbb{R}^N).$$

Théorème 2.1.24. *(Semi-continuité inférieure de la variation totale)*

On suppose que $u_n \in BV(\Omega)$ *(n=1,2,...) et* $u_n \to u \in L^1(\Omega)$*. Alors*

$$\|Du\|(\Omega) \leq \liminf_{n\to+\infty} \|Du_n\|(\Omega)$$

Théorème 2.1.25. *(Compacité)*

Soit $\Omega \subset \mathbb{R}^N$ *un ensemble ouvert borné à frontière* $\partial\Omega$ *lipschitzienne. On suppose que* $\{u_n\}_{n=1}^{\infty}$ *est une suite de* $BV(\Omega)$ *satisfaisant*

$$\sup_n \|u_n\|_{BV(\Omega)} < \infty.$$

Alors il existe une sous-suite $\{u_{n_j}\}_{j=1}^{\infty}$ *et une fonction* $u \in BV(\Omega)$ *telles que*

$$u_{n_j} \underset{j\to+\infty}{\longrightarrow} u \text{ dans } L^1(\Omega).$$

Théorème 2.1.26. *(Approximation régulière)*

Soit $\Omega \subset \mathbb{R}^N$ *un ensemble ouvert borné. Soit* $u \in BV(\Omega)$*, alors il existe une suite de fonctions* $(u_k)_{k\in\mathbb{N}}$ *de* $BV(\Omega) \cap \mathcal{C}^\infty(\Omega)$ *telles que*

1. $u_k \underset{k\to+\infty}{\longrightarrow} u$ *dans* $L^1(\Omega)$*,*
2. $\|Du_k\|(\Omega) \underset{k\to+\infty}{\longrightarrow} \|Du\|(\Omega)$*.*

Théorème 2.1.27. *(Injection)*

Soit $\Omega \subset \mathbb{R}^N$ *un ensemble ouvert borné à frontière* $\partial\Omega$ *lipschitzienne. Alors l'injection* $BV(\Omega) \subset L^{N/(N-1)}(\Omega)$ *est continue et l'injection* $BV(\Omega) \subset L^p(\Omega)$ *est compacte pour tout* $1 \leq p < \frac{N}{N-1}$*.*

Théorème 2.1.28. *(Inégalité de Poincaré-Wirtinger)*

Soit $\Omega \subset \mathbb{R}^N$ *un ensemble ouvert borné et connexe à frontière* $\partial\Omega$ *lipschitzienne. Alors il existe une constante* $C > 0$ *dépendant uniquement de* N *et* Ω *telle que*

$$\left\| u - \frac{1}{|\Omega|}\int_\Omega u(x)dx \right\|_{L^p(\Omega)} \leq C\|Du\|(\Omega), \quad \forall u \in BV(\Omega), \quad 1 \leq p \leq \frac{N}{N-1}.$$

Théorème 2.1.29. *(Formule de la co-aire)*

Soit $u \in BV(\Omega)$ *avec* $\Omega \subset \mathbb{R}^N$ *un ensemble ouvert. Alors*

1. $E_t = \{x \in \Omega \,|\, u(x) > t\}$ *a un périmètre fini, i.e.,* $\chi_{E_t} \in BV(\Omega)$ *et* $\|D\chi_{E_t}\|(\Omega)$ *est le périmètre* $P_\Omega(E_t)$ *de* E_t *dans* Ω*, pour presque tout* $t \in \mathbb{R}$*.*
2. $\|Du\|(\Omega) = \displaystyle\int_{-\infty}^{+\infty} \|D\chi_{E_t}\|(\Omega)dt.$
3. *Réciproquement, si* $u \in L^1(\Omega)$ *et* $\displaystyle\int_{-\infty}^{+\infty} \|D\chi_{E_t}\|(\Omega)dt < \infty$*, alors* $u \in BV(\Omega)$*.*

2.2 Calculs des variations

Cette section s'appuie sur le livre de B. Dacorogna ([35]) et vise à examiner l'existence voire l'unicité des minimiseurs de problèmes définis comme

$$\inf_{u\in X} I(u) = \int_\Omega f(x, u(x), \nabla u(x))dx, \tag{2.1}$$

où

- $\Omega \subset \mathbb{R}^N, N \geq 1$, est un ensemble ouvert et borné. De plus, un point dans Ω est désigné par $x = (x_1, \ldots, x_N)$;

- $u : \Omega \to \mathbb{R}^M, M \geq 1$, est la fonction inconnue avec

$$\nabla u = \left(\frac{\partial u^j}{\partial x_i}\right)_{1\leq i\leq N}^{1\leq j\leq M} \in \mathbb{R}^{M\times N};$$

- X est l'espace des fonctions admissibles ;

- $f : \Omega \times \mathbb{R}^M \times \mathbb{R}^{M\times N} \to \mathbb{R}$, $f = f(x, u, \xi)$, une fonction donnée.

Avant de présenter les étapes de la méthode directe du calcul des variations, nous redonnons la définition d'une fonction convexe.

Définition 2.2.1. (Fonction convexe)

— Une fonction $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ est dite convexe si

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y),$$

$\forall x \in \mathbb{R}^N, \forall y \in \mathbb{R}^N$, et $\forall t \in [0, 1]$.

— Une fonction $f : E \subset \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ est dite strictement convexe sur l'ensemble convexe E si

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y),$$

$\forall (x, y) \in E^2, x \neq y$, et $\forall t \in (0, 1)$.

2.2.1 Méthode directe du calcul des variations

La méthode directe du calcul des variations se décline comme suit en trois étapes, afin de prouver l'existence d'une solution au problème (2.1),

1. On construit d'abord une suite minimisante $u_n \in X$, *i.e.*, une suite satisfaisant
$$\lim_{n\to+\infty} I(u_n) \leq \inf_{x\in X} I(u).$$
2. On obtient une borne uniforme de $\|u_n\|_X$ par le biais d'une inégalité de coercivité. En effet, si I est coercive, signifiant que $\lim\limits_{\|u\|_X\to+\infty} I(u_n) = +\infty$, cette borne uniforme est directement extraite. (En argumentant par contradiction, on suppose que $\forall C > 0, \exists n \in \mathbb{N}, \|u_n\|_X > C$. On prouve, par construction qu'il existe un sous-suite (u_{n_k}) de (u_n) telle que $\lim\limits_{k\to+\infty} I(u_{n_k}) = +\infty$ en raison de la coercivité de I, ce qui contredit le fait que $\lim\limits_{k\to+\infty} I(u_{n_k}) = \inf\limits_{u\in X} I(u)$).
Si X est réflexif, alors d'après le Théorème 2.1.11, on peut trouver $\bar{u} \in X$ et une sous-suite (u_{n_k}) de (u_n) telles que $u_{n_k} \underset{X}{\rightharpoonup} \bar{u}$ quand $k \longrightarrow +\infty$.
3. Pour prouver que $\bar{u}$ est un minimiseur de I, il suffit d'établir l'inégalité
$$I(\bar{u}) \leq \liminf_{k\to+\infty} I(u_{n_k}).$$

La dernière propriété, que nous étudions juste après, est appelée semi-continuité inférieure faible. Nous nous plaçons dans le cas courant où l'espace des fonctions admissibles X est l'espace de Sobolev $u_0 + W_0^{1,p}(\Omega)$, avec $u_0 \in W^{1,p}(\Omega)$ une fonction donnée.

Définition 2.2.2. (Semi-continuité inférieure faible)

Soit $p \geq 1$ et Ω, u, f comme précédemment. On dit que I est faiblement semi-continue inférieurement dans $W^{1,p}(\Omega)$ si pour toute suite $u_n \underset{W^{1,p}(\Omega)}{\rightharpoonup} \bar{u}$, alors
$$I(\bar{u}) \leq \liminf_{n\to+\infty} I(u_n).$$

Si $p = \infty$, on dit que I est faiblement-$*$ semi-continue inférieurement dans $W^{1,\infty}(\Omega)$ si la même inégalité est valable pour toute suite $u_n \underset{W^{1,\infty}(\Omega)}{\overset{*}{\rightharpoonup}} \bar{u}$.

2.2.2 Γ-convergence

Définition 2.2.3. (Γ-convergence)

Soit (X, D) un espace métrique. On dit qu'une suite $F_j : X \to [-\infty, +\infty]$ Γ-converge vers $F : X \to [-\infty, +\infty]$ (lorsque $j \to +\infty$) si pour tout $u \in X$ on a

1. (inégalité lim inf) pour toute suite $(u_j) \subset X$ qui converge vers u,
$$F(u) \leq \liminf_{j\to+\infty} F_j(u_j);$$
2. (suite recouvrante) il existe une suite $(u_j) \subset X$ qui converge vers u telle que
$$F(u) \geq \limsup_{j\to+\infty} F_j(u_j).$$

La fonction F est appelée la Γ-limite de (F_j) (par rapport à D) et on écrit $F = \Gamma - \lim_j F_j$.

Le théorème suivant est fondamental pour la convergence de certaines approximations.

Théorème 2.2.4. *(Théorème fondamental de Γ-convergence)*

On suppose que $F = \Gamma - \lim_j F_j$, et qu'il existe un ensemble compact $C \subset F$ tel que $\inf_X F_j = \inf_C F_j$ pour tout j. Alors il existe un minimum de F sur X tel que

$$\min_X F = \lim_j \inf_X F_j,$$

et si $(u_j) \subset X$ est une suite convergente telle que $\lim_j F_j(u_j) = \lim_j \inf_X F_j$, alors sa limite est le point minimum de F.

2.3 Élasticité non linéaire

Dans cette partie, nous présentons des éléments issus de la théorie de l'élasticité non linéaire, sur laquelle est bâti le modèle du Chapitre 4, et en particulier les théorèmes de J. Ball [8].

Remarque 2.3.1. La préservation de l'orientation correspond à la condition

$$\det \nabla u(x) > 0 \text{ presque partout}, x \in \Omega.$$

Théorème 2.3.2. *([8, Théorème 1])*

Soit $\Omega \subset \mathbb{R}^N$ un ensemble ouvert non vide, borné, connexe, fortement Lipschitz. Soit $u_0 : \bar{\Omega} \mapsto \mathbb{R}^N$ continue dans $\bar{\Omega}$ et bijective dans Ω. Soit $p > N$ et soit $u \in W^{1,p}(\Omega, \mathbb{R}^N)$ à valeurs dans $\mathbb{R}^N$ et qui satisfait $u|_{\partial\Omega} = u_0|_{\partial\Omega}$, $\det \nabla u(x) > 0$ presque partout dans Ω. Alors

1. *$u(\bar{\Omega}) = u_0(\bar{\Omega})$,*
2. *u fait correspondre des ensembles mesurables dans $\bar{\Omega}$ à des ensembles mesurables dans $u_0(\bar{\Omega})$, et la formule de changement de variable*

$$\int_A f(u(x)) \det \nabla u(x)\, dx = \int_{u(A)} f(v)\, dv \tag{2.2}$$

   est valable pour tout ensemble mesurable $A \subset \bar{\Omega}$ et toute fonction mesurable $f : \mathbb{R}^N \to \mathbb{R}$, à condition que l'une des intégrales dans (2.2) existe.

3. *u est bijective presque partout, i.e., l'ensemble*

$$S = \Big\{ v \in u_0(\bar{\Omega}) : u^{-1}(v) \text{ contient plus d'un élément} \Big\}$$

   est de mesure nulle,

4. si $v \in u_0(\Omega)$, alors $u^{-1}(v)$ est un continuum contenu dans Ω, tandis que si $v \in \partial u_0(\Omega)$ alors chaque composante connexe de $u^{-1}(v)$ intersecte $\partial\Omega$.

Théorème 2.3.3. *([8, Théorème 2])*

On suppose que les hypothèses du Théorème 2.3.2 sont vérifiées. Soit $u_0(\Omega)$ satisfaisant la condition du cône. Supposons que pour un certain $q > N$,

$$\int_\Omega |\nabla u^{-1}(x)|^q \det \nabla u(x) dx < \infty. \tag{2.3}$$

Alors u est un homéomorphisme de Ω dans $u_0(\Omega)$, et la fonction inverse $x(u)$ appartient à $W^{1,q}(u_0(\Omega))$. La matrice des dérivées faibles de $x(\cdot)$ est donnée par

$$\nabla x(v) = \nabla u^{-1}(x(v)) \text{ presque partout dans } u_0(\Omega).$$

Si, en outre, $u_0(\Omega)$ est fortement Lipschitz, alors u est un homéomorphisme de $\bar{\Omega}$ dans $u_0(\bar{\Omega})$.

Théorème 2.3.4. *(Extrait de [35, Théorème 8.20])*

Soit $\Omega \subset \mathbb{R}^N$ un ensemble ouvert borné, $1 < p < \infty$, et soit

$$u_n \rightharpoonup u \text{ dans } W^{1,p}(\Omega, \mathbb{R}^N).$$

Soit $N = 2$, alors

— *si $p \geq 2$*

$$\det \nabla u_n \rightharpoonup \det \nabla u \text{ dans } \mathcal{D}'(\Omega);$$

— *si $p > 2$*

$$\det \nabla u_n \rightharpoonup \det \nabla u \text{ dans } L^{p/2}(\Omega).$$

2.4 Éléments d'optimisation convexe

Pour finir, nous rappelons quelques éléments sur les méthodes proximales pour l'optimisation convexe et fortement inspirés de [6, 34]. Ces éléments irriguent les Chapitres 3 et 4.

Nous désignons par $\mathbb{R}^N$ l'espace euclidien classique de dimension N muni de la norme usuelle $\|\cdot\|$.

Définition 2.4.1.

- Le domaine d'une fonction $f : \mathbb{R}^N \to]-\infty, +\infty]$ est défini par

$$\operatorname{dom} f = \left\{ x \in \mathbb{R}^N \mid f(x) < +\infty \right\}.$$

- La classe des fonctions convexes semi-continues inférieurement de $\mathbb{R}^N$ dans $]-\infty,+\infty]$ qui sont propres (*i.e.*, avec un domaine non vide) est désignée par $\Gamma_0(\mathbb{R}^N)$.
- Le sous-différentiel de f est l'opérateur défini par

$$\begin{array}{rccl} \partial f: & \mathbb{R}^N & \to & 2^{\mathbb{R}^N} \\ & x & \mapsto & \left\{u \in \mathbb{R}^N \,|\, \forall y \in \mathbb{R}^N,\, f(y) \geq f(x) + (y-x)^T u\right\}. \end{array}$$

- Soit C un sous-ensemble convexe non vide de $\mathbb{R}^N$. La fonction indicatrice de C est

$$i_C : x \mapsto \left\{ \begin{array}{ll} 0 & \text{if} \quad x \in C \\ +\infty & \text{if} \quad x \notin C. \end{array} \right.$$

- La distance de $x \in \mathbb{R}^N$ à C est définie par la fonction $d_C : \mathbb{R}^N \to [0,+\infty[: x \mapsto \inf_{y \in C} \|x-y\|$. Si l'ensemble C est fermé et convexe, la projection de $x \in \mathbb{R}^N$ sur C est l'unique point $P_C\, x$ tel que $d_C(x) = \|x - P_C\, x\|$.

La projection $P_C\, x$ de $x \in \mathbb{R}^N$ sur l'ensemble fermé convexe non vide $C \subset \mathbb{R}^N$ peut être appréhendée comme la solution du problème

$$\arg\min_{y \in \mathbb{R}^N} i_C(y) + \frac{1}{2}\|x-y\|^2,$$

la fonction i_C étant un élément de $\Gamma_0(\mathbb{R}^N)$ du fait des hypothèses sur C.
Cette formulation conduit Moreau ([97]) à étendre la notion de projection dans laquelle une fonction arbitraire $f \in \Gamma_0(\mathbb{R}^N)$ est maintenant un substitut à i_C. Nous obtenons ainsi la définition de l'opérateur dit proximal ou de proximité.

Définition 2.4.2. (Opérateur proximal, extraite de [34, Définition 10.1])

Soit $f \in \Gamma_0(\mathbb{R}^N)$. Pour tout $x \in \mathbb{R}^N$, le problème de minimisation

$$\min_{y \in \mathbb{R}^N} f(y) + \frac{1}{2}\|x-y\|^2$$

admet une solution unique désignée par $\text{prox}_f\, x$. L'opérateur $\text{prox}_f : \mathbb{R}^N \to \mathbb{R}^N$ définit ce qu'on appelle un opérateur proximal.

Soit $f \in \Gamma_0(\mathbb{R}^N)$. L'opérateur proximal de f est caractérisé par :

$$\left(\forall (x,p) \in \mathbb{R}^N \times \mathbb{R}^N\right),\ \ p = \text{prox}_f\, x \iff x - p \in \partial f(p),$$

qui se réduit à

$$\left(\forall (x,p) \in \mathbb{R}^N \times \mathbb{R}^N\right),\ \ p = \text{prox}_f\, x \iff x - p \in \nabla f(p),$$

si f est différentiable. L'opérateur proximal jouit de bonnes propriétés particulièrement bien adaptées aux algorithmes itératifs de minimisation. Par exemple, $\mathrm{prox}_f\, x$ est fermement non expansif, *i.e.*,

$$(\forall x \in \mathbb{R}^N),\ (\forall y \in \mathbb{R}^N),\ \|\mathrm{prox}_f\, x - \mathrm{prox}_f\, y\|^2 + \|(x - \mathrm{prox}_f\, x) - (y - \mathrm{prox}_f\, y)\|^2 \leq \|x - y\|^2,$$

et son ensemble de points fixes est précisément l'ensemble des minimiseurs de f.

Nous terminons par l'introduction d'un algorithme itératif classique pour un problème de minimisation. L'algorithme *forward-backward*, présenté dans l'Algorithme 1, s'appuie sur les opérateurs proximaux pour minimiser un problème de la forme

$$\min_{x \in \mathbb{R}^N}\ f_1(x) + f_2(x),$$

avec $f_1 \in \Gamma_0(\mathbb{R}^N)$ et f_2 une fonction de $\mathbb{R}^N$ dans $]-\infty, +\infty]$, convexe et différentiable à gradient continu β-lipschitzien. La majorité des algorithmes itératifs de minimisation découle de celui-ci.

Algorithme 1 Algorithme forward-backward (Extrait de [34, Algorithme 3.2]) : problème générique $\min_{x \in \mathbb{R}^N}\ f_1(x) + f_2(x)$ avec $f_1 \in \Gamma_0(\mathbb{R}^N)$ et f_2 une fonction de $\mathbb{R}^N$ dans $]-\infty, +\infty]$, convexe et différentiable à gradient continu β-lipschitzien.

Fixer $\varepsilon \in]0, \min\{1, 1/\beta\}[$, $x_0 \in \mathbb{R}^N$
pour $n = 0, 1, \cdots$ **faire**
  $\gamma_n \in [\varepsilon, 2/\beta - \varepsilon]$
  $y_n = x_n - \gamma_n \nabla f_2(x_n)$
  $\lambda_n \in [\varepsilon, 1]$
  $x_{n+1} = x_n + \lambda_n \left(\mathrm{prox}_{\gamma_n f_1}\, y_n - x_n\right)$.
fin pour

CHAPITRE

3

Inclusion de contraintes géométriques dans les réseaux de neurones convolutifs

L'inclusion d'informations *a priori* dans un processus de segmentation, qu'il s'agisse de contraintes géométriques comme la pénalisation de volume, la convexité partielle de la frontière d'un objet, ou de prescriptions topologiques pour préserver les relations contextuelles entre les objets, permet d'améliorer la précision des segmentations d'images médicales. En particulier lorsqu'on fait face au problème de contours flous et peu contrastés. Motivée par ce constat, la contribution proposée dans ce chapitre vise à fournir un cadre variationnel unifié dans l'apprentissage des réseaux de neurones convolutifs pour inclure des contraintes géométriques sous la forme d'une pénalité dans la fonction de perte. Ces contraintes géométriques prennent plusieurs formes et englobent l'alignement des courbes de niveau par l'intégration de la variation totale pondérée, une pénalisation de la surface formulée comme une contrainte dure dans la modélisation, et un critère d'homogénéité de l'intensité lumineuse basé sur une combinaison de la fonction perte standard de Dice avec le modèle constant par morceaux de Mumford-Shah. La formulation mathématique conduit à un problème d'optimisation non convexe et non lisse, ce qui exclut de fait les techniques d'optimisation lisses classiques, et nous amène à adopter un cadre lagrangien. L'application s'inscrit dans le cadre de la segmentation des organes à risque dans des scanners thoraciques, dans un contexte de planification d'un traitement de radiothérapie. Nos simulations numériques démontrent que notre méthode fournit des améliorations significatives (i) par rapport aux approches non contraintes existantes, tant en termes de critères quantitatifs, tels que la mesure de superposition, que d'évaluations qualitatives (régularisation spatiale/cohérence, réduction des valeurs aberrantes) et (ii) par rapport aux réseaux convolutifs profonds directement contraints dans une couche connectée. Elle présente de plus un certain degré de flexibilité dans la mesure où elle peut s'adapter à d'autres contraintes.

Cette contribution a conduit à deux publications :

1. Z. Lambert, C. Le Guyader, C. Petitjean. Enforcing Geometrical Priors in Deep Networks for Semantic Segmentation Applied to Radiotherapy Planning. Journal of Mathematical Imaging and Vision (JMIV), pages 1-24.

2. Z. Lambert, C. Le Guyader, C. Petitjean. A geometrically-constrained deep network for CT image segmentation. Dans 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pages 29-33. [1]

1. Cet article est une version préliminaire de la contribution précédente, dans lequel seul le cas binaire est traité et les résultats théoriques manquent entre autres.

3.1 Introduction

Aujourd'hui les méthodes *Deep-Learning* (DL), et en particulier les réseaux de neurones convolutifs profonds (CNN), font figure d'état de l'art pour de nombreuses applications de segmentation d'images médicales ([9,11]). L'entraînement d'un CNN consiste en l'apprentissage de ses paramètres via l'optimisation d'une fonction de perte différentiable, telle que le Dice ou l'entropie croisée, avec une technique de descente de gradient. Le rôle de la fonction de perte est de contraindre la carte de segmentation prédite à se rapprocher le plus possible de la carte vérité terrain et par conséquent, son expression est principalement dictée par les données. Néanmoins, la littérature scientifique montre que l'incorporation d'une petite quantité d'informations haut-niveau dans un processus de segmentation permet d'obtenir des résultats plus précis. Comme nous l'avons développé dans le Chapitre 1, l'inclusion d'*a priori* géométriques dans un contexte d'apprentissage profond peut s'effectuer dans un schéma séquentiel, c'est-à-dire comme étape de pré- ou post-traitement à la segmentation [12]. Tandis que, dans [129], Taylor *et al.* imbriquent le cadre variationnel avec celui de l'apprentissage profond : en effet, dans la phase d'entraînement du CNN, ils remplacent la descente de gradient par la méthode des multiplicateurs à direction alternée (ADMM). Par contraste, nous proposons dans notre approche, qui articule les deux cadres de manière entremêlée, un modèle variationnel unifié dans le processus d'apprentissage et incluant des *a priori* géométriques. Pour ce faire, nous nous appuyons sur une fonction de perte ad hoc impliquant que l'apprentissage du réseau est dicté et guidé par ces contraintes.
Motivés par le travail de Peng *et al.* [107], nous proposons d'introduire des contraintes géométriques dans l'entraînement d'un CNN grâce à la conception d'une fonction de perte appropriée, qui inclut (i) un critère d'alignement des contours via un terme de variation totale pondérée, (ii) une pénalisation d'aire, (iii) une composante assurant l'homogénéité de l'intensité lumineuse des pixels, en supplément de l'appariement standard des intensités, classiquement modélisé par une fonction objectif de Dice ou d'entropie croisée. L'optimisation de cette fonction de perte donne un problème de minimisation non convexe et non lisse, que nous proposons de résoudre comme suit.

Dans un premier temps, nous séparons le problème non convexe de celui non lisse en introduisant une variable auxiliaire, et nous adoptons ensuite une approche de séparation des variables, réalisée avec un algorithme ADMM [51]. Le schéma alternatif résultant se montre facile à implémenter (le plus souvent avec des solutions *closed form*) et consiste en deux sous-problèmes : (i) un premier lisse et non convexe : il s'agit de l'optimisation des paramètres du réseau réalisée par une méthode de descente de gradient stochastique (SGD) ; (ii) un second convexe mais non lisse, en raison de l'application des contraintes dures et du terme de variation totale pondérée. L'optimisation de ce dernier sous-problème se montre difficile et nous proposons deux approches distinctes pour le résoudre : l'une offre plus de flexibilité dans la conception de la fonction objectif équivalente (puisque la différentiabilité n'est pas requise) ainsi qu'une garantie de convergence, tandis que

la seconde hérite également de cette bonne propriété de convergence mais se révèle de surcroît plus efficace en termes de calcul. La première approche s'appuie sur une méthode proximale de séparation des variables [34], appelée algorithme de Douglas-Rachford (DR). L'adjectif *proximal* signifie que chaque fonction non lisse est impliquée via son opérateur proximal, garantissant la convergence théorique du sous-problème. La seconde approche repose sur l'algorithme hybride Primal-Dual (PD), et se trouve également motivée par des résultats mathématiques théoriques.

Ce schéma alternatif unifie donc les deux formalismes, approches d'apprentissage profond et modèles variationnels, dans un cadre unique : la chaîne de traitement n'est pas un enchaînement séquentiel entre une partie fondée sur l'apprentissage profond, qui fournirait une segmentation estimée, et une étape de post-traitement réalisée dans un cadre variationnel. En partageant les représentations entre les tâches et en les entrelaçant judicieusement, nous créons une synergie qui accroît la précision des résultats, tout en obtenant de meilleures capacités de généralisation. Un autre point intéressant tient au fait que cette approche, à la différence de [107], concilie la nature intrinsèquement discrète de la segmentation —qui consiste à attribuer une étiquette à chaque pixel de l'image —avec la dimension continue des méthodes variationnelles, comme nous le verrons plus loin. Pour le dire autrement, les étiquettes qui sont discrètes par essence se trouvent approximées par des variables continues.

Une tâche en segmentation d'images médicales qui peut bénéficier de telles contraintes géométriques dans la fonction de perte est la délinéation des organes à risque thoraciques dans des images de scanner. Cette segmentation est nécessaire pour planifier la radiothérapie, dans le but de déterminer la dosimétrie et de manière à épargner les OAR des rayons. Comme observé sur la Figure 3.1, les OAR montrent des tailles inégales : le cœur est volumineux comparé aux trois autres organes. Ainsi, en pénalisant l'aire, nous pouvons espérer guider le processus de segmentation et accroître sa précision. Par ailleurs, les OAR sont localisés dans une région qui exhibe un faible contraste : par exemple les frontières de l'œsophage apparaissent presque invisibles. L'approche contrainte proposée, en imposant l'homogénéité et l'alignement des bords par rapport à la vérité terrain, devrait pouvoir se montrer bénéfique pour la segmentation des OAR, et elle sera évaluée sur le dataset public SegTHOR contenant des scanners de patients atteints de cancer du poumon.

Ainsi, nos contributions revêtent différentes formes : (i) d'abord, de nature méthodologique. Dans un cadre imbriqué *'CNN/algorithmes proximaux'*, et non pas dans une simple articulation séquentielle, nous concevons une fonction de perte adaptée encodant plusieurs critères géométriques. Elle inclut l'alignement des contours via un terme de variation totale pondérée, l'application de l'homogénéité de l'intensité lumineuse des pixels par le biais d'un terme constant par morceaux de type Mumford-Shah (MS), et le respect de contraintes d'aire prescrites, formulées comme des contraintes d'égalité (*i.e.* contraintes dures) dans le problème de minimisation (où classiquement, des contraintes souples sont souvent un substitut à ces contraintes dures, engendrant des algorithmes

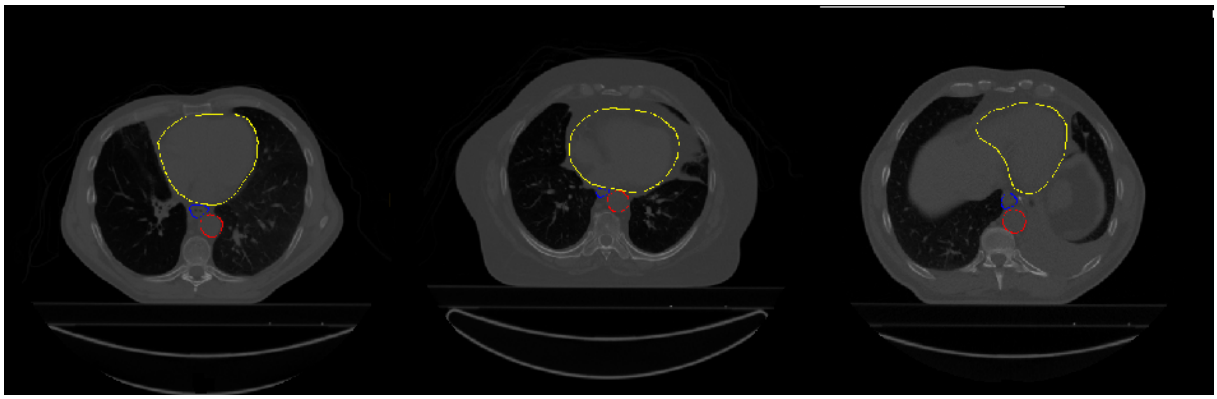


FIGURE 3.1 – Exemple d'images de scanner de trois patients différents issues du dataset SegTHOR, avec la superposition des contours de trois organes (ici œsophage (bleu), cœur (jaune) et aorte (rouge)) délimités par un expert.

moins complexes). L'objectif sous-jacent réside dans l'évaluation du bénéfice d'appliquer des contraintes géométriques dans la phase d'entraînement ; (ii) ensuite, de nature plus théorique. Notre algorithme imbriqué CNN+ADMM implique un sous-problème convexe pour lequel l'existence de minimiseurs est assurée, ainsi qu'un algorithme afférent réalisable sur le plan computationnel, et pour lequel plusieurs résultats théoriques sont fournis (existence d'un point-selle, résultat de convergence, etc.) garantissant son caractère bien posé. À noter cependant, que la question de la convergence globale de notre combinaison descente de gradient par lots + ADMM impliquant des fonctions objectifs non convexes reste toujours une question ouverte, bien au-delà de la portée de la contribution proposée ; (iii) enfin, de nature plus appliquée, avec des évaluations approfondies sur des images thoraciques de tomodensitométrie issues du jeu de données SegTHOR décrit dans le Chapitre 1, prenant en compte la précision par rapport aux mesures quantitatives classiques, une évaluation qualitative et des comparaisons avec des stratégies récentes (régularisation et contraintes incluses dans la structure du réseau).

3.2 Comparaison avec l'état de l'art

Dans la Section 1.2.4, nous avons mis en exergue le bien-fondé de l'inclusion d'informations *a priori* dans les algorithmes de segmentation d'images, notamment dans un contexte médical dans le but d'assurer une certaine cohérence anatomique. Une difficulté à cette inclusion au sein d'un processus d'apprentissage profond se situe dans la nécessité de formuler des éléments différentiables, afin d'effectuer efficacement la rétropropagation de l'erreur du gradient à travers l'ensemble du réseau. Malgré cet obstacle, un nombre croissant de travaux bénéficiant de l'intégration de connaissances préalables émerge. Nous rappelons très brièvement les trois grandes familles de stratégies recensées dans le Chapitre 1, inspirées de méthodes variationnelles ayant démontré leur efficacité, et nous soulignons

les principales différences et nouveautés avec notre méthode :

1. **L'utilisation des méthodes traditionnelles de modèles de formes et de modèles déformables dans un schéma séquentiel ([12,67,117]).** Contrairement à ces approches qui articulent les techniques fondées sur l'apprentissage profond et les modèles variationnels dans un schéma séquentiel —que la partie variationnelle serve d'étape de prédiction/estimation, ou qu'elle permette d'affiner le résultat produit par le CNN —, le travail que nous proposons vise à créer une synergie en entrelaçant les deux formalismes apprentissage profond/méthodes variationnelles. Les contraintes géométriques imposées dans le bloc variationnel alimentent l'entraînement du CNN. Cette démarche est en phase avec la philosophie des travaux récents qui tentent de tirer parti des deux techniques dans un cadre unifié permettant un entraînement de bout en bout. Zheng *et al.* ont été pionniers et proposent dans [148] de formuler les CRF comme des réseaux de neurones récurrents à insérer en fin de CNN, donnant lieu à un réseau profond entraînable de bout en bout. Fu *et al.* [48] appliquent cette idée à la segmentation des vaisseaux de la rétine.
2. **La conception d'une fonction de perte pouvant intégrer des connaissances préalables ([32, 39, 45, 53, 69, 70, 107]).** C'est le paradigme dans lequel s'inscrit notre contribution. En comparaison avec les travaux susmentionnés, la nouveauté de notre contribution repose sur plusieurs points : (i) d'abord, la conception d'une fonction de perte englobant des critères locaux et globaux. En effet, nous nous appuyons sur (a) la variation totale pondérée qui promeut à la fois l'alignement entre la carte de probabilités reconstruite et la vérité terrain ainsi qu'une régularisation sur les contours, (b) un terme fondé sur l'intensité lumineuse des régions, inspiré du modèle de Mumford-Shah constant par morceaux, qui garantit la cohérence spatiale/homogénéité et (c) le terme de Dice qui tient compte des informations à la fois localement et globalement ; (ii) ensuite, le problème de minimisation résultant est sujet à des contraintes géométriques (pénalisation d'aire, critère de non-chevauchement des phases) qui sont exprimées comme des contraintes dures (contrairement par exemple à [69] dans lequel la contrainte de taille est relaxée). Cela signifie, qu'en pratique, ces contraintes sont théoriquement exactement remplies par la variable auxiliaire mimant la carte de probabilités reconstruite, en vertu du résultat de convergence déduit. Par ailleurs, le problème de minimisation est résolu à l'aide de techniques d'optimisation originales, par comparaison avec les papiers cités ci-dessus, et plus précisément, avec des algorithmes qui s'appuient sur les opérateurs proximaux, ce qui diffère du travail de Kim *et al.* [70] par exemple. En effet, dans [70], la fonction de perte formulée à partir de la fonctionnelle de Mumford-Shah est différentiable (—en particulier, la semi-norme $W^{1,1}$ est un substitut à la variation totale (TV) classique, et par conséquent une approximation de la variation totale est traitée numériquement —), ce qui aboutit à une minimisation traditionnelle par rétropropagation pendant l'apprentissage ; (iii) enfin, plusieurs résultats théoriques mettent en évidence le caractère bien posé de notre modèle, aspect qui fait quelque peu défaut dans les articles susmentionnés.

3. **La modification de la structure du réseau pour intégrer les divers *a priori* envisagés ([44,52,131]), et ainsi entrelacer les cadres variationnel et d'apprentissage profond le cas échéant.** Cette stratégie permet de tirer profit des informations préalables durant la phase d'inférence, ce qui constitue un intérêt majeur. En particulier, Liu *et al.* façonnent une nouvelle fonction d'activation softmax dans [87], en s'appuyant sur une variable duale, pour régulariser les sorties du CNN via la TV ou une contrainte de volume. Les avantages de cette approche, par rapport à la conception de fonctions de perte spécifiques, résident dans le fait que la sortie de l'ultime couche du réseau répond théoriquement aux contraintes prescrites. Il faut noter cependant, que si dans la phase d'apprentissage les *a priori* géométriques tels que la prescription de volume peuvent être calculés à partir de la vérité terrain, ce n'est plus le cas dans la phase d'inférence, nécessitant alors l'utilisation de données statistiques, et donc de valeurs approximatives. Les inconvénients reposent sur le fait que la différentiabilité par rapport à l'ensemble des paramètres est nécessaire, ce qui, en termes de modélisation mathématique, peut être un facteur limitant et conduit à des approximations successives (puisqu'en pratique la règle de la chaîne peut être appliquée plusieurs fois), ainsi qu'à des algorithmes imbriqués.
   Afin d'établir des comparaisons avec notre proposition, cette idée d'intégrer la régularité spatiale directement dans le CNN est réinvestie dans la section 3.5.

Nous introduisons maintenant de manière détaillée le modèle conjoint proposé et fondé sur l'apprentissage profond et les approches variationnelles.

3.3 Modèle conjoint proposé fondé sur l'apprentissage profond et les approches variationnelles

3.3.1 Notations

Cadre discret Nous commençons par introduire le cadre discret que nous utiliserons tout au long des sections suivantes, et nous suivons principalement les notations de [18]. Pour simplifier, nous appréhendons les images comme des matrices en deux dimensions et définies sur une grille cartésienne régulière $\mathcal{G}$ de taille $N \times N$:

$$\mathcal{G} = \{(ih, jh) \mid 1 \leq i \leq N,\, 1 \leq j \leq N\},$$

h désigne la taille de l'espacement (classiquement pris à 1), tandis que la paire (i, j) indique les indices de la localisation discrète (ih, jh) dans le domaine image. Nous prenons donc $h = 1$, de sorte que $\mathcal{G}$ est réduit à $\mathcal{G} = \{(i, j) \mid 1 \leq i \leq N,\, 1 \leq j \leq N\}$. L'adaptation aux autres cas (3D par exemple) s'effectue directement.

Nous définissons par X l'espace euclidien $\mathbb{R}^{N\times N}$ muni du produit scalaire standard

$$\langle u, v\rangle_X = \sum_{(i,j)\in\mathcal{G}} u_{i,j}v_{i,j}.$$

Ensuite, si $u \in X$, un gradient discrétisé (par différences finies *forward*), ∇u, est un vecteur de $Y = X \times X$ défini par $(\nabla u)_{i,j} = \left((\nabla u)^1_{i,j}, (\nabla u)^2_{i,j}\right)$ avec

$$(\nabla u)^1_{i,j} = \begin{cases} u_{i+1,j} - u_{i,j} & \text{if} \quad i < N \\ 0 & \text{if} \quad i = N, \end{cases} \tag{3.1}$$

$$(\nabla u)^2_{i,j} = \begin{cases} u_{i,j+1} - u_{i,j} & \text{if} \quad j < N \\ 0 & \text{if} \quad j = N \end{cases}. \tag{3.2}$$

Nous définissons également le produit scalaire sur Y par

$$\langle p, q\rangle_Y = \sum_{(i,j)\in\mathcal{G}} \left(p^1_{i,j}q^1_{i,j} + p^2_{i,j}q^2_{i,j}\right),$$

avec $p = (p^1, p^2) \in Y$ et $q = (q^1, q^2) \in Y$. Enfin, nous posons $|z| := \sqrt{z_1^2 + z_2^2}$ pour tout $z = (z_1, z_2) \in \mathbb{R}^2$.

L'opérateur de divergence discret $\operatorname{div} : Y \to X$ (défini de manière analogue au cadre continu) est l'opposé de l'opérateur adjoint de l'opérateur de gradient ∇, c'est-à-dire :

$$\forall p \in Y,\ \forall u \in X,\ \langle -\operatorname{div} p, u\rangle_X = \langle p, \nabla u\rangle_Y,$$

$\operatorname{div} p$ étant ainsi donné par

$$(\operatorname{div} p)_{i,j} = \begin{cases} p^1_{i,j} - p^1_{i-1,j} & \text{if} \quad 1 < i < N \\ p^1_{i,j} & \text{if} \quad i = 1 \\ -p^1_{i-1,j} & \text{if} \quad i = N \end{cases} + \begin{cases} p^2_{i,j} - p^2_{i,j-1} & \text{if} \quad 1 < j < N \\ p^2_{i,j} & \text{if} \quad j = 1 \\ -p^2_{i,j-1} & \text{if} \quad j = N \end{cases}. \tag{3.3}$$

Formalisme CNN Si le formalisme des réseaux de neurones convolutifs a déjà été introduit plus tôt dans la sous-section 1.2.2, nous y revenons rapidement afin d'introduire les notations utilisées dans notre modélisation. Ainsi, dans sa forme la plus simple un CNN se schématise comme suit (se référer à [66, 87] pour plus de détails).

En considérant $i^0 : \mathcal{G} \to \mathbb{R}$ l'entrée du réseau, alors la sortie du réseau $i^T : \mathcal{G} \to [0, 1]^L$, avec L le nombre de classes, peut être appréhendée comme la résultante de l'action d'un opérateur non linéaire paramétré $\mathcal{F}_\theta$, soit $i^T = \mathcal{F}_\theta(i^0)$, avec θ l'ensemble des paramètres inconnus à apprendre. Cette variable i^T représente une fonction de classification dont la composante $i^T_l(x)$ retourne la probabilité d'un pixel x d'appartenir à la $l^{\text{ième}}$ classe. Plus

précisément, la sortie i^T d'un tel réseau résulte de l'application de T couches connectées récursives qui peuvent être exprimées comme suit :

$$\left\{ \begin{array}{rcl} o^s & = & \mathcal{T}_{\theta^{s-1}}(i^{s-1}) \\ i^s & = & \mathcal{A}^s(o^s) \end{array} \right. , \ s = 1, \cdots, T,$$

avec $\mathcal{A}^s$, soit une fonction d'activation, soit une opération de sous-échantillonnage/sur-échantillonnage ou encore une composition des deux éléments, et avec $\mathcal{T}_{\theta^{s-1}}$ un opérateur donné réalisant la connexion entre la $s^{\text{ième}}$ couche et la couche précédente.
Dans notre problème spécifique, ainsi que dans le reste de ce chapitre, nous modélisons la sortie du réseau par une fonction de segmentation $s(\theta)$, de sorte que sa $l^{\text{ième}}$ composante représente la probabilité $s^l(\theta)_{i,j}$ de chaque pixel (i, j) du domaine image discret $\mathcal{G}$ d'appartenir à la classe l.

De plus, l'entraînement d'un CNN nécessite de considérer un jeu de données de K images 2D appartenant à X, accompagnées de leurs cartes de vérité terrain associées, et désignées par $\left\{y^k\right\}_{k=1,\cdots,K}$, avec $y^k \in \{1, \cdots, L\}$ où L correspond toujours au nombre de classes. Ainsi $y^{l,k} \in \{0, l\}$ représente la version binaire de la vérité terrain avec $l \in \{1, \cdots, L\}$. Le problème étant séparable par rapport à la variable k, nous omettons la dépendance à k à partir de maintenant. Finalement, l'optimisation des paramètres du réseau s'effectue par rétropropagation de l'erreur calculée entre la carte de probabilité prédite $s^l(\theta)$ et la vérité terrain correspondante y^l à travers toutes les couches du réseau.

Munis de ces notations, nous pouvons à présent introduire le problème d'optimisation lié à la fonction de perte.

3.3.2 Conception d'une fonction de perte fondée sur la variation totale pondérée et sujette à une contrainte d'aire

Nous formulons un problème d'optimisation fondé sur la version discrète de la variation totale pondérée (WTV, *Weighted Total Variation*) dans le but de favoriser l'alignement des courbes de niveau, soit l'alignement des frontières entre la segmentation prédite et la vérité terrain, et sujet à des contraintes discrètes incluant une pénalisation d'aire par le biais d'une contrainte dure.

Variation totale pondérée Le terme de variation totale pondérée, mathématiquement désigné par TV_{g^l}, permet de procéder à l'unification du modèle de Rudin-Osher-Fatemi pour la restauration d'images ([115]) et du modèle de contour actif ([17]), comme démontré par Bresson *et al.* dans [14], lorsque le terme de pondération est donné par une fonction de détection de bords. Notons $g^l : \mathbb{R}^+ \to \mathbb{R}^+$ une fonction de détection de bords

relativement à la classe numérotée l satisfaisant $g^l(0) = 1$, avec g^l strictement décroissante et $\lim_{r\to+\infty} g^l(r) = 0$. Un choix commun pour une telle fonction de détection est $g^l(\zeta) = \frac{1}{1+\beta|\zeta|^2}$, où β représente un paramètre qui détermine le niveau de détail de la segmentation. Nous appliquons cette fonction de détection de bords à la norme du gradient de la vérité terrain binaire en chaque nœud $(i,j) \in \mathcal{G}$, $g^l(|(\nabla y^l)_{i,j}|)$. Dans un souci de concision, nous posons $g^l := (g^l_{i,j})_{i,j} = (g^l(|(\nabla y^l)_{i,j}|))_{i,j}$. Aussi, à des fins théoriques, nous supposons que $\forall l \in \{1, \cdots, L\}$, $\forall (i,j) \in \mathcal{G}$, $g^l_{i,j} \geq \varepsilon$, avec ε un paramètre positif assez petit, ce qui n'est pas restrictif. Les fonctions g^l, $l \in \{1, \cdots, L\}$, forment donc des nouvelles données en entrée de notre algorithme. Ces fonctions devraient permettre d'influencer la mise à jour des paramètres du réseau lors de la phase d'entraînement.

Avant d'énoncer une définition théorique de la WTV dans la Définition 3.3.1, nous en donnons d'abord une représentation intuitive. À noter que des travaux antérieurs (tels que [14]) affirment que par comparaison à la variation totale seule, la variation totale pondérée contribue à générer des résultats qui préservent mieux la géométrie et qui se révèlent plus fidèles aux caractéristiques originales. Dans notre cas, la variation totale pondérée TV_{g^l} impose la nature constante par morceaux de l'inconnue $s^l(\theta)$. De plus, elle requiert également que l'ensemble des sauts de $s^l(\theta)$ corresponde à l'ensemble des discontinuités de la vérité terrain y^l, pour laquelle g^l est proche de zéro. Ce composant favorise donc l'alignement des formes entre la carte de probabilités reconstruite $s^l(\theta)$ et la vérité terrain associée y^l. Pour se convaincre de cette propriété et obtenir une caractérisation plus intuitive du rôle de ce terme, considérons le cas où v est la fonction caractéristique, $1_{\Omega_\mathcal{C}}$, d'un ensemble fermé $\Omega_\mathcal{C} \subset \Omega$, Ω étant un sous-ensemble ouvert borné de $\mathbb{R}^n$, avec des bords réguliers $\mathcal{C}$ de classe $\mathcal{C}^2$. Alors, avec g^l la fonction de détection de bords définie ci-dessus, nous pouvons montrer ([7, Remarque 10]) que :

$$TV_{g^l}(v = 1_{\Omega_\mathcal{C}}) = \int_\mathcal{C} g^l \, ds,$$

le terme $\int_\mathcal{C} g^l \, ds$ constituant la nouvelle définition de la longueur de courbe pondérée par le facteur g^l. Nous cherchons donc à localiser la courbe $\mathcal{C}$ là où g^l se rapproche de zéro, c'est-à-dire, sur les frontières de la forme contenue dans l'image de vérité terrain y^l. Par conséquent, lorsqu'il s'agit de segmentation d'images constantes par morceaux, cette propriété constitue un point clé pour faire respecter l'alignement des bords. Encore une fois, à notre connaissance, c'est la première fois qu'un tel terme de régularisation est introduit dans un cadre combinant l'apprentissage profond et les approches variationnelles.

Pénalisation d'aire et autres contraintes Toujours avec l'objectif de rendre la prédiction de segmentation davantage similaire à la vérité terrain, nous proposons différents types de contraintes. Tout d'abord, nous appliquons une pénalisation d'aire en restreignant l'aire de la segmentation obtenue, $\sum_{(i,j)\in\mathcal{G}} (s^l(\theta))_{i,j}$, à être égale à l'aire de la vérité terrain symbolisée par α^l. Cette dernière peut se déterminer directement à partir de l'image de vérité terrain associée par la relation $\alpha^l = \frac{1}{l}\sum_{(i,j)\in\mathcal{G}} y^l_{i,j}$, et donc connue *a priori*. Nous

souhaitons également que $(s^l(\theta))_{i,j}$ devienne binaire. Finalement, afin d'éviter les superpositions entre les différentes classes l, nous imposons qu'en chaque pixel la somme des l probabilités $(s^l(\theta))_{i,j}$ reste toujours égale à 1.

Le problème de minimisation de la fonction de perte qui en résulte se décline comme suit :

$$\begin{aligned} &\inf_{\theta} \mathcal{L}(\theta) = \mathcal{F}(s(\theta), y) + \sum_{l=1}^{L} TV_{g^l}(s^l(\theta)) \\ &\text{s.c} \qquad (s^l(\theta))_{i,j} \in \{0,1\}, \quad \sum_{(i,j)\in\mathcal{G}} (s^l(\theta))_{i,j} = \alpha^l \quad \text{et} \quad \sum_{l=1}^{L} (s^l(\theta))_{i,j} = 1, \end{aligned} \tag{3.4}$$

où $\mathcal{F}$ désigne une fonction coût classique en segmentation d'images telle que la mesure de Dice ou l'entropie croisée, que nous pouvons combiner à un terme de fidélité aux données de Mumford-Shah [70, 98]

$$\sum_{l=1}^{L} \sum_{(i,j)\in\mathcal{G}} \left(y^l_{i,j} - l\right)^2 (s^l(\theta))_{i,j}.$$

La prochaine partie est dédiée aux deux caractérisations de la variation totale pondérée discrète que nous envisageons dans nos schémas d'optimisation.

3.3.3 Caractérisations de la variation totale pondérée discrète

Nous fournissons maintenant deux caractérisations de la variation totale pondérée discrète (dWTV) : une première motivée par le concept de consistance en analyse numérique (la consistance fait référence à une mesure quantitative de la proportion dans laquelle la solution exacte satisfait le problème discret), et une seconde érigée sur la formulation duale. Chacune d'elle sera incorporée dans une fonction de perte, aboutissant à deux modèles différents qui seront comparés et résolus par des techniques d'optimisation adaptées.

Dans cette optique, nous introduisons tout d'abord, dans un cadre continu, la généralisation de la notion de fonctions à variation bornée au cadre des espaces BV associés à une pondération w comme introduit par Baldi *et al.* dans [7, Définition 2]. Nous voyons ensuite comment étendre cette définition au cadre discret afin qu'elle hérite des bonnes propriétés du cadre continu.

Définition 3.3.1. (Extraite de [7, Definition 2])

Soit w un facteur de pondération positif suffisamment lisse (se référer à [7, Definition 1] pour les hypothèses exactes). Nous désignons par $BV(\Omega, w)$ l'ensemble des fonctions $u \in$

$L^1(\Omega, w)$ telles que

$$\sup \left\{ \int_\Omega u \operatorname{div}(\varphi)\, dx \ : \ |\varphi| \leq w \text{ partout}, \ \varphi \in \operatorname{Lip}_0(\Omega, \mathbb{R}^2) \right\} < \infty, \tag{3.5}$$

avec $\operatorname{Lip}_0(\Omega, \mathbb{R}^2)$ l'espace des fonctions continues lipschitziennes à support compact. Nous désignons par $TV_w(u)$ la quantité (3.5).

Remarque 3.3.2. Dans [7], Baldi définit l'espace BV en choisissant comme fonctions tests des éléments de $\operatorname{Lip}_0(\Omega, \mathbb{R}^2)$, tandis que classiquement dans la littérature, les fonctions tests sont prises dans $\mathcal{C}_0^1(\Omega, \mathbb{R}^2)$. En fait, ces définitions coïncident par des arguments de densité (cf. Annexe).

Remarque 3.3.3. Nous rappelons que l'espace de Sobolev pondéré $W^{1,1}(\Omega, w)$ est défini par

$$W^{1,1}(\Omega, w) = \left\{ u \in L^1(\Omega, w) \,|\, \nabla u \in L^1(\Omega, w) \right\},$$

muni de la norme $\|u\|_{W^{1,1}(\Omega,w)} = \|u\|_{L^1(\Omega,w)} + \|\nabla u\|_{L^1(\Omega,w)}$. Lorsque w est suffisamment régulier et u est une fonction dans $W^{1,1}(\Omega, w)$, les deux normes $BV(\Omega, w)$ et $W^{1,1}(\Omega, w)$ sont équivalentes puisque $TV_w(u) = \int_\Omega w\, |\nabla u|\, dx$.
La première version discrète de la variation totale pondérée que nous fournissons repose sur cette observation.

Première caractérisation de la dWTV Une première version de la variation totale pondérée discrète est maintenant conçue à des fins algorithmiques. Dans un souci de simplicité, nous utilisons la même notation pour définir la variation totale pondérée discrète, avec un facteur de pondération positif w.
Une première définition naturelle, inspirée par la Remarque 3.3.3, se décline comme

$$TV_w(u) = \sum_{(i,j)\in\mathcal{G}} w_{i,j}\, |(\nabla u)_{i,j}|, \tag{3.6}$$

pour $u \in X$. Cette première caractérisation de la dWTV sera utilisée dans l'algorithme de Douglas-Rachford introduit dans la sous-section 3.4.2. De plus, la motivation pour cette caractérisation réside dans le concept de consistance en analyse numérique (cf. [96]).

Seconde caractérisation de la dWTV Pour éviter le souci de non-différentiabilité inhérent à la formulation ci-dessus, une alternative consiste à fournir une caractérisation mimant la définition continue (3.5).

Pour ce faire nous adaptons les arguments de Chambolle dans [18].

La fonctionnelle TV_w est positivement homogène de degré un. En effet, $\forall u \in X$, $\forall \lambda > 0$, $TV_w(\lambda\, u) = \lambda\, TV_w(u)$, de sorte que sa transformée de Legendre-Fenchel ([6]), TV_w^*, définie par

$$TV_w^*(v) = \sup_{u \in X} \langle u, v \rangle_X - TV_w(u)$$

est la fonction indicatrice d'un ensemble convexe $\mathcal{K}$:

$$TV_w^*(v) = \begin{cases} 0 & \text{si} \quad v \in \mathcal{K} \\ +\infty & \text{sinon.} \end{cases}$$

Puisque $TV_w^{**} = TV_w$, il vient aisément que

$$TV_w(u) = \sup_{v \in \mathcal{K}} \langle u, v \rangle_X . \tag{3.7}$$

Il reste alors à expliciter l'ensemble $\mathcal{K}$. En revenant à la définition (3.6), clairement, $\forall u \in X$,

$$TV_w(u) = \sup_{p} \langle \nabla u, p \rangle_Y \tag{3.8}$$

avec le supremum pris sur $p \in Y$ tel que $\forall (i,j) \in \mathcal{G}$, $|p_{i,j}| \leq w_{i,j}$. En utilisant l'opérateur de divergence discret introduit précédemment dans (3.3), nous avons $\langle \nabla u, p \rangle_Y = -\langle u, \operatorname{div} p \rangle_X$, menant à

$$\mathcal{K} = \left\{ \operatorname{div} p \,|\, \forall (i,j) \in \mathcal{G},\ |p_{i,j}| \leq w_{i,j} \right\}. \tag{3.9}$$

Cet ensemble apparaît comme une variante discrète de l'ensemble des fonctions tests impliquées dans (3.5). Cette caractérisation de la dWTV est appelée formulation duale de la variation totale pondérée et sera utilisée dans le modèle décrit dans la sous-section 3.4.3.

Pour les lecteurs intéressés par les liens entre les approximations discrètes et leurs homologues continus, nous renvoyons à [22]. Dans cet article, Chambolle *et al.* fournissent une connexion entre l'approximation discrète de la variation totale et sa variante isotrope continue au moyen d'un résultat de Γ-convergence.

La section suivante se consacre à la solution numérique du problème d'optimisation original (3.4).

3.4 Schémas d'optimisation proposés

Notre objectif vise à séparer le problème original (3.4) en problèmes plus facilement résolubles. Pour ce faire, nous introduisons une variable auxiliaire $u = (u^l)$, et le problème

de minimisation (3.4) peut être reformulé de manière équivalente comme suit :

$$\begin{aligned} \inf_{\theta,u=(u^l)} \quad & \mathcal{L}(\theta,u) = \mathcal{F}(s(\theta),y) + \sum_{l=1}^{L} TV_{g^l}(u^l) \\ \text{s.c} \quad & u^l = s^l(\theta),\ u^l_{i,j} \in \{0,1\}\,,\ \sum_{(i,j)\in\mathcal{G}} u^l_{i,j} = \alpha^l \text{ et } \sum_{l=1}^{L} u^l_{i,j} = 1. \end{aligned} \tag{3.10}$$

Nous résolvons ce problème à l'aide d'une méthode de Lagrangien augmenté, plus précisément la forme réduite [13], énoncée comme :

$$\begin{aligned} \max_{w=(w^l)} \min_{\theta,u=(u^l)} \quad & \mathcal{L}(\theta,u,w) = \mathcal{F}(s(\theta),y) + \sum_{l=1}^{L} TV_{g^l}(u^l) + \sum_{l=1}^{L} \frac{\mu}{2} \left\| s^l(\theta) - u^l + w^l \right\|^2 \\ \text{s.c} \quad & u^l_{i,j} \in \{0,1\}\,,\ \sum_{(i,j)\in\mathcal{G}} u^l_{i,j} = \alpha^l \text{ et } \sum_{l=1}^{L} u^l_{i,j} = 1, \end{aligned} \tag{3.11}$$

où $\mu > 0$ est le paramètre de Lagrange augmenté, et $\|f\|$ désigne la quantité $\|f\| = \left(\sum_{(i,j)\in\mathcal{G}} |f_{i,j}|^2\right)^{\frac{1}{2}}$.

L'algorithme ADMM constitue une méthode efficace pour résoudre ce problème d'optimisation, et consiste en la mise à jour alternative de chaque variable alors que les autres sont considérées comme fixes.

3.4.1 Algorithme ADMM pour actualiser les inconnues introduites θ, u et w

Nous détaillons maintenant le procédé de mise à jour des trois paramètres θ, $u = (u^l)$ et $w = (w^l)$ considérés dans notre problème, en suivant le schéma ADMM décrit dans l'Algorithme 2. Pour des raisons de lisibilité, nous explicitons d'abord les actualisations de θ et $w = (w^l)$, puis nous terminons par la mise à jour de $u = (u^l)$, qui concentre la nouveauté de notre travail. Cependant, en pratique et comme indiqué dans l'Algorithme 2, ces mises à jour ne s'opèrent pas dans cet ordre. En effet, initialement, nous choisissons de traiter θ en premier pour obtenir une prédiction grossière $s(\theta)$, et d'actualiser u dans un second temps, via ADMM, puisqu'il est supposé simuler $s(\theta)$. Enfin, et de manière classique, la variable lagrangienne w est mise à jour en dernière.

Tout d'abord, pour actualiser les paramètres du réseau θ, nous utilisons simplement une technique de descente de gradient par lots sur la fonction objectif (la fonctionnelle $\bar{\mathcal{L}}$ à minimiser par rapport à θ étant lisse (non convexe) ici) réduite à :

$$\bar{\mathcal{L}}(\theta) = \mathcal{F}(s(\theta),y) + \frac{\mu}{2} \sum_{l=1}^{L} \left\| s^l(\theta) - u^l + w^l \right\|^2. \tag{3.12}$$

Algorithme 2 Algorithme ADMM pour actualiser θ, u et w

Initialiser θ_0 aléatoirement et $u_0 = w_0 = 0$
Fixer $\mu > 0$
pour $n = 1, 2, \cdots$ **faire**
$\bar{\mathcal{L}}(\theta_n) = \mathcal{F}(s(\theta_n), y) + \frac{\mu}{2} \sum_{l=1}^{L} \|s^l(\theta_n) - u_n^l + w_n^l\|^2$
$\theta_{n+1} = \theta_n - \eta \nabla_\theta \bar{\mathcal{L}}(\theta_n)$
$u_{n+1} = \mathrm{DR}(s(\theta_{n+1}), u_n, w_n)$ ou $\mathrm{PD}(s(\theta_{n+1}), u_n, w_n)$
$w_{n+1} = w_n + \mu \left(s(\theta_{n+1}) - u_{n+1}\right)$
fin pour

Le gradient de cette fonction perte se calcule comme

$$\nabla_\theta \bar{\mathcal{L}}(\theta) = \nabla_\theta \mathcal{F}(s(\theta), y) + \mu \sum_{l=1}^{L} \left(s^l(\theta) - u^l + w^l\right) \nabla_\theta s^l(\theta). \tag{3.13}$$

Les paramètres sont alors actualisés en prenant l'opposé du gradient, *i.e.*, $\theta_{n+1} = \theta_n - \eta \nabla_\theta \bar{\mathcal{L}}(\theta_n)$, où η désigne le taux d'apprentissage et n représente l'ensemble de toutes les étapes successives de descente de gradient réalisées durant une époque. En ce qui concerne la variable lagrangienne $w = (w^l)$, la mise à jour s'effectue par le biais d'une technique de gradient ascendant, *i.e.*, $w_{n+1} = w_n + \mu(s(\theta_{n+1}) - u_{n+1})$. Nous passons maintenant à l'actualisation de $u = (u^l)$.

Deux algorithmes, désignés par les acronymes DR et PD dans l'Algorithme 2, sont analysés et implémentés pour actualiser la variable $u = (u^l)$.

Tous deux s'inscrivent dans le cadre des méthodes proximales ([34]) introduites dans la section 2.4. Leur caractéristique principale a trait au fait de procéder par séparation des variables, en exploitant la séparabilité de la fonction objectif f, dans le cas général, en fonctions $f_1, \cdots, f_m$ employées individuellement, et conduisant à un algorithme facile à mettre en œuvre. À noter qu'il existe un lien étroit entre les deux algorithmes DR et PD comme mis en exergue dans [20, Section 4.] : sous certaines hypothèses, l'algorithme Primal-Dual se ramène à l'approche de séparation des variables de Douglas-Rachford. Il s'agit de méthodes dites proximales, dans la mesure où chaque fonction non différentiable se voit impliquée dans la résolution par le biais de son opérateur de proximité.

Le premier algorithme traité est donc celui de Douglas-Rachford (DR) conçu pour étendre l'algorithme classique forward-backward (Chap. 2, Algorithme 1) aux fonctions non différentiables. En effet, l'algorithme de séparation de Douglas-Rachford offre une certaine

souplesse dans la conception de la fonction coût puisque la différentiabilité n'est maintenant plus requise. Cette caractéristique se révèle très intéressante dans notre cas, du fait de l'absence de solution *closed-form* pour la projection sur l'ensemble des contraintes apparaissant dans notre modélisation. Ce verrou théorique est contourné en dupliquant la variable auxiliaire u en une variable v (donc via un couplage fort entre u et v), chacune d'entre elles supportant une partie des contraintes dures. Nous obtenons ainsi une fonction objectif composée de deux termes non lisses mais pour lesquels les correspondances proximales peuvent être explicitement calculées.

Le second algorithme est celui Primal-Dual (PD) ([20,21]) qui s'appuie sur la reformulation du problème de minimisation considéré comme un problème de point-selle. Une nouvelle fois, l'absence de solution *closed-form* pour la projection sur l'ensemble des contraintes prescrites nous empêche d'appliquer l'algorithme Primal-Dual standard de Chambolle et Pock ([20]). Par conséquent, nous assouplissons le couplage entre u et v (en le transformant en une contrainte relaxée par le biais d'une pénalisation L^2), tout en maintenant les contraintes dures liées à la pénalisation de l'aire et au critère de non-superposition. Cette relaxation conduit finalement un problème qui s'inscrit dans le cadre Primal-Dual hybride de Chambolle et Pock ([21]).

D'une part, intuitivement, nous pouvons nous attendre à ce que la solution DR produise des résultats de segmentation plus précis par rapport à la solution PD, puisque nous conservons l'ensemble des contraintes dures avec celle-ci, tandis que dans la seconde l'une des contraintes se trouve relâchée. D'autre part, nous pouvons anticiper le fait que la première méthode sera plus intensive en termes de calcul puisqu'elle implique la résolution de systèmes linéaires (en comparaison, PD implique le plus souvent des étapes de solutions *closed-form*).

Notre objectif réside alors dans la recherche d'un compromis entre la précision quantitative et l'efficacité algorithmique —même si la charge de calcul principale s'effectue hors ligne —. Cette question de l'équilibre à trouver constitue le premier angle d'investigation pour l'appréciation de nos résultats dans la Section 3.5.

Des éléments d'analyse convexe communs aux deux approches ont été introduits dans la Section 2.4, munis de ces derniers nous pouvons maintenant détailler nos deux solutions.

3.4.2 Algorithme de Douglas-Rachford

Le premier algorithme envisagé impliquant des opérateurs proximaux s'appuie sur la stratégie de *splitting* de Douglas-Rachford. Comme spécifié précédemment, cette technique a d'abord été élaborée pour étendre l'algorithme classique forward-backward aux fonctions non différentiables. Dans notre cas, cette non-différentiabilité apparaît lorsque

nous ajoutons explicitement certaines contraintes dans la fonction objectif en utilisant des fonctions indicatrices, comme décrit par la suite.

Pour actualiser la variable $u = (u^l)$, nous relâchons d'abord légèrement la contrainte binaire et la convertissons en $\forall l \in \{1, \cdots, L\}$, $\forall (i,j) \in \mathcal{G}$, $u^l_{i,j} \in [0,1]$, menant à un problème d'optimisation convexe non lisse en $u = (u^l)$:

$$\begin{aligned} \min_{u=(u^l)} \quad & \sum_{l=1}^{L} \left(TV_{g^l}(u^l) + \frac{\mu}{2} \left\| s^l(\theta) - u^l + w^l \right\|^2 \right) \\ \text{s.c} \quad & u^l_{i,j} \in [0,1], \sum_{(i,j)\in\mathcal{G}} u^l_{i,j} = \alpha^l \text{ et } \sum_{l=1}^{L} u^l_{i,j} = 1. \end{aligned} \tag{3.14}$$

Dans un contexte multi-classes, un obstacle mathématique se situe dans le fait qu'il n'existe pas de solution *closed form* pour la projection sur l'ensemble des contraintes apparaissant dans (3.14). La forme canonique de l'algorithme de Douglas-Rachford permet de dupliquer la variable auxiliaire u en une nouvelle variable v et de séparer les contraintes de manière à ce que u encode la contrainte du simplexe, tandis que v supporte la contrainte d'aire. Nous introduisons donc les variables auxiliaires $z = (z^l)$ et $v = (v^l)$ —la dernière étant relative à la contrainte d'aire —telles que $\forall l \in \{1, \cdots, L\}$, $z^l = \nabla u^l$ et $v^l = u^l$, ainsi que les ensembles convexes

$$\mathcal{C}_1 = \left\{ u = (u^l)_{l\in\{1,\cdots,L\}} \,|\, \forall (i,j) \in \mathcal{G}, \forall l \in \{1, \cdots, L\}\,,\, u^l_{i,j} \in [0,1] \text{ et } \forall (i,j) \in \mathcal{G}, \sum_{l=1}^{L} u^l_{i,j} = 1 \right\},$$

$$\mathcal{C}_2^l = \left\{ (u^l, z^l) \mid \forall (i,j) \in \mathcal{G},\, z^l_{i,j} = (z^{l,1}_{i,j}, z^{l,2}_{i,j})^T = (\nabla u^l)_{i,j} \right\},$$

$$\mathcal{C}_3^l = \left\{ (u^l, v^l) \mid \forall (i,j) \in \mathcal{G},\, u^l_{i,j} = v^l_{i,j} \right\},$$

$$\mathcal{C}_4^l = \left\{ v = (v^l)_{l\in\{1,\cdots,L\}} \,|\, \sum_{(i,j)\in\mathcal{G}} v^l_{i,j} = \alpha^l \right\}.$$

Munis de ces éléments et nous appuyant sur la première caractérisation de la variation totale pondérée discrète (3.6), nous cherchons à résoudre, en utilisant l'algorithme de Douglas-Rachford, le problème équivalent

$$\begin{aligned} \min_{\substack{u=(u^l),\\ z=(z^l), v=(v^l)}} \quad & \sum_{l=1}^{L} \sum_{(i,j)\in\mathcal{G}} g^l_{i,j} \left| z^l_{i,j} \right| + \frac{\mu}{2} \sum_{l=1}^{L} \left\| u^l - s^l(\theta) - w^l \right\|^2 \\ & + i_{\mathcal{C}_1}(u = (u^l)) + \sum_{l=1}^{L} \left(i_{\mathcal{C}_3^l}(u^l, v^l) + i_{\mathcal{C}_2^l}(u^l, z^l) + i_{\mathcal{C}_4^l}(v^l) \right). \end{aligned} \tag{3.15}$$

Algorithme 3 Algorithme de Douglas-Rachford ([34, Algorithme 10.15]) : problème générique $\min_{x\in\mathbb{R}^N} f_1(x) + f_2(x)$ avec f_1 et f_2 fonctions convexes semi-continues inférieures de $\mathbb{R}^N$ dans $]-\infty, +\infty]$.

Fixer $\varepsilon \in]0,1[$, $\gamma > 0$, $y_0 \in \mathbb{R}^N$
pour $n = 0, 1, \cdots$ **faire**
  $x_n = \text{prox}_{\gamma f_2}\, y_n$
  $\lambda_n \in [\varepsilon, 2-\varepsilon]$
  $y_{n+1} = y_n + \lambda_n \left(\text{prox}_{\gamma f_1}\, (2x_n - y_n) - x_n\right)$.
fin pour

L'algorithme de Douglas-Rachford prend la forme d'un schéma itératif, présenté dans l'Algorithme 3 dans sa forme générale, pour minimiser la somme de fonctions convexes. Les solutions sont obtenues grâce aux opérateurs proximaux qui peuvent être appréhendés comme une extension naturelle de la notion d'opérateur de projection sur un ensemble convexe ([34]). En outre, le résultat de convergence suivant est valable, rappelé dans un souci d'exhaustivité.

Théorème 3.4.1. *(Extrait de [34, Proposition 10.16])*

Toute suite $(x_n)_{n\in\mathbb{N}}$ générée par l'Algorithme 3 converge vers une solution au problème générique $\min_{x\in\mathbb{R}^N} f_1(x) + f_2(x)$.

Conformément aux notations ci-dessus, nous définissons la fonction $f_1(u, v, z) = \phi(u, v) + h(z)$ telle que

$$\phi(u,v) = i_{\mathcal{C}_1}(u) + \sum_{l=1}^{L} \left(\frac{\mu}{2}\left\|u^l - s^l(\theta) - w^l\right\|^2 + i_{\mathcal{C}_3^l}(u^l, v^l)\right),$$

$$h(z) = \sum_{l=1}^{L} \sum_{(i,j)\in\mathcal{G}} g_{i,j}^l \left|z_{i,j}^l\right|.$$

L'indépendance des variables de f_1 permet de conclure que

$$\text{prox}_{\gamma f_1}(u,v,z) = \begin{pmatrix} \text{prox}_{\gamma\,\phi}(u,v) \\ \text{prox}_{\gamma\, h}(z) \end{pmatrix}.$$

Concentrons-nous d'abord sur $\text{prox}_{\gamma h}(z) = \left(\text{prox}_{\gamma g_{i,j}^l |\cdot|}(z_{i,j}^l)\right)_{l\in\{1,\cdots,L\},(i,j)\in\mathcal{G}}$, en raison encore de l'indépendance des variables.
En supposant que $\forall l \in \{1, \cdots, L\}$, $\forall (i,j) \in \mathcal{G}$, $g_{i,j}^l \geq \varepsilon$, avec ε un paramètre positif assez petit, nous obtenons

$$\text{prox}_{\gamma g_{i,j}^l |\cdot|}(z_{i,j}^l) = \begin{cases} \left(1 - \dfrac{\gamma\, g_{i,j}^l}{|z_{i,j}^l|}\right) z_{i,j}^l & \text{si} \quad |z_{i,j}^l| \geq \gamma\, g_{i,j}^l \\ 0 & \text{sinon} \end{cases}.$$

Par définition,

$$\begin{aligned}\text{prox}_{\gamma\phi}(u,v) = \underset{\substack{\tilde{u}=(\tilde{u}^l)\\ \tilde{v}=(\tilde{v}^l)}}{\arg\min}\ & \frac{\mu\gamma}{2}\sum_{l=1}^{L}\|\tilde{u}^l - s^l(\theta) - w^l\|^2 + i_{\mathcal{C}_1}(\tilde{u}=(\tilde{u}^l)) \\ & + \sum_{l=1}^{L}\left(i_{\mathcal{C}_3^l}(\tilde{u}^l,\tilde{v}^l) + \frac{1}{2}\|\tilde{u}^l - u^l\|^2 + \frac{1}{2}\|\tilde{v}^l - v^l\|^2\right).\end{aligned}$$

Nous posons $\tilde{v}^l = \tilde{u}^l$ et résolvons

$$\begin{aligned}&\underset{\tilde{u}\in\mathcal{C}_1}{\arg\min}\ \frac{\mu\gamma}{2}\sum_{l=1}^{L}\|\tilde{u}^l - s^l(\theta) - w^l\|^2 + \frac{1}{2}\sum_{l=1}^{L}\|\tilde{u}^l - u^l\|^2 + \frac{1}{2}\sum_{l=1}^{L}\|\tilde{u}^l - v^l\|^2, \\ =&\underset{\tilde{u}\in\mathcal{C}_1}{\arg\min}\ \frac{\mu\gamma+2}{2}\sum_{l=1}^{L}\|\tilde{u}^l - \frac{\mu\gamma}{\mu\gamma+2}\left(s^l(\theta)+w^l\right) - \frac{1}{\mu\gamma+2}u^l - \frac{1}{\mu\gamma+2}v^l\|^2, \\ =&\underset{\tilde{u}\in\mathcal{C}_1}{\arg\min}\ \frac{\mu\gamma+2}{2}\sum_{l=1}^{L}\|\tilde{u}^l - x^l\|^2,\end{aligned}$$

avec

$$x^l = \frac{\mu\gamma}{\mu\gamma+2}\left(s^l(\theta)+w^l\right) + \frac{1}{\mu\gamma+2}u^l + \frac{1}{\mu\gamma+2}v^l.$$

Ce dernier problème, séparable par rapport à $(i,j)\in\mathcal{G}$, peut être décomposé en N^2 problèmes plus petits dans $\mathbb{R}^L$ et la solution est donnée par

$$\left(\text{prox}_{\gamma\phi}(u,v)\right)_{(i,j)\in\mathcal{G},\, l\in\{1,\cdots,L\}} = \begin{pmatrix} P_{[0,1]}(x_{i,j}^l - \lambda) \\ P_{[0,1]}(x_{i,j}^l - \lambda)\end{pmatrix},$$

où $\lambda(=\lambda_{i,j})$ est la solution de l'équation

$$\sum_{l=1}^{L} P_{[0,1]}(x_{i,j}^l - \lambda) = 1.$$

Similairement, nous définissons $f_2(u,z,v) = \sum_{l=1}^{L}\left(i_{\mathcal{C}_2^l}(u^l,z^l) + i_{\mathcal{C}_4^l}(v^l)\right)$, et en vertu des propriétés de séparabilité, nous avons

$$\text{prox}_{\gamma f_2}(u,z,v) = \begin{pmatrix}\left(P_{\mathcal{C}_2^l}(u^l,z^l)\right)_{l\in\{1,\cdots,L\}} \\ \left(P_{\mathcal{C}_4^l}(v^l)\right)_{l\in\{1,\cdots,L\}}\end{pmatrix}.$$

Par définition, pour un $l\in\{1,\cdots,L\}$ donné,

$$\begin{aligned}P_{\mathcal{C}_2^l}(u^l,z^l) &= \underset{(\tilde{u}^l,\tilde{z}^l)\in\mathcal{C}_2^l}{\arg\min}\ \frac{1}{2}\|\tilde{u}^l - u^l\|^2 + \frac{1}{2}\|\tilde{z}^l - z^l\|^2, \\ &= \underset{\tilde{u}^l}{\arg\min}\ \frac{1}{2}\|\tilde{u}^l - u^l\|^2 + \frac{1}{2}\|\nabla\tilde{u}^l - z^l\|^2, \\ &= \underset{\tilde{u}^l}{\arg\min}\ \frac{1}{2}\|\tilde{u}^l - u^l\|^2 + \frac{1}{2}\|\nabla_x\tilde{u}^l - z^{l,1}\|^2 + \frac{1}{2}\|\nabla_y\tilde{u}^l - z^{l,2}\|^2,\end{aligned}$$

en utilisant le fait que $\nabla\tilde{u}^l = \tilde{z}^l$ et en prenant $z^l = (z^{l,1}, z^{l,2})^T$. À noter que nous utilisons la même notation pour désigner à la fois la norme euclidienne dans $X = \mathbb{R}^{N\times N}$ et la norme euclidienne dans $\mathbb{R}^{2\times N\times N}$. La condition d'optimalité du premier ordre (condition nécessaire et suffisante) mène à

$$\tilde{u}^l = (I + \nabla_x^T\nabla_x + \nabla_y^T\nabla_y)^{-1}(u^l + \nabla_x^T z^{l,1} + \nabla_y^T z^{l,2}).$$

Alors, nous obtenons $P_{\mathcal{C}_2^l}(u^l, z^l) = (\tilde{u}^l, \nabla\tilde{u}^l)^T$. Enfin,

$$P_{C_4^l}(v^l) = v^l + \frac{\alpha^l - \sum\limits_{(i,j)\in\mathcal{G}} v_{i,j}^l}{N^2}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}.$$

Bien qu'il présente de bonnes propriétés de séparabilité et produise des résultats satisfaisants, cet algorithme nécessite la résolution d'un large système linéaire, ce qui est réalisé en interfaçant GPU et CPU. Plus précisément, le calcul de la solution donnée avec l'algorithme DR est implémenté sur CPU pour profiter des librairies adéquates d'algèbre linéaire. L'indépendance entre chaque image de l'ensemble de données permet de procéder à la parallélisation du code de calcul via MPI. L'apprentissage des paramètres du réseau profond s'effectue sur GPU. Afin de surmonter cet inconvénient, nous introduisons un algorithme alternatif fondé sur la formulation duale de la variation totale pondérée, qui conduit à un problème de type min-max.

3.4.3 Formulation duale de la variation totale pondérée

Dans le second algorithme envisagé, nous introduisons toujours une variable $v = (v^l)$ relative à la contrainte d'aire, telle que v^l reste proche de u^l. L'idée sous-jacente étant toujours de décomposer le problème original en problèmes plus facilement résolubles. Cette fois, un couplage plus souple entre u^l et v^l, par le biais d'une pénalisation L^2, est un substitut à la contrainte dure liant v^l et u^l dans l'algorithme de Douglas-Rachford. Cette version de (3.14) s'énonce donc comme suit

$$\min_{\substack{u=(u^l)\in\mathcal{C}_1\\ v=(v^l)\in\mathcal{C}_2}} \mathcal{J}(u,v) = \sum_{l=1}^{L}\left(TV_{g^l}(u^l) + \frac{\mu}{2}\left\|s^l(\theta) - u^l + w^l\right\|^2\right) + \frac{1}{2}\sum_{l=1}^{L}\|u^l - v^l\|^2, \qquad (\mathcal{P})$$

avec maintenant $\mathcal{C}_1$ et $\mathcal{C}_2$ les ensembles convexes définis par

$$\mathcal{C}_1 = \left\{ u = (u^l)_{l\in\{1,\cdots,L\}} \;\middle|\; \begin{array}{c} \forall (i,j)\in\mathcal{G}, \forall l\in\{1,\cdots,L\}, \\ u_{i,j}^l \in [0,1] \text{ et } \sum\limits_{l=1}^{L} u_{i,j}^l = 1 \end{array} \right\},$$

$$\mathcal{C}_2 = \left\{ v = (v^l)_{l\in\{1,\cdots,L\}} \,|\, \forall (i,j)\in\mathcal{G}, \forall l\in\{1,\cdots,L\}, \sum_{(i,j)\in\mathcal{G}} v_{i,j}^l = \alpha^l \right\}.$$

Un premier résultat théorique assure l'existence et l'unicité du minimiseur du problème $(\mathcal{P})$. Ce résultat fait sens lorsque nous introduisons le problème associé min-max.

Théorème 3.4.2. *Le problème $(\mathcal{P})$ admet une solution unique.*

Démonstration. Les ensembles $\mathcal{C}_1$ et $\mathcal{C}_2$ sont fermés et convexes, tandis que la fonction objectif est continue et coercive. En effet, $\forall l \in \{1, \cdots, L\}$,

$$\begin{cases} \frac{\mu}{2}\|u^l - s^l(\theta) - w^l\|^2 & \geq \frac{\mu}{4}\|u^l\|^2 - \frac{\mu}{2}\|s^l(\theta) + w^l\|^2, \\ \frac{1}{2}\|u^l - v^l\|^2 & \geq \frac{1}{2}\|u^l\|^2 + \frac{1}{2}\|v^l\|^2 - \|u^l\|\,\|v^l\|. \end{cases}$$

En utilisant l'inégalité de Young avec ϵ (valide pour $\epsilon > 0$) et formulée comme $ab \leq \frac{a^2}{2\epsilon} + \frac{\epsilon b^2}{2}$, nous obtenons :

$$\mathcal{J}(u,v) \geq \left(\frac{\mu+2}{4} - \frac{1}{2\epsilon}\right)\sum_{l=1}^{L}\|u^l\|^2 - \frac{\mu}{2}\sum_{l=1}^{L}\|s^l(\theta) + w^l\|^2 + \left(\frac{1}{2} - \frac{\epsilon}{2}\right)\sum_{l=1}^{L}\|v^l\|^2,$$

ou de manière équivalente,

$$\mathcal{J}(u,v) \geq \left(\frac{\mu+2}{4} - \frac{1}{2\epsilon}\right)\|u\|^2 - \frac{\mu}{2}\|s(\theta) + w\|^2 + \left(\frac{1}{2} - \frac{\epsilon}{2}\right)\|v\|^2.$$

Pour alléger les notations, quand il n'y a pas d'ambiguïté sur la dimension des objets mathématiques que nous manipulons, nous omettons l'indice rendant cette dimension explicite dans la définition de la norme euclidienne, ainsi que pour le produit scalaire associé.
En choisissant un ϵ tel que $\frac{2}{\mu+2} < \epsilon < 1$, cela nous permet d'établir à la fois que $\frac{\mu+2}{4} - \frac{1}{2\epsilon} > 0$ et $\frac{1}{2} - \frac{\epsilon}{2} > 0$. Nous obtenons alors le résultat désiré, et par conséquent la fonctionnelle $\mathcal{J}$ admet un minimum.

Pour conclure, la fonctionnelle $\mathcal{J}$ est strictement convexe, du fait de la convexité stricte de la fonctionnelle $\mathcal{H}$ définie par $\mathcal{H}(u,v) = \frac{\mu}{2}\|u - s(\theta) - w\|^2 + \frac{1}{2}\|u - v\|^2$, u désignant la concaténation des éléments u^l, et similairement pour $s(\theta)$, w et v. Un calcul simple permet d'obtenir que $\forall (u_1, v_1) \in \mathcal{C}_1 \times \mathcal{C}_2$, $\forall (u_2, v_2) \in \mathcal{C}_1 \times \mathcal{C}_2$,

$$\langle \nabla\mathcal{H}(u_1,v_1) - \nabla\mathcal{H}(u_2,v_2), \begin{pmatrix} u_1 - u_2 \\ v_1 - v_2 \end{pmatrix}\rangle = \mu\|u_1 - u_2\|^2 + \|(u_1 - u_2) - (v_1 - v_2)\|^2,$$

cette quantité s'annule si et seulement si $u_1 = u_2$ et $v_1 = v_2$, justifiant l'unicité du minimum. □

Plutôt que d'introduire la variable auxiliaire z^l dans le problème $(\mathcal{P})$, nous considérons maintenant la variation totale pondérée TV_g comme la quantité définie par (3.8). Par conséquent, nous réécrivons ce terme comme :

$$TV_{g^l}(u^l) = \max_{p^l} \langle \nabla u^l, p^l \rangle_Y - i_{\mathcal{B}_l}(p^l), \tag{3.16}$$

où l'ensemble convexe $\mathcal{B}_l$ est donné par

$$\mathcal{B}_l = \left\{ p^l \in Y \,|\, \forall (i,j) \in \mathcal{G}, |p^l_{i,j}| \leq g^l_{i,j} \right\}$$

(désigné par $\mathcal{K}$ dans l'exposé du cadre théorique, sous-section 3.3.3). Une fois encore, pour alléger la notation, dans la suite nous utiliserons

$$\sum_{l=1}^{L} TV_{g^l}(u^l) = \max_{p=(p^l)\in Y^L} \langle \nabla u, p \rangle - i_{\mathcal{B}}(p),$$

avec $\mathcal{B}$ l'ensemble convexe donné par

$$\mathcal{B} = \left\{ p = (p^l) \in Y^L \,|\, \forall l \in \{1, \cdots, L\}, \forall (i,j) \in \mathcal{G}, \, |p^l_{i,j}| \leq g^l_{i,j} \right\}.$$

Ces éléments nous permettent de reformuler notre problème d'optimisation $(\mathcal{P})$ comme un problème de type min-max

$$\min_{\substack{u=(u^l)\in\mathcal{C}_1 \\ v=(v^l)\in\mathcal{C}_2}} \max_{p=(p^l)\in\mathcal{B}} \mathcal{L}(u,v,p) := \langle \nabla u, p \rangle + \frac{\mu}{2} \|u - s(\theta) - w\|^2 + \frac{1}{2} \|u - v\|^2. \tag{$\bar{\mathcal{P}}$}$$

Le résultat qui suit assure que le Lagrangien $\mathcal{L}$ possède un point-selle. Avant d'énoncer ce résultat théorique, nous rappelons quelques notions de base à propos de la dualité par le théorème minimax, principalement tirées de [43, Chapitre VI]. En particulier, nous établissons la connexion explicite entre un point-selle de $\mathcal{L}$ et un minimiseur de $(\mathcal{P})$, motivant ainsi le modèle introduit.

Nous utilisons les mêmes notations qu'Ekeland et Temam et considérons le problème général de minimisation formulé par

$$\inf_{u\in V} \Phi(u) \tag{3.17}$$

et appelé *problème primal*. Φ est ici une fonctionnelle générique. Nous supposons que $\Phi(u)$ peut être réécrite comme un supremum en p d'une fonction $L(u,p)$ pour avoir

$$\inf_{u\in V} \sup_{p\in Z} L(u,p). \tag{3.18}$$

Un problème relatif appelé *dual de (3.17)* se définit comme

$$\sup_{p\in Z} \inf_{u\in V} L(u,p). \tag{3.19}$$

Nous rappelons d'abord la définition d'un point-selle.

Définition 3.4.3. (Extraite de [43, Chapitre VI, Définition 1.1])

On dit qu'une paire $(\bar{u}, \bar{p}) \in \mathcal{A} \times \mathcal{B}$ est un point-selle de L sur $\mathcal{A} \times \mathcal{B}$ si, $\forall u \in \mathcal{A}$, $\forall p \in \mathcal{B}$,

$$L(\bar{u}, p) \leq L(\bar{u}, \bar{p}) \leq L(u, \bar{p}).$$

À présent, afin de légitimer l'introduction de notre algorithme min-max, il reste à établir la connexion entre un point-selle de L et un minimiseur du problème primal (3.17). Pour ce faire, nous supposons que $(\bar{u}, \bar{p})$ est un point-selle de L. À partir du côté gauche de l'inégalité dans la Définition 3.4.3., nous avons :

$$\inf_{u \in \mathcal{A}} \Phi(u) \leq \Phi(\bar{u}) = \sup_{p \in \mathcal{B}} L(\bar{u}, p) \leq L(\bar{u}, \bar{p}).$$

Maintenant, à partir du côté droit de l'inégalité dans la Définition 3.4.3., il vient

$$L(\bar{u}, \bar{p}) \leq \inf_{u \in \mathcal{A}} L(u, \bar{p}).$$

Mais, $\forall u \in \mathcal{A}$, $L(u, \bar{p}) \leq \sup_{p \in \mathcal{B}} L(u, p)$, ce qui donne $\inf_{u \in \mathcal{A}} L(u, \bar{p}) \leq \inf_{u \in \mathcal{A}} \sup_{p \in \mathcal{B}} L(u, p) = \inf_{u \in \mathcal{A}} \Phi(u)$. Par conséquent, si $(\bar{u}, \bar{p})$ est un point-selle de L, alors $\bar{u}$ est un minimiseur de Φ.

À la lumière de cette liaison entre le problème min-max et le problème original (primal) $(\mathcal{P})$, nous sommes maintenant prêts à énoncer le résultat de l'existence de points-selle. (À noter que cette question de l'existence de points-selle est rarement abordée.) Ainsi, par souci d'exhaustivité et pour assurer le caractère bien posé de notre problème, il nous semble pertinent de la vérifier.

Théorème 3.4.4. *Le Lagrangien $\mathcal{L}$ du problème $(\bar{\mathcal{P}})$ possède au moins un point-selle $(\bar{u}, \bar{v}, \bar{p}) \in \mathcal{C}_1 \times \mathcal{C}_2 \times \mathcal{B}$.*

Démonstration. La preuve est une adaptation de celle de [43, Chapitre VI, proposition 2.1] au cas non borné.

Tout d'abord, nous rappelons le Lagrangien $\mathcal{L}$ sur lequel nous travaillons

$$\mathcal{L}(u, v, p) := \langle \nabla u, p \rangle + \frac{\mu}{2} \|u - s(\theta) - w\|^2 + \frac{1}{2} \|u - v\|^2.$$

Pour tout $p \in \mathcal{B}$, la fonctionnelle $(u, v) \mapsto \mathcal{L}(u, v, p)$ est strictement convexe, du fait de la stricte convexité de la fonctionnelle $\mathcal{H} = \frac{\mu}{2} \|u - s(\theta) - w\|^2 + \frac{1}{2} \|u - v\|^2$. De plus, pour tout $p \in \mathcal{B}$, la fonctionnelle $(u, v) \mapsto \mathcal{L}(u, v, p)$ est continue et coercive. Pour établir une

telle inégalité de coercivité, en notant $\kappa = \|\mathrm{div}\| = \sup_{\|p\|_{Y^L} \leq 1} \|\mathrm{div}\, p\|_{X^L}$, nous remarquons d'abord qu'avec la convention $(p^1)^1_{0,j} = (p^1)^1_{N,j} = (p^1)^2_{i,0} = (p^1)^2_{i,N} = 0$ (et identiquement pour p^l avec $l \in \{2, \cdots, L\}$) et en appliquant deux fois l'inégalité $(a+b)^2 \leq 2(a^2+b^2)$,

$$\|\mathrm{div}\, p\|^2_{X^L} \leq 8\, \|p\|^2.$$

Ainsi $\kappa \leq 2\sqrt{2}$ et avec un ϵ adapté,

$$\begin{aligned}\mathcal{L}(u,v,p) &\geq \left(\frac{\mu+2}{4} - \frac{1}{2\epsilon}\right) \|u\|^2 - \frac{\mu}{2}\|s(\theta)+w\|^2 + \left(\frac{1}{2} - \frac{\epsilon}{2}\right)\|v\|^2 - \|\mathrm{div}\|\|p\|\|u\|, \\ &\geq \left(\frac{\mu+2}{4} - \frac{1}{2\epsilon}\right) \|u\|^2 - \frac{\mu}{2}\|s(\theta)+w\|^2 + \left(\frac{1}{2} - \frac{\epsilon}{2}\right)\|v\|^2 - 2\sqrt{2L}N\|u\|,\end{aligned}$$

puisque $p \in \mathcal{B}$, entraînant que $\|p\|^2_{Y^L} \leq LN^2$.
En appliquant une nouvelle fois l'inégalité de Young avec $\epsilon' > 0$,

$$\mathcal{L}(u,v,p) \geq \left(\frac{\mu+2}{4} - \frac{1}{2\epsilon} - \frac{\epsilon'}{2}\right) \|u\|^2 + \left(\frac{1}{2} - \frac{\epsilon}{2}\right)\|v\|^2 - \frac{\mu}{2}\|s(\theta)+w\|^2 - \frac{4LN^2}{\epsilon'}.$$

Cette dernière inégalité montre qu'en choisissant judicieusement ϵ' —ce qui est toujours possible —, la propriété de coercivité est garantie.
De plus, la quantité $\mathcal{L}(u,v,p)$ est bornée inférieurement (indépendamment de p) et nous remarquons qu'en prenant la variable $\tilde{u}$ telle que $\tilde{u}^1 \equiv 1$ et $\forall l \in \{2, \cdots, L\}$, $\tilde{u}^l \equiv 0$, et $\tilde{v}$ telle que $\forall l \in \{1, \cdots, L\}$, $\forall (i,j) \in \mathcal{G}$, $\tilde{v}^l_{i,j} = \dfrac{\alpha^l}{N^2}$, alors la fonctionnelle $\mathcal{L}(\tilde{u}, \tilde{v}, p)$ est indépendante de p, ce qui montre que l'infimum est fini.
Donc pour tout $p \in \mathcal{B}$, la fonctionnelle $\mathcal{L}(\cdot, \cdot, p)$ est continue, coercive et strictement convexe, de sorte qu'elle admet un unique minimiseur dans $\mathcal{C}_1 \times \mathcal{C}_2$ —$\mathcal{C}_1$ et $\mathcal{C}_2$ étant des ensembles convexes fermés —désigné par $(e_1(p), e_2(p))$.
Nous notons ce minimum par $f(p)$, c'est-à-dire,

$$f(p) = \min_{(u,v) \in \mathcal{C}_1 \times \mathcal{C}_2} \mathcal{L}(u,v,p) = \mathcal{L}(e_1(p), e_2(p), p).$$

La fonction $p \mapsto f(p)$ est concave en tant qu'infimum ponctuel de fonctions concaves ($\forall (u,v) \in \mathcal{C}_1 \times \mathcal{C}_2$, $p \mapsto \mathcal{L}(u,v,p)$ est concave puisqu'en fait linéaire). Cela peut se prouver en utilisant l'hypographe de f. Aussi, f est semi-continue supérieurement en tant qu'infimum ponctuel de fonctions continues. Elle est alors bornée supérieurement et atteint sa borne supérieure car l'ensemble $\mathcal{B}$ est compact en un point désigné par $\bar{p}$. Donc

$$f(\bar{p}) = \max_{p \in \mathcal{B}} f(p) = \max_{p \in \mathcal{B}} \min_{(u,v) \in \mathcal{C}_1 \times \mathcal{C}_2} \mathcal{L}(u,v,p).$$

De plus, comme $f(\bar{p}) = \min_{(u,v) \in \mathcal{C}_1 \times \mathcal{C}_2} \mathcal{L}(u,v,\bar{p})$, nous avons, $\forall (u,v) \in \mathcal{C}_1 \times \mathcal{C}_2$,

$$f(\bar{p}) \leq \mathcal{L}(u,v,\bar{p}). \tag{3.20}$$

Par concavité de $\mathcal{L}$ par rapport au troisième argument (en fait, linéarité), $\forall (u,v) \in \mathcal{C}_1 \times \mathcal{C}_2$, $\forall p \in \mathcal{B}$, $\forall \lambda \in]0,1[$,

$$\mathcal{L}(u,v,(1-\lambda)\,\bar{p}+\lambda p) = (1-\lambda)\,\mathcal{L}(u,v,\bar{p}) + \lambda\,\mathcal{L}(u,v,p).$$

En prenant comme valeur particulière $(u,v) = (e_1((1-\lambda)\,\bar{p}+\lambda p), e_2((1-\lambda)\,\bar{p}+\lambda p)) = (e_\lambda^1, e_\lambda^2)$, il vient, en utilisant à nouveau le fait que $f(\bar{p}) = \max\limits_{p\in\mathcal{B}} f(p)$ et la concavité (même la linéarité) de $\mathcal{L}$ par rapport au troisième argument,

$$\begin{aligned} f(\bar{p}) &\geq f((1-\lambda)\,\bar{p}+\lambda p) = \mathcal{L}(e_\lambda^1, e_\lambda^2, (1-\lambda)\,\bar{p}+\lambda p), \\ &\geq (1-\lambda)\,\mathcal{L}(e_\lambda^1, e_\lambda^2, \bar{p}) + \lambda\,\mathcal{L}(e_\lambda^1, e_\lambda^2, p). \end{aligned}$$

Comme $\mathcal{L}(e_\lambda^1, e_\lambda^2, \bar{p}) \geq f(\bar{p}) = \min\limits_{(u,v)\in\mathcal{C}_1\times\mathcal{C}_2} \mathcal{L}(u,v,\bar{p})$, la dernière inégalité implique que $\forall p \in \mathcal{B}$,

$$f(\bar{p}) \geq \mathcal{L}(e_\lambda^1, e_\lambda^2, p). \tag{3.21}$$

En vertu de la propriété de coercivité établie précédemment, nous avons, $\forall p \in \mathcal{B}$ (les paramètres ϵ et ϵ' étant convenablement choisis),

$$\begin{aligned} \left(\frac{\mu+2}{4} - \frac{1}{2\epsilon} - \frac{\epsilon'}{2}\right) \|e_\lambda^1\|^2 + \left(\frac{1}{2} - \frac{\epsilon}{2}\right) \|e_\lambda^2\|^2 - \frac{\mu}{2}\,\|s(\theta)+w\|^2 - \frac{4LN^2}{\epsilon'} \\ \leq \mathcal{L}(e_\lambda^1, e_\lambda^2, (1-\lambda)\,\bar{p}+\lambda p) \leq \mathcal{L}(\tilde{u},\tilde{v},(1-\lambda)\,\bar{p}+\lambda p), \end{aligned}$$

avec les variables $\tilde{u} \in \mathcal{C}_1$ et $\tilde{v} \in \mathcal{C}_2$ telles que définies précédemment, rendant la partie droite indépendante de p, $\bar{p}$ et λ —constituant ainsi une borne uniforme —, et montrant que e_λ^1 est uniformément bornée (cela était déjà connu en raison de la définition de $\mathcal{C}_1$) tout comme e_λ^2. Il est alors possible d'extraire des sous-suites $e_{\lambda_n}^1$ et $e_{\lambda_n}^2$ qui convergent vers des limites $\bar{u}$ et $\bar{v}$ quand $\lambda_n \underset{n\to+\infty}{\to} 0$. Nous montrons ensuite que $\bar{u} = e_1(\bar{p})$ et $\bar{v} = e_2(\bar{p})$. Comme par définition, $\mathcal{L}(e_\lambda^1, e_\lambda^2, (1-\lambda)\,\bar{p}+\lambda p) = \min\limits_{(u,v)\in\mathcal{C}_1\times\mathcal{C}_2} \mathcal{L}(u,v,(1-\lambda)\,\bar{p}+\lambda p)$, $\forall (u,v) \in \mathcal{C}_1 \times \mathcal{C}_2$,

$$\mathcal{L}(e_\lambda^1, e_\lambda^2, (1-\lambda)\,\bar{p}+\lambda p) \leq \mathcal{L}(u,v,(1-\lambda)\,\bar{p}+\lambda p),$$

et par linéarité de $\mathcal{L}$ par rapport au troisième argument, il s'ensuit que $\forall (u,v) \in \mathcal{C}_1 \times \mathcal{C}_2$,

$$(1-\lambda)\,\mathcal{L}(e_\lambda^1, e_\lambda^2, \bar{p}) + \lambda\,\mathcal{L}(e_\lambda^1, e_\lambda^2, p) \leq \mathcal{L}(u,v,(1-\lambda)\,\bar{p}+\lambda p).$$

La quantité $\mathcal{L}(e_\lambda^1, e_\lambda^2, p)$ est bornée inférieurement par $f(p)$, de sorte que le passage à la limite dans l'inégalité précédente, lorsque λ_n tend vers 0, conduit à, en utilisant la continuité de $\mathcal{L}$,

$$\mathcal{L}(\bar{u},\bar{v},\bar{p}) \leq \mathcal{L}(u,v,\bar{p}),$$

cela étant vrai pour tout $(u, v) \in \mathcal{C}_1 \times \mathcal{C}_2$.
Par unicité du minimiseur de $\min_{(u,v)\in\mathcal{C}_1\times\mathcal{C}_2} \mathcal{L}(u, v, \bar{p})$, nous déduisons que $(\bar{u}, \bar{v}) = (e_1(\bar{p}), e_2(\bar{p}))$.
Pour finir, en passant à la limite dans (3.21), il vient $\mathcal{L}(\bar{u}, \bar{v}, p) \leq f(\bar{p}), \forall p \in \mathcal{B}$, qui combiné avec (3.20) et l'invocation de [43, Chapitre VI, Proposition 1.3] nous permet de conclure que $(\bar{u}, \bar{v}, \bar{p})$ est un point-selle de $\mathcal{L}$. □

Notre problème $(\bar{\mathcal{P}})$ s'inscrit dans le cadre général des problèmes d'optimisation convexe avec une structure de point-selle connue, abordés dans [21] et exprimés comme

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \mathcal{L}(x, y) = \langle Kx, y\rangle + f(x) + g(x) - h^*(y), \tag{CP}$$

(i) $\mathcal{X}$ et $\mathcal{Y}$ étant, dans le cadre le plus général, des espaces de Banach réflexifs réels avec les normes correspondantes $\|\cdot\|_{\mathcal{X}}$ et $\|\cdot\|_{\mathcal{Y}}$,

(ii) $K : \mathcal{X} \to \mathcal{Y}^*$ (espace dual de $\mathcal{Y}$) étant un opérateur linéaire borné dont l'opérateur adjoint correspondant $K^* : \mathcal{Y} \to \mathcal{X}^*$ est défini par $\langle Kx, y\rangle = \langle K^*y, x\rangle$, $\forall(x, y) \in \mathcal{X} \times \mathcal{Y}$,

(iii) f est une fonction convexe propre semi-continue inférieurement (s.c.i) avec ∇f continu lipschitzien sur $\mathcal{X}$, la constante de Lipschitz étant notée L_f,

(iv) g et h^* sont des fonctions convexes, propres, s.c.i, avec une structure simple dans le sens où leur opérateur proximal peut se calculer explicitement.

Alors l'algorithme général donné dans [21] se décline comme suit.

Algorithme 4 Itération générale de l'algorithme de Chambolle-Pock [21, Algorithme 1] - Fonction de proximité euclidienne

Entrée : Norme de l'opérateur $L := \|K\|$, constante de Lipschitz L_f de ∇f
Initialisation : Choisir $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$, τ, $\sigma > 0$
Itérations : Pour chaque $n \geq 0$, soit

$$(x^{n+1}, y^{n+1}) = PD_{\tau,\sigma}(x^n, y^n, 2x^{n+1} - x^n, y^n)$$

avec l'itération $(\hat{x}, \hat{y}) = PD_{\tau,\sigma}(\bar{x}, \bar{y}, \tilde{x}, \tilde{y})$ telle que

$$\begin{cases} \hat{x} &= \arg\min_x f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + g(x) + \langle Kx, \tilde{y}\rangle \\ & \qquad + \dfrac{1}{2\tau}\|x - \bar{x}\|^2, \\ \hat{y} &= \arg\min_y h^*(y) - \langle K\tilde{x}, y\rangle + \dfrac{1}{2\sigma}\|y - \bar{y}\|^2. \end{cases}$$

Remarque 3.4.5. Dans notre cas, les espaces $\mathcal{X}$ et $\mathcal{Y}$ sont supposés de dimension finie (espaces euclidiens classiques).

Remarque 3.4.6. Il peut être observé que les points fixes de l'Algorithme 4 sont des points-selle du Lagrangien associé. Dans la suite, nous développons cette observation (qui n'est pas explicitement établie dans [21]).

Commentaire sur le fait que les points fixes de l'Algorithme 4 sont des points–selle du Lagrangien associé :

Tout d'abord, nous remarquons que

$$\begin{aligned}\hat{x} &= \arg\min_x\ g(x) + \langle \nabla f(\bar{x}) + K^*\tilde{y}, x\rangle + \frac{1}{2\tau}\,\|x-\bar{x}\|^2,\\ &= \arg\min_x\ g(x) + \frac{1}{2\tau}\,\|x - (\bar{x} - \tau\ (\nabla f(\bar{x}) + K^*\tilde{y}))\,\|^2,\\ &= \mathrm{prox}_{\tau\, g}(\bar{x} - \tau\,(\nabla f(\bar{x}) + K^*\tilde{y})).\end{aligned}$$

Ainsi, la première étape de l'algorithme peut être reformulée comme

$$x^{n+1} = \mathrm{prox}_{\tau\, g}(x^n - \tau\,(\nabla f(x^n) + K^* y^n)). \tag{3.22}$$

De la même manière,

$$\begin{aligned}\hat{y} &= \arg\min_y\ h^*(y) - \langle K\tilde{x}, y\rangle + \frac{1}{2\sigma}\,\|y-\bar{y}\|^2,\\ &= \arg\min_y\ h^*(y) + \frac{1}{2\sigma}\,\|y - (\bar{y} + \sigma\, K\tilde{x})\,\|^2,\\ &= \mathrm{prox}_{\sigma\, h^*}(\bar{y} + \sigma\, K\tilde{x}).\end{aligned}$$

La seconde étape de l'algorithme s'exprime donc comme

$$y^{n+1} = \mathrm{prox}_{\sigma\, h^*}(y^n + \sigma\, K(2x^{n+1} - x^n)). \tag{3.23}$$

Maintenant, en considérant un point fixe (x^*, y^*) de l'algorithme et en raison du fait que

$$\begin{aligned} r = \mathrm{prox}_f(s) &\iff s - r \in \partial f(r)\\ &\iff \forall t,\ f(t) \geq f(r) + \langle s - r, t - r\rangle,\end{aligned}$$

la relation (3.22) donne que $\forall x \in \mathcal{X}$,

$$g(x) \geq g(x^*) - \langle x - x^*, \nabla f(x^*) + K^* y^*\rangle,$$

tandis que la relation (3.23) mène à $\forall y \in \mathcal{Y}$,

$$h^*(y) \geq h^*(y^*) + \langle y - y^*, Kx^*\rangle.$$

En sommant les deux inégalités, il vient

$$g(x^*) - h^*(y) - \langle x - x^*, \nabla f(x^*)\rangle - \langle Kx, y^*\rangle \leq g(x) - h^*(y^*) - \langle Kx^*, y\rangle.$$

Mais la convexité de f, *i.e.*, $f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \leq f(x)$, implique que

$$g(x^*) - h^*(y) + f(x^*) + \langle Kx^*, y \rangle \leq g(x) - h^*(y^*) + \langle Kx, y^* \rangle + f(x).$$

D'où $\forall x \in \mathcal{X}$, $\forall y \in \mathcal{Y}$, $\mathcal{L}(x^*, y) \leq \mathcal{L}(x, y^*)$, montrant que (x^*, y^*) est un point-selle du Lagrangien associé, désigné ici par $\mathcal{L}$, ce qui conclut ce commentaire.

Nous rendons maintenant plus explicite le lien entre notre problème (de dimension finie) $(\bar{\mathcal{P}})$ et le cadre général de [21]. Le problème $(\bar{\mathcal{P}})$ est un cas particulier de (CP) avec :

- $x = (u, v) \in X^L \times X^L$ et $y = p \in Y^L$,
- $K = \begin{pmatrix} \nabla & 0 \end{pmatrix}$,
- $f(u, v) = \frac{1}{2} \|u - v\|^2$ (avec $\nabla_u f(u, v) = u - v$ et $\nabla_v f(u, v) = v - u$),
- $g(u, v) = \frac{\mu}{2} \|u - s(\theta) - w\|^2 + i_{\mathcal{C}_1}(u) + i_{\mathcal{C}_2}(v)$,
- $h^* = i_{\mathcal{B}}$.

Un calcul simple permet d'obtenir $\|K\| \leq 2\sqrt{2}$ et $L_f = 2$.

Dans notre cas, la première étape de l'algorithme est la suivante

$$\begin{pmatrix} u^{n+1} \\ v^{n+1} \end{pmatrix} = \mathrm{prox}_{\tau g} \begin{pmatrix} u^n - \tau \nabla_u f(u^n, v^n) + \mathrm{div}\, p^n \\ v^n - \tau \nabla_v f(u^n, v^n) \end{pmatrix}.$$

La propriété de séparabilité des variables u et v de g permet de conclure que

$$\mathrm{prox}_{\tau g}(u, v) = \begin{pmatrix} \mathrm{prox}_{\tau \phi}(u) \\ P_{\mathcal{C}_2}(v) \end{pmatrix},$$

avec $\phi(u) = \frac{\mu}{2} \|u - s(\theta) - w\|^2 + i_{\mathcal{C}_1}(u)$.

D'une part, $P_{\mathcal{C}_2}(v) = (P_{\mathcal{C}_2}(v^l))_{l \in \{1, \cdots, L\}}$ avec

$$P_{\mathcal{C}_2}(v^l) = v^l + \frac{\alpha^l - \sum\limits_{(i,j) \in \mathcal{G}} v^l_{i,j}}{N^2} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

de sorte que

$$(v^l)^{n+1} = (1 - \tau)(v^l)^n + \tau (u^l)^n + \frac{\alpha^l - \sum\limits_{(i,j) \in \mathcal{G}} ((1 - \tau)(v^l_{i,j})^n + \tau (u^l_{i,j})^n)}{N^2} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

D'autre part,

$$\begin{aligned}\mathrm{prox}_{\tau\,\phi}(u) &= \underset{\tilde{u}=(\tilde{u}^l)\in\mathcal{C}_1}{\arg\min}\ \frac{\tau\,\mu}{2}\,\|\tilde{u}-s(\theta)-w\|^2+\frac{1}{2}\,\|u-\tilde{u}\|^2,\\ &= \underset{\tilde{u}=(\tilde{u}^l)\in\mathcal{C}_1}{\arg\min}\ \frac{\tau\,\mu+1}{2}\,\|\tilde{u}-x\|^2,\end{aligned}$$

avec $x=\dfrac{\tau\,\mu}{\tau\,\mu+1}\,(s(\theta)+w)+\dfrac{1}{\tau\,\mu+1}\,u$. Ce dernier problème, séparable par rapport à chaque $\tilde{u}_{i,j}$, $(i,j)\in\mathcal{G}$, peut se décomposer en N^2 problèmes plus petits dans $\mathbb{R}^L$ et la solution est donnée par

$$\Big(\mathrm{prox}_{\tau\,\phi}(u)\Big)_{(i,j)\in\mathcal{G},\,l\in\{1,\cdots,L\}}=P_{[0,1]}(x^l_{i,j}-\lambda),$$

où $\lambda(=\lambda_{i,j})$ est la solution de l'équation

$$\sum_{l=1}^{L}P_{[0,1]}(x^l_{i,j}-\lambda)=1.$$

La seconde étape de l'algorithme s'exprime comme

$$\begin{aligned}p^{n+1} &= \mathrm{prox}_{\sigma\,h^*}\left(p^n+\sigma\,\nabla(2u^{n+1}-u^n)\right),\\ &= P_{\mathcal{B}}\left(p^n+\sigma\,\nabla(2u^{n+1}-u^n)\right).\end{aligned}$$

Invoquant à nouveau des propriétés de séparabilité, il vient avec $z\in Y^L$,

$$P_{\mathcal{B}}(z)=\Big(P_{\mathcal{B}_l}(z^l)\Big)_{l\in\{1,\cdots,L\}}$$

où

$$\Big(P_{\mathcal{B}_l}(z^l)\Big)_{i,j}=\frac{z^l_{i,j}}{\max\left(1,\dfrac{|z^l_{i,j}|}{g^l_{i,j}}\right)}.$$

Par souci de clarté et d'exhaustivité, nous terminons cette section avec un résultat de convergence dont la preuve est seulement esquissée dans [21] en tant que remarque ([21, Remarque 3]). Nous pensons qu'il s'agit d'un résultat intéressant qui mérite d'être détaillé. Nous énonçons ce théorème, dérivé du critère de la règle de descente [21, Lemme 1], en utilisant les notations générales de Chambolle et Pock dans le cas de la dimension finie, soit le cadre dans lequel nous travaillons.

Théorème 3.4.7.

Soit (x_n, y_n) une suite générée par l'Algorithme 4. Supposons que les paramètres de pas τ, $\sigma > 0$ sont choisis tels que

$$\left(\frac{1}{\tau} - L_f\right) \frac{1}{\sigma} > \|K\|^2, \tag{H}$$

avec $L_f > 0$. Alors il existe un point-selle (x^*, y^*) tel que $x^n \to x^*$ et $y^n \to y^*$.

Démonstration. Nous rappelons d'abord que la structure de point-selle considéré s'exprime comme

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \mathcal{L}(x, y) = \langle Kx, y\rangle + f(x) + g(x) - h^*(y), \tag{CP}$$

et que l'itération générale de l'algorithme est donnée

$$(\hat{x}, \hat{y}) = PD_{\tau,\sigma}(\bar{x}, \bar{y}, \tilde{x}, \tilde{y}),$$

avec

$$\begin{cases} \hat{x} &= \arg\min_x f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + g(x) + \langle Kx, \tilde{y}\rangle + \frac{1}{\tau} D_x(x, \bar{x}), \\ \hat{y} &= \arg\min_y h^*(y) - \langle K\tilde{x}, y\rangle + \frac{1}{\sigma} D_y(y, \bar{y}). \end{cases}$$

Les entrées de l'algorithme sont donc les points $(\bar{x}, \bar{y})$, ainsi que les points intermédiaires $(\tilde{x}, \tilde{y})$, tandis que les sorties sont les points engendrés $(\hat{x}, \hat{y})$.
Ici D_x et D_y sont des fonctions de proximité/distance de Bregman (voir [21, p. 256] pour de plus amples informations) choisies de manière à ce que $D_x(x, \bar{x}) = \frac{1}{2}\|x - \bar{x}\|^2$ (respectivement $D_y(y, \bar{y}) = \frac{1}{2}\|y - \bar{y}\|^2$) dans notre cadre, ce choix demeurant celui le plus commun. Aussi, sur le plan algorithmique, une itération est appliquée avec $\bar{x} = x^n$, $\bar{y} = y^n$, $\tilde{x} = 2x^{n+1} - x^n$ et $\tilde{y} = y^n$. Le Lemme 1 extrait de [21, Lemme 1, p. 257 avec preuve] stipule que, à condition que l'itération générale précédente soit valable, pour tout $x \in \mathcal{X}$ et pour tout $y \in \mathcal{Y}$, nous avons

$$\begin{aligned} \mathcal{L}(\hat{x}, y) - \mathcal{L}(x, \hat{y}) \leq \quad & \frac{1}{\tau} D_x(x, \bar{x}) - \frac{1}{\tau} D_x(x, \hat{x}) - \frac{1}{\tau} D_x(\hat{x}, \bar{x}) + \frac{L_f}{2}\|\hat{x} - \bar{x}\|^2 \\ & + \frac{1}{\sigma} D_y(y, \bar{y}) - \frac{1}{\sigma} D_y(y, \hat{y}) - \frac{1}{\sigma} D_y(\hat{y}, \bar{y}) + \langle K(x - \hat{x}), \tilde{y} - \hat{y}\rangle \\ & - \langle K(\tilde{x} - \hat{x}), y - \hat{y}\rangle. \end{aligned}$$

En appliquant [21, Lemme 1] avec $\hat{x} = x^{n+1}$, $\hat{y} = y^{n+1}$ et D_x, D_y comme définies ci-dessus,

cela donne que $\forall x \in \mathcal{X}$ et $\forall y \in \mathcal{Y}$, nous avons :

$$\begin{aligned}\mathcal{L}(x^{n+1}, y) - \mathcal{L}(x, y^{n+1}) \leq & \frac{1}{2\tau}\|x - x^n\|^2 - \frac{1}{2\tau}\|x - x^{n+1}\|^2 - \frac{1}{2\tau}\|x^{n+1} - x^n\|^2 \\ & + \frac{L_f}{2}\|x^{n+1} - x^n\|^2 + \frac{1}{2\sigma}\|y - y^n\|^2 - \frac{1}{2\sigma}\|y - y^{n+1}\|^2 \\ & - \frac{1}{2\sigma}\|y^{n+1} - y^n\|^2 + \langle K(x - x^{n+1}), y^n - y^{n+1}\rangle \\ & - \langle K(x^{n+1} - x^n), y - y^{n+1}\rangle.\end{aligned}$$

En remarquant ensuite que $\left\{ \begin{array}{rcl} x - x^{n+1} & = & x - x^n + x^n - x^{n+1} \\ x^{n+1} - x^n & = & x^{n+1} - x + x - x^n \end{array} \right.$, et par conséquent que

$$\begin{aligned}&\langle K(x - x^{n+1}), y^n - y^{n+1}\rangle - \langle K(x^{n+1} - x^n), y - y^{n+1}\rangle \\ &\quad = \langle K(x - x^n), y^n - y^{n+1}\rangle + \langle K(x^n - x^{n+1}), y^n - y^{n+1}\rangle \\ &\qquad - \langle K(x^{n+1} - x), y - y^{n+1}\rangle - \langle K(x - x^n), y - y^{n+1}\rangle, \\ &\quad = - \langle K(x - x^n), y - y^n\rangle + \langle K(x^{n+1} - x^n), y^{n+1} - y^n\rangle + \langle K(x - x^{n+1}), y - y^{n+1}\rangle,\end{aligned}$$

il vient

$$\begin{aligned}\mathcal{L}(x^{n+1}, y) - \mathcal{L}(x, y^{n+1}) \leq & \left[\frac{1}{2\tau}\|x - x^n\|^2 + \frac{1}{2\sigma}\|y - y^n\|^2 - \langle K(x - x^n), y - y^n\rangle\right] \\ & - \left[\frac{1}{2\tau}\|x - x^{n+1}\|^2 + \frac{1}{2\sigma}\|y - y^{n+1}\|^2 - \langle K(x - x^{n+1}), y - y^{n+1}\rangle\right] \\ & - \left[\frac{1}{2\tau}\|x^{n+1} - x^n\|^2 + \frac{1}{2\sigma}\|y^{n+1} - y^n\|^2 - \langle K(x^{n+1} - x^n), y^{n+1} - y^n\rangle\right. \\ & \left. - \frac{L_f}{2}\|x^{n+1} - x^n\|^2\right]. \qquad (3.24)\end{aligned}$$

En raison de l'hypothèse (H), selon laquelle $\left(\frac{1}{\tau} - L_f\right)\frac{1}{\sigma} > \|K\|^2$, les quantités entre crochets sont positives. Pour montrer cela, concentrons nous uniquement sur la dernière quantité entre crochets, le même raisonnement s'appliquant aux deux premières. En utilisant l'inégalité de Cauchy-Schwarz combinée à l'inégalité de Young avec un paramètre $\varepsilon > 0$, nous obtenons

$$\begin{aligned}\langle K(x^{n+1} - x^n), y^{n+1} - y^n\rangle &\leq \|K\|\,\|x^{n+1} - x^n\|\,\|y^{n+1} - y^n\| \\ &\leq \frac{\|K\|}{2\varepsilon}\|x^{n+1} - x^n\|^2 + \frac{\|K\|\varepsilon}{2}\|y^{n+1} - y^n\|^2,\end{aligned}$$

de sorte qu'en posant $\varepsilon = \frac{1}{\sqrt{\sigma\left(\frac{1}{\tau}-L_f\right)}}$, cela nous mène à l'inégalité

$$\left(\frac{1}{2\tau}-\frac{L_f}{2}\right)\|x^{n+1}-x^n\|^2+\frac{1}{2\sigma}\|y^{n+1}-y^n\|^2-\langle K(x^{n+1}-x^n),y^{n+1}-y^n\rangle$$
$$\geq\left(\frac{1}{2\tau}-\frac{L_f}{2}-\sqrt{\sigma\left(\frac{1}{\tau}-L_f\right)}\frac{\|K\|}{2}\right)\|x^{n+1}-x^n\|^2+\left(\frac{1}{2\sigma}-\frac{\|K\|}{2\sqrt{\sigma\left(\frac{1}{\tau}-L_f\right)}}\right)\|y^{n+1}-y^n\|^2.$$

L'hypothèse (H) permet de conclure que les facteurs de pondération de $\|x^{n+1}-x^n\|^2$ et $\|y^{n+1}-y^n\|^2$ sont positifs, ou de manière équivalente qu'il existe $\xi>0$ de façon à ce que

$$\left(\frac{1}{2\tau}-\frac{L_f}{2}\right)\|x^{n+1}-x^n\|^2+\frac{1}{2\sigma}\|y^{n+1}-y^n\|^2-\langle K(x^{n+1}-x^n),y^{n+1}-y^n\rangle$$
$$\geq\xi\left(\|x^{n+1}-x^n\|^2+\|y^{n+1}-y^n\|^2\right). \tag{3.25}$$

Une conséquence immédiate est que l'inégalité (3.24) se réduit à

$$\mathcal{L}(x^{n+1},y)-\mathcal{L}(x,y^{n+1})\leq\left[\frac{1}{2\tau}\|x-x^n\|^2+\frac{1}{2\sigma}\|y-y^n\|^2-\langle K(x-x^n),y-y^n\rangle\right]$$
$$-\left[\frac{1}{2\tau}\|x-x^{n+1}\|^2+\frac{1}{2\sigma}\|y-y^{n+1}\|^2-\langle K(x-x^{n+1}),y-y^{n+1}\rangle\right]. \tag{3.26}$$

Nous prenons comme (x,y) particulier un point-selle (x^*,y^*) du Lagrangien $\mathcal{L}$, dont l'existence est garantie dans notre cas par le Théorème 3.4.4. Cela implique, par définition d'un point-selle, que $\mathcal{L}(x^{n+1},y^*)-\mathcal{L}(x^*,y^{n+1})\geq 0$. Alors l'inégalité (3.26) conduit à :

$$\frac{1}{2\tau}\|x^*-x^{n+1}\|^2+\frac{1}{2\sigma}\|y^*-y^{n+1}\|^2-\langle K(x^*-x^{n+1}),y^*-y^{n+1}\rangle$$
$$-\frac{1}{2\tau}\|x^*-x^n\|^2-\frac{1}{2\sigma}\|y^*-y^n\|^2+\langle K(x^*-x^n),y^*-y^n\rangle\leq 0.$$

En sommant de $n=0$ à $N-1$, nous obtenons :

$$\frac{1}{2\tau}\|x^*-x^N\|^2-\frac{1}{2\tau}\|x^*-x^0\|^2+\frac{1}{2\sigma}\|y^*-y^N\|^2-\frac{1}{2\sigma}\|y^*-y^0\|^2$$
$$+\langle K(x^*-x^0),y^*-y^0\rangle-\langle K(x^*-x^N),y^*-y^N\rangle\leq 0,$$

et à l'aide d'une estimation équivalente à (3.25), il s'ensuit, toujours en utilisant les inégalités de Cauchy-Schwarz et de Young, que

$$\xi\left(\|x^*-x^N\|^2+\|y^*-y^N\|^2\right)\leq\left(\frac{1}{2\tau}+\frac{\|K\|}{2}\right)\|x^*-x^0\|^2+\left(\frac{1}{2\sigma}+\frac{\|K\|}{2}\right)\|y^*-y^0\|^2,$$

montrant que la suite (x^n, y^n) est une suite bornée. Il est alors possible d'extraire une suite $(x^{\Psi(n)}, y^{\Psi(n)})$ qui converge (fortement) vers $(\hat{\hat{x}}, \hat{\hat{y}})$ (puisque nous travaillons en dimension finie). Revenons maintenant à l'inégalité (3.24). En procédant comme précédemment avec $(x, y) = (x^*, y^*)$ et en sommant les inégalités pour n allant de 0 à $N-1$, couplé avec l'estimation (3.25), nous pouvons montrer que

$$\begin{aligned}
&\xi \left(\sum_{n=0}^{N-1} \|x^{n+1} - x^n\|^2 + \sum_{n=0}^{N-1} \|y^{n+1} - y^n\|^2 \right) \\
\leq &\frac{1}{2\tau} \|x^* - x^0\|^2 + \frac{1}{2\sigma} \|y^* - y^0\|^2 - \langle K(x^* - x^0), y^* - y^0 \rangle \\
&- \frac{1}{2\tau} \|x^* - x^N\|^2 - \frac{1}{2\sigma} \|y^* - y^N\|^2 + \langle K(x^* - x^N), y^* - y^N \rangle, \qquad (3.27)
\end{aligned}$$

la dernière ligne contenant une quantité négative toujours d'après l'hypothèse (H). D'où

$$\begin{aligned}
&\xi \left(\sum_{n=0}^{N-1} \|x^{n+1} - x^n\|^2 + \sum_{n=0}^{N-1} \|y^{n+1} - y^n\|^2 \right) \\
\leq &\frac{1}{2\tau} \|x^* - x^0\|^2 + \frac{1}{2\sigma} \|y^* - y^0\|^2 - \langle K(x^* - x^0), y^* - y^0 \rangle.
\end{aligned}$$

La suite $(\mathcal{S}(N))_{N \in \mathbb{N}^*}$, de terme général $\mathcal{S}(N) = \sum_{n=0}^{N-1} \|x^{n+1} - x^n\|^2$, est donc croissante et bornée supérieurement. Par conséquent, cette suite converge et $\lim_{n \to +\infty} (x^{n+1} - x^n) = 0$. Nous procédons de la même manière pour obtenir que $\lim_{n \to +\infty} (y^{n+1} - y^n) = 0$. Ces observations impliquent que $(x^{\Psi(n)-1}, y^{\Psi(n)-1})$ converge également vers $(\hat{\hat{x}}, \hat{\hat{y}})$, en prenant $n := \Psi(n) - 1$ dans le résultat précédent, qui est alors un point fixe d'une itération de l'Algorithme 4, donc un point-selle du Lagrangien $\mathcal{L}$ d'après la Remarque 3.4.6.

Pour la dernière fois, nous revenons à l'inégalité (3.26) avec $(x, y) = (\hat{\hat{x}}, \hat{\hat{y}})$ et sommons les inégalités de $n = \Psi(n)$ à $N-1$ avec $N > \Psi(n)$. Il vient

$$\xi \left(\|\hat{\hat{x}} - x^N\|^2 + \|\hat{\hat{y}} - y^N\|^2 \right) \leq \left(\frac{1}{2\tau} + \frac{\|K\|}{2} \right) \|\hat{\hat{x}} - x^{\Psi(n)}\|^2 + \left(\frac{1}{2\sigma} + \frac{\|K\|}{2} \right) \|\hat{\hat{y}} - y^{\Psi(n)}\|^2,$$

ce qui prouve que $x^N \to \hat{\hat{x}}$ et $y^N \to \hat{\hat{y}}$ lorsque N tend vers $+\infty$. □

Cette preuve clôt la partie théorique de ce travail, et nous permet d'aborder la partie expérimentale.

3.5 Expériences et résultats

Nous évaluons notre méthode régularisée incluant des contraintes géométriques sur le jeu de données SegTHOR introduit dans la sous-section 1.3.2. Pour rappel, il se compose

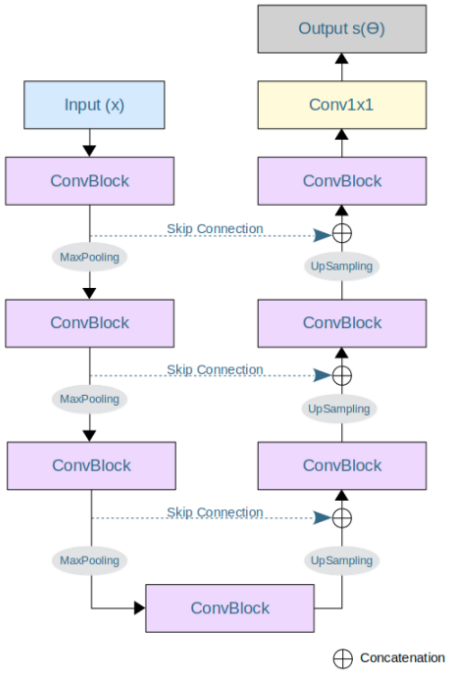


FIGURE 3.2 – L'architecte réseau de sU-Net, une version simplifiée de U-Net, où chaque ConvBlock consiste en deux opérations de convolution avec des noyaux de taille 3×3 suivies par une fonction d'activation ReLU et une normalisation par lots. La résolution spatiale est réduite par max-pooling puis récupérée par un sur-échantillonnage bilinéaire.

de 60 images de scanner et l'objectif est de segmenter quatre organes à risque : l'œsophage, le cœur, la trachée et l'aorte. Les difficultés spécifiques de ce dataset, telles que le manque de contraste ou encore le déséquilibre entre les classes, fait de lui un bon candidat pour tirer profit de cette méthode. Nous exposons dans la suite les conditions expérimentales.

3.5.1 Implémentation

Dans toutes nos expériences, l'architecture réseau que nous utilisons est une version simplifiée de U-Net, appelée sU-Net [76], précédemment introduite dans la sous-section 1.3.3 et rappelée sur la Figure 3.2, qui détient moins de couches connectées que l'architecture originale. Nous évaluons la fonction de perte $\mathcal{F}$ comme, d'une part, le Dice généralisé (Eq. 1.9) seul, et d'autre part, combiné avec le terme de Mumford-Shah (MS). Pour des raisons de complexité algorithmique, les images sont normalisées et rognées à partir du centre pour obtenir des images de taille 304×240 pixels. Dans certaines expériences, une technique d'augmentation de données est mise en œuvre pour tripler artificiellement le nombre d'images d'entraînement. Nous procédons de la même manière que pour les premières expériences (sous-section 1.3.4), et comme décrit dans [93], chaque image est modifiée par une transformation affine aléatoire d'une part, et est déformée en utilisant

un champ de déformation dense obtenu par le biais d'une grille de points de contrôle $2 \times 2 \times 2$ puis une interpolation B-Spline d'autre part. Les paramètres de sU-Net sont initialisés aléatoirement par une technique d'initialisation de Glorot [55] et actualisés en utilisant une technique SGD. Concernant les hyperparamètres du réseau, l'optimisation s'effectue en suivant une technique de quadrillage, en particulier nous fixons le taux d'apprentissage initial à 10^{-3} et la taille du lot à 4 pour la totalité des expériences décrites. Ensuite, de la même façon, le terme de pondération μ est fixé à 0.5. Les paramètres γ et λ de l'algorithme de Douglas-Rachford sont tous deux fixés à 1. Quant aux paramètres σ et τ de l'algorithme Primal-Dual, ils sont respectivement réglés à 0.01 et 0.4 pour respecter les conditions énoncées dans le Théorème 3.4.7. L'ensemble du processus d'apprentissage se déroule sur un maximum de 150 époques. Pour résoudre les sous-problèmes, les algorithmes DR ou PD sont exécutés pendant 100 itérations afin de conserver un temps d'exécution raisonnable, ce qui paraît suffisant pour converger vers le résultat souhaité. L'ensemble du code est développé en Python avec la librairie Pytorch.

3.5.2 Protocole et métriques d'évaluation

Le dataset se divise en deux : 40 images sont utilisées pour l'entraînement et les 20 restantes sont conservées pour l'étape d'inférence, en suivant le protocole standardisé donné dans le challenge SegTHOR. Dans cette section, tous les résultats de segmentation sont évalués en 3D pour chaque patient, en utilisant deux métriques populaires, à savoir le score de Dice et la distance de Hausdorff (HD). Cette dernière mesure quantifie l'aptitude du modèle à éliminer les valeurs aberrantes. Ces indicateurs doivent être évalués à la lumière de critères plus qualitatifs (cohérence spatiale, régularité, respect des contraintes géométriques prescrites, segmentations plus fidèles à la réalité anatomique, moins d'excroissances et détections erronées, etc.) puisque le coefficient de Dice, par exemple, par son effet de moyenne sous-jacent, tend à masquer les disparités visuelles que nous pouvons observer entre les cas contraints et non contraints.

3.5.3 Résultats

Algorithmes de Douglas-Rachford vs Primal-Dual Commençons tout d'abord par comparer les performances des deux algorithmes proposés, celui de Douglas-Rachford et Primal-Dual. L'algorithme de Douglas-Rachford est plutôt gourmand en temps et en mémoire car il nécessite la résolution d'un système linéaire, ce qui requiert en terme d'implémentation l'interfaçage entre GPU et CPU, afin d'utiliser les outils appropriés d'algèbre linéaire *sparse* et de gérer le plus efficacement possible les ressources mémoire. Par conséquent, les résultats sont obtenus en entraînant le modèle proposé sur le dataset SegTHOR initial, sans augmentation de données. Comme l'algorithme DR solutionne le problème de minimisation sujet aux contraintes dures, tandis que celui de type PD traite un problème min-max avec une variable (duale) additionnelle et une contrainte relâchée, nous pouvons anticiper une meilleure précision de segmentation de la part de l'algorithme

TABLE 3.1 – Résultats de segmentation (moyenne±écart-type) obtenus avec notre méthode contrainte, résolue avec les algorithmes DR et PD. La fonction de perte $\mathcal{F}$ combine les termes de Dice et MS. Métriques utilisées : score de Dice et distance de Hausdorff (HD). Aucune augmentation de données n'est utilisée pendant l'entraînement.

Méthode	Algorithme	Métriques	Aorte	Cœur	Trachée	Œsophage
Proposition : sU-Net +geom. cont.	DR	Dice %	94.07 ± 1.57	91.43 ± 9.49	88.75 ± 4.42	82.26 ± 4.80
		HD (mm)	10.83 ± 5.03	29.90 ± 16.89	21.95 ± 8.94	37.98 ± 28.37
	PD	Dice %	93.85 ± 1.73	89.66 ± 3.70	88.65 ± 4.29	82.04 ± 4.07
		HD (mm)	13.69 ± 9.15	38.06 ± 16.74	23.32 ± 8.76	40.29 ± 24.22

DR. Toutefois, en pratique, comme le montrent la Table 3.1 et la Figure 3.3, les résultats obtenus avec l'algorithme DR sont globalement similaires par comparaison avec l'algorithme PD, ou légèrement améliorés pour le cœur. Ainsi, dans les expériences restantes, nous adoptons l'algorithme PD, ce qui constitue un bon compromis entre précision et efficacité de calcul.

Évaluation de la contrainte géométrique proposée Il a été clairement démontré que l'augmentation de données améliore les performances dans cette application [76] (cf. sous-section 1.3.4) : nous en tirons parti dans ce qui suit. Dans ce cas, uniquement la méthode contrainte impliquant la formulation duale de la TV_g avec l'algorithme PD, présentée dans la sous-section 3.4.3, peut être mise en œuvre et donc entraînée. Dans les résultats reportés dans la Table 3.2, nous présentons également ceux obtenus avec un réseau profond non contraint, soit avec la même architecture sU-Net, dans laquelle les paramètres sont mis à jour seulement avec une technique de rétropropagation du gradient, en dehors de tout autre schéma d'optimisation (ligne 'non contraint' dans la Table 3.2). Nous évaluons aussi la contribution du terme MS, dans les lignes 'Dice' vs 'Dice + MS' dans la Table 3.2. De plus, nous visualisons les résultats de segmentation de six patients différents sur la Figure 3.4.

Concernant la trachée (en vert clair sur la Figure 3.4), facilement distinguable dans l'imagerie scanner, les scores de Dice et les distances de Hausdorff sont similaires pour toutes les méthodes. De même, pour l'aorte (en jaune sur la Figure 3.4), les valeurs de Dice sont sensiblement équivalentes pour toutes les méthodes tandis que les distances de Hausdorff moyennées sur les vingt patients s'avèrent être plus faibles grâce au terme de MS. Les anomalies que ces modèles sont capables de corriger sont illustrées sur la Figure 3.4(e). Pour les deux derniers OAR, le cœur et l'œsophage, l'ajout du terme de MS en combinaison avec les contraintes géométriques améliorent les valeurs de Dice ainsi que les distances de Hausdorff. Dans la suite, nous approfondissons notre analyse sur ces deux organes.

Le cœur (en vert foncé sur la Figure 3.4) est un organe de taille importante qui englobe

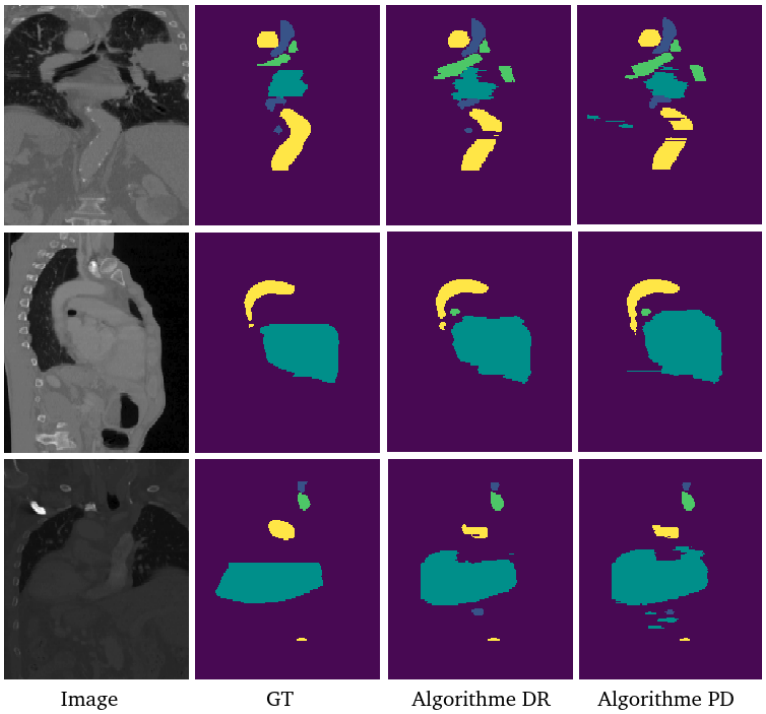


FIGURE 3.3 – Résultats de segmentation pour trois patients de l'aorte (jaune), la trachée (vert clair), le cœur (vert foncé) et l'œsophage (bleu) obtenus avec les algorithmes de Douglas-Rachford et Primal-Dual.

de nombreux tissus et structures différents, rendant ses contours flous et sa segmentation difficile. Cela explique pourquoi il résulte de la méthode non contrainte sans le terme d'homogénéité un organe largement sur-segmenté comme illustré sur la Figure 3.4, troisième colonne. L'ajout du terme MS ou des contraintes et de la régularisation spatiale améliore clairement les résultats sur les plans quantitatif (Tab. 3.2, lignes 2 et 3) et qualitatif (Fig. 3.4, colonnes 4 et 5). La combinaison de ces deux éléments corrige les valeurs aberrantes persistantes (colonne 6). En effet, un score de Dice moyen de pratiquement 93% est atteint, soit 1 à 2 pp de plus que les deux méthodes précédentes et plus de 7 pp supérieur à la première méthode sans contrainte. Le résultat apparaît encore plus évident pour la distance de Hausdorff qui chute à 20 mm, soit environ 6, 1.5 et 2 fois moins que, respectivement, les méthodes non contraintes sans et avec le terme MS et la méthode contrainte sans le terme de MS. Ces analyses sont étayées par le test des rangs signés de Wilcoxon qui confirme que les méthodes susmentionnées (en gras dans la Table 3.2) sont significativement différentes entre les approches non contraintes et contraintes, sans et avec le terme de MS (p-value < 0.05). Cela reflète les propriétés fines et qualitatives qui peuvent être observées dans le cas contraint, en opposition au cas non contraint, et spécifiquement une segmentation plus fidèle à la réalité anatomique, moins d'excroissances et de détections erronées.

TABLE 3.2 – Résultats de segmentation (moyenne±écart-type) obtenus sans et avec les contraintes géométriques proposées. $\mathcal{F}$ est la fonction de perte. Métriques utilisées : score de Dice et distance de Hausdorff (HD). Les résultats les plus significatifs sont en gras.

Méthode	$\mathcal{F}$	Métriques	Aorte	Cœur	Trachée	Œsophage
sU-Net (non contraint).	Dice	Dice %	94.17 ± 2.28	85.44 ± 5.26	90.46 ± 2.51	81.14 ± 4.96
		HD (mm)	16.25 ± 12.65	116.14 ± 39.19	19.59 ± 8.28	49.25 ± 32.29
	Dice + MS	Dice %	94.39 ± 1.85	91.97 ± 3.36	90.73 ± 2.44	82.76 ± 4.45
		HD (mm)	10.53 ± 6.66	28.57 ± 18.11	19.06 ± 9.38	35.01 ± 24.64
Proposition : sU-Net +geom. cont.	Dice	Dice %	94.39 ± 1.93	91.40 ± 3.34	90.53 ± 2.39	82.02 ± 4.71
		HD (mm)	16.64 ± 12.09	36.90 ± 17.25	18.96 ± 8.60	34.72 ± 24.87
	Dice + MS	Dice %	**94.75 ± 1.53**	**93.04 ± 3.00**	91.05 ±2.55	82.91 ± 4.73
		HD (mm)	9.73 ± 5.54	**20.47 ± 10.52**	18.05 ± 9.63	**23.97 ± 17.69**

Comme déjà expliqué dans le Chapitre 1, l'œsophage (en bleu sur la Figure 3.4) est un organe très petit qui se révèle particulièrement difficile à segmenter en raison du manque de contraste, en plus de sa grande variabilité inter- et intra-patient. Les méthodes tendent à le sous-segmenter mais également à détecter des pixels isolés. La pénalisation d'aire aide les méthodes contraintes à contrecarrer cet effet, de même que l'homogénéisation des régions assurée par le terme de MS, comme nous pouvons le voir par exemple sur la Figure 3.4(b-d-e), permettant d'obtenir une distance de Hausdorff plus faible en moyenne.

Enfin, nous relevons un fait intéressant en matière de convergence : les méthodes contraintes convergent plus rapidement comme l'illustre la Figure 3.5, ce qui permet de procéder à l'entraînement sur moins d'époques.

Comparaison avec l'inclusion de la TV_g en tant que couche connectée Nous comparons notre modèle avec une version incluant la variation totale pondérée directement dans la structure du réseau, c'est-à-dire comme une couche connectée. Cela signifie que l'intégration de cet *a priori* de régularité s'effectue en amont de l'évaluation de la fonction de perte. À la manière de Liu *et al.* [87], nous incorporons la régularisation spatiale à l'intérieur du CNN par le biais d'une couche de softmax régularisée. La fonction d'activation est maintenant envisagée comme une solution au problème de minimisation :

$$\min_{u=(u^l)} \langle -o^T, u\rangle + \langle u, \ln u\rangle + TV_g(u) \quad \text{s.c} \quad \sum_{l=1}^{L} u^l_{i,j} = 1.$$

Ce problème se résout itérativement en utilisant l'algorithme du gradient Primal-Dual et il vient :

$$\begin{cases} p^{n+1} = P_{\mathcal{B}}\left(p^n + \sigma\, \nabla u^n\right) \\ u^{n+1} = S(o^T + \operatorname{div} p^{n+1}) \end{cases},$$

où S est la fonction d'activation softmax et o^T la sortie du réseau.

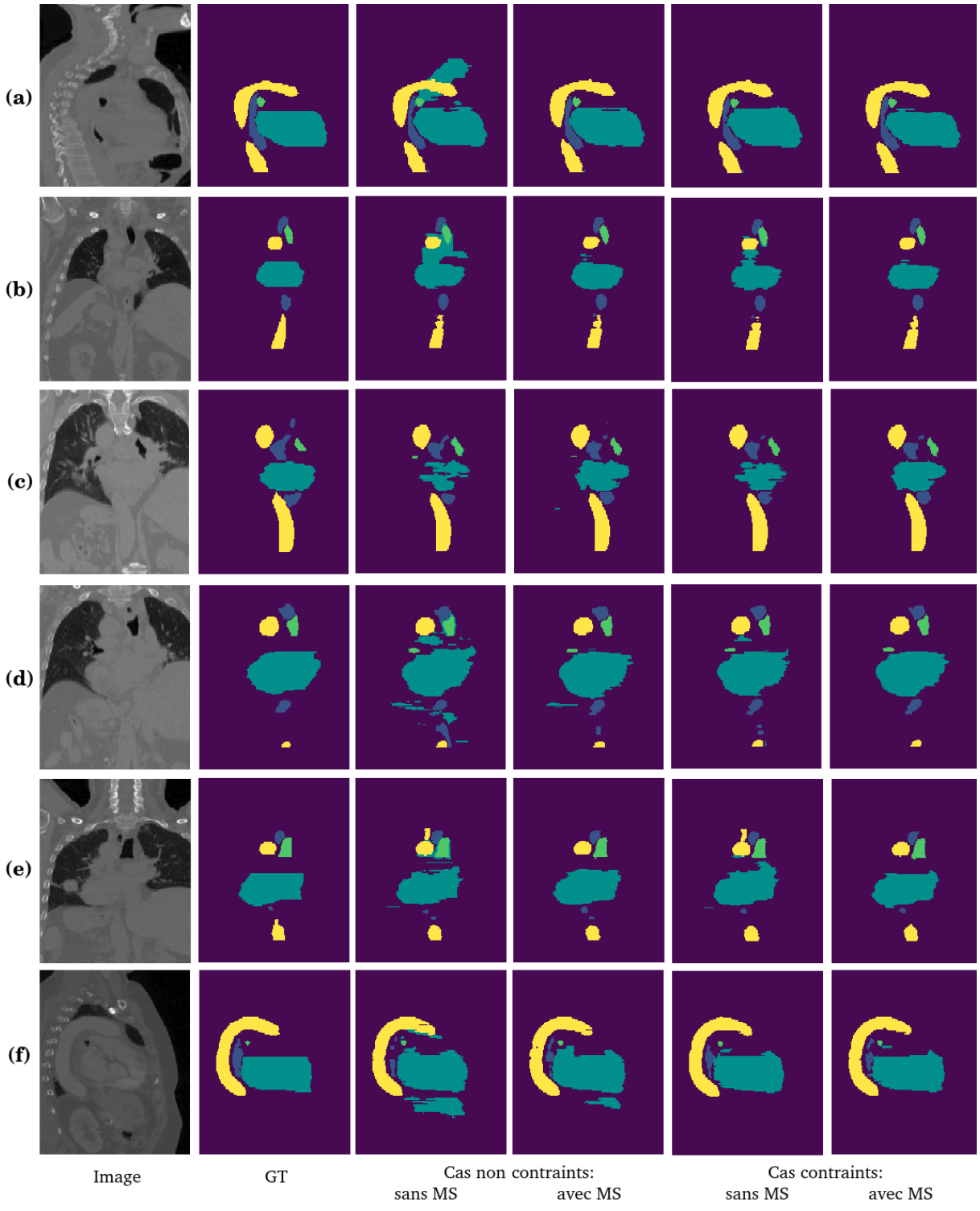


FIGURE 3.4 – Résultats de segmentation pour six patients de l'aorte (jaune), la trachée (vert clair), le cœur (vert foncé) et l'œsophage (bleu) obtenus avec les cas non contraints et contraints, sans ou avec le terme de Mumford-Shah (MS).

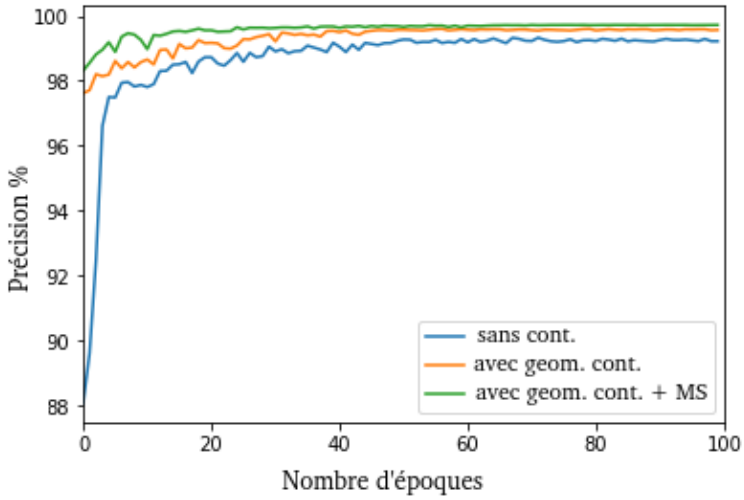


FIGURE 3.5 – Évolution de la précision sur l'ensemble de validation au fil des époques, pour les approches non contrainte (ligne bleue) et contraintes (lignes orange et verte).

La solution du problème de minimisation incluant la TV_g s'apparente à un substitut à la fonction d'activation, donnant lieu à un cadre d'optimisation 'déroulé' dans lequel les itérations s'interprètent comme des couches profondes successives d'un réseau. En conséquence, non seulement les informations spatiales peuvent être pleinement intégrées dans le processus de rétropropagation du gradient pendant la phase d'apprentissage, mais cet *a priori* contribue également à la prédiction dans la phase de test. Les détails complets sont fournis dans [66, Section 3.2]. Comme nous l'avons souligné précédemment, nous pouvons attendre de meilleurs résultats avec ce schéma 'déroulé' puisque les contraintes géométriques sont explicitement appliquées. Cependant, il faut préciser que pour les critères géométriques impliqués, comme par exemple l'aire ou le volume, certaines statistiques peuvent être nécessaires pour les approximer, en particulier dans la phase d'inférence, induisant une certaine incertitude. Notamment, lorsque la variabilité inter/intra-patient apparaît élevée.
En contrepartie, dans la conception de telles architectures, la différentiabilité par rapport aux paramètres du réseau est requise, ce qui, en terme de modélisation mathématique, peut se révéler être un facteur limitant et conduit à des approximations ainsi qu'à des algorithmes imbriqués. Dans cette perspective, notre méthodologie se montre davantage "plug-and-play", dans le sens où de nouveaux *a priori* géométriques peuvent facilement s'intégrer au processus.

De plus, comme mentionné dans [66], en raison des ressources de calcul et de mémoire nécessaires pour la rétropropagation, nous ne pouvons pas effectuer beaucoup d'itérations pour résoudre ce problème de minimisation. Cela signifie que le minimum n'est pas néces-

TABLE 3.3 – Résultats de segmentation (moyenne±écart-type) obtenus avec la TV incluse dans le CNN (inclusion structurale) et dans la fonction de perte (notre proposition). La fonction coût $\mathcal{F}$ combine les termes de Dice et MS. Métriques utilisées : score de Dice et distance de Hausdorff (HD).

Méthode	Métriques	Aorte	Cœur	Trachée	Œsophage
Inclusion structurale	Dice %	94.13 ± 2.17	92.22 ± 3.26	90.34 ± 2.54	82.87 ± 4.23
	HD (mm)	10.84 ± 7.18	24.11 ± 13.30	20.32 ± 8.72	29.86 ± 21.73
Notre proposition	Dice %	94.75 ± 1.53	93.04 ± 3.00	91.05 ±2.55	82.91 ± 4.73
	HD (mm)	9.73 ± 5.54	20.47 ± 10.52	18.05 ± 9.63	23.97 ± 17.69

sairement atteint. Pour une comparaison équitable, nous conservons la même architecture sU-Net et nous ajoutons à la fonction de perte le terme d'homogénéité et la pénalisation d'aire, de la même manière que dans notre proposition. Cela évite les calculs statistiques sur les images, et en particulier sur la taille des organes. Durant l'entraînement du réseau, le facteur de pondération g est calculé à partir de la vérité terrain, tandis qu'en phase d'inférence, il s'obtient à partir de l'image elle-même. Les résultats résumés dans la Table 3.3 suggèrent que l'inclusion d'*a priori* géométriques dans la fonction de perte seule produit des résultats au moins équivalents à ceux obtenus avec des stratégies structurales, ce qui fait de notre modèle un bon compromis entre précision et propriété "plug-and-play".

3.6 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle fonction de perte qui incorpore des contraintes géométriques durant l'entraînement d'un CNN 2D. Cette fonction repose sur un terme de Dice encourageant l'appariement des intensités, la variation totale pondérée induisant l'alignement des contours, un terme de fidélité de Mumford-Shah (MS) constant par morceaux incitant l'homogénéité des intensités lumineuses des pixels, et une pénalisation de l'aire. Le problème de minimisation résultant est séparé en deux sous-problèmes de manière à être résolu par un algorithme ADMM. Le premier sous-problème se formule par rapport à l'inconnue $s(\theta)$, la carte de probabilités prédite de la segmentation, et est lisse. Tandis que le second sous-problème, exprimé par rapport à une variable auxiliaire u mimant $s(\theta)$ et encodant les contraintes géométriques, est convexe et non lisse. L'optimisation du premier sous-problème s'appuie sur une technique SGD, tandis que deux approches sont étudiées pour réaliser l'optimisation du second sous-problème. Un obstacle scientifique réside dans le fait qu'il n'existe pas d'expression *closed form* pour la projection sur l'ensemble réalisable. La première approche duplique la variable auxiliaire u en une nouvelle variable v, chacune d'entre elles supportant une partie des contraintes, inclut la contrainte dure $u = v$, et se résout à l'aide de l'algorithme de Douglas-Rachford (DR). Dans la seconde approche, un couplage plus souple entre u et v, par le biais d'une

pénalisation L^2, est maintenant un substitut à la contrainte dure liant u et v, et mène à un algorithme non classique Primal-Dual (PD). Pour ce dernier, nous avons pu prouver des résultats théoriques, alors qu'ils n'étaient qu'esquissés dans [21].

L'application proposée s'inscrit dans le périmètre complexe de la segmentation d'organes à risque sur des images de tomodensitométrie. Ces images présentent un faible contraste, des contours d'organes parfois pratiquement invisibles, une topologie dépendant de la coupe, une grande variabilité inter-/intra-patient et, en tant que telles, elles constituent un cas d'utilisation parfait pour obtenir un aperçu pratique de notre cadre contraint. Une première expérience montre que l'algorithme PD est le meilleur compromis entre précision et efficacité de calcul. Dans une seconde série d'expériences, il est démontré que l'introduction de cette nouvelle fonction de perte au sein d'un CNN très basique améliore la segmentation de l'aorte, de l'œsophage et du cœur du jeu de données SegTHOR, par rapport à celle uniquement fondée sur la minimisation de la fonction de perte de Dice, tant d'un point de vue qualitatif (segmentations plus fidèles à la réalité anatomique) que quantitatif (scores améliorés). De plus, les résultats obtenus sont au moins aussi précis que ceux obtenus avec un réseau qui inclut une régularisation spatiale dans son architecture, et donc à la fois dans la phase d'apprentissage et celle d'inférence du CNN, faisant de notre modèle un bon compromis entre précision et propriété "plug-and-play". Des travaux futurs pourraient porter sur la généralisation à des cadres faiblement et semi-supervisés ainsi que sur l'inclusion de connaissances préalables topologiques afin de tirer le meilleur parti des relations contextuelles entre les formes contenues dans les images et de contrôler, par exemple, le nombre de composantes connexes. En effet, si ce chapitre met en avant le bénéfice apporté par une régularisation et des contraintes de nature géométrique, quid de celles topologiques ?

CHAPITRE

4 Inclusion de contraintes topologiques dans les réseaux de neurones convolutifs

Dans le même esprit que le chapitre précédent, nous cherchons à perfectionner une tâche de segmentation par apprentissage profond en y intégrant une quantité d'informations connues au préalable. Cette fois, et pour pallier le problème de non-respect du nombre de composantes connexes, nous souhaitons contrôler la topologie des segmentations prédites par le réseau. L'idée consiste à construire des segmentations homéomorphes à un *a priori* connu. Le cadre hybride variationnel/apprentissage profond proposé, incluant des contraintes géométriques et topologiques, aborde le problème de segmentation comme une tâche de recalage appariant la vérité terrain et l'image à étiqueter, fondée sur des principes d'élasticité non linéaire. L'application de conditions d'incompressibilité, qui se modélisent par une contrainte sur le déterminant de la matrice jacobienne de la déformation, garantit la préservation du volume et de la topologie, sans auto-intersection de matière. La méthode mise en place est (i) justifiée théoriquement, via notamment un résultat d'existence de minimiseurs et une analyse asymptotique, (ii) numériquement, en montrant que les sous-problèmes introduits admettent, pour la plupart, des solutions *closed form*, et (iii) validée sur deux jeux de données d'images médicales acquises avec différentes modalités, montrant son caractère généralisable.

4.1 Introduction et motivations

L'inclusion de contraintes géométriques comme connaissances *a priori* dans des réseaux de neurones convolutifs apporte non seulement des améliorations qualitatives et quantitatives significatives (*e.g.* cohérence spatiale), mais atténue également l'aspect boîte noire du processus en lui conférant un plus haut degré d'explicabilité. S'il subsiste des attentes similaires en matière de topologie, il faut cependant souligner que le contrôle de la topologie (englobant le nombre de composantes connexes à segmenter ou les relations contextuelles entre les objets) est plus complexe à réaliser, et par conséquent moins utilisé. Or, le contrôle de la topologie s'avère pertinent, en particulier en imagerie médicale, lorsque l'exigence anatomique n'est pas en accord visuel avec les données : par exemple, malgré sa nature très plissée exhibant des circonvolutions, la structure dépliée intrinsèque du cortex correspond à celle d'un feuillet 2D.

Cette difficulté à faire respecter les exigences en matière de topologie dans les chaînes de traitement s'explique essentiellement par la nature même du concept de topologie qui est intrinsèquement duale, puisqu'il s'agit d'une propriété à la fois globale et locale : de petits changements localisés sur une forme géométrique peuvent modifier sa connectivité globale. De plus, la topologie est un concept défini dans le domaine continu dont les propriétés sont difficiles à transposer au cas discret.

La topologie digitale réconcilie la nature discrète intrinsèque des images numériques avec leur représentation continue, en comblant le fossé entre leurs propriétés et caractéristiques, et leurs propriétés topologiques (*e.g.* connectivité) ou caractéristiques topologiques (*e.g.* frontières) correspondantes. En effet, elle offre un large panel de caractérisations des points ou ensembles de points pertinents sur une image, caractérisations sur lesquelles de nombreux algorithmes sont fondés. Parmi les catégories possibles, nous pouvons citer par exemple : (i) les points simples ([10]), identifiés comme les points dont l'ajout ou la suppression laisse la topologie inchangée et qui reposent sur le calcul de deux nombres topologiques ; (ii) les points multisimples ([120]), vus comme les points dont l'ajout ou la suppression ne crée ou ne supprime pas de poignées ; (iii) les ensembles *well-composed* ([133]), en n dimensions, un ensemble numérique étant bien composé si et seulement si la frontière de son homologue continu est une $(n-1)$-variété. Ces concepts sont généralement entrelacés avec des approches variationnelles (propagation de fronts par ensembles de niveaux comme dans [58, 120]) ou appliqués *a posteriori* pour corriger des défauts topologiques et géométriques de manière séquentielle à l'instar de [121] ou encore [146]. Certains travaux intègrent également ces concepts dans les CNN. Dans [63], Hu *et al.* exploitent l'homologie persistante ([42]) et notamment les nombres de Betti pour concevoir une fonction objectif fondée sur des diagrammes persistants (cf. Fig. 4.1), afin de segmenter des membranes sur des images neuronales de microscope électronique. Les auteurs précisent cependant que l'optimisation de cette fonction se révèle assez coûteuse, complexe et assujettie à des erreurs lorsqu'ils travaillent avec des images de taille réelle.

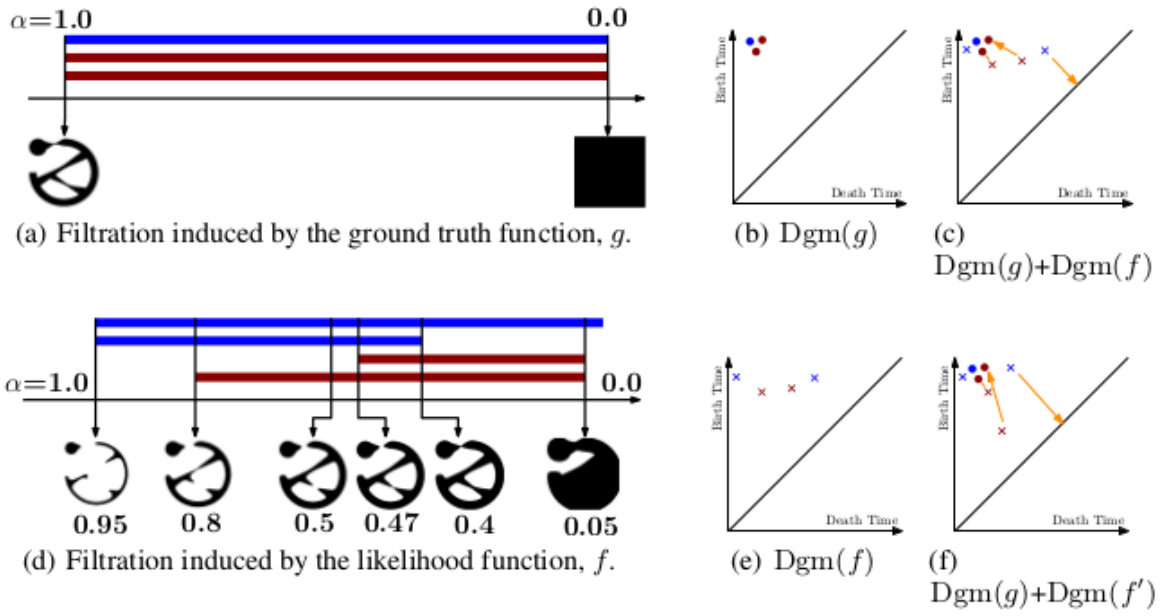


FIGURE 4.1 – Une illustration d'homologie persistante. À gauche les filtrations sur la vérité terrain g et la carte de probabilités f. Les barres de couleurs bleue et bordeaux sont les composantes connexes et les poignées respectivement. (a) Pour g, toutes les structures naissent pour $\alpha = 1.0$ et meurent pour $\alpha = 0$. (d) Pour f, de gauche à droite, naissance de deux composantes, naissance de la poignée plus longue, segmentation pour $\alpha = 0.5$, naissance de la poignée la plus courte, mort de la composante additionnelle, mort des deux poignées. (b) et (e) Les diagrammes persistants de g et f. (c) La superposition des deux diagrammes. Les flèches orange désignent la correspondance entre les points persistants. La composante excédentaire (croix bleue) de f est mise en correspondance avec la ligne diagonale et sera supprimée en déplaçant $Dgm(f)$ vers $Dgm(g)$. (f) La superposition des diagrammes de g et d'une prédiction moins bonne f'. L'appariement est évidemment plus coûteux. (Extrait de [63]).

Clough *et al.* [32] s'appuient également sur la théorie de l'homologie persistante pour mesurer la correspondance entre les nombres de Betti de la prédiction et ceux connus *a priori* via des codes-barres persistants. Cet outil leur permet d'intégrer une fonction de perte, dite topologique, dans l'entraînement d'un CNN pour améliorer, en particulier, la précision des segmentations d'images IRM cardiaques. Ils nuancent cependant leur propos en signalant que la fonction de perte ainsi établie n'est pas pertinente dans certains contextes, notamment si la segmentation prédite exhibe déjà une topologie correcte ou si, à l'inverse, elle est trop éloignée de la topologie attendue. Enfin, la fonction objectif clDice [122] de Shit *et al.* repose sur la notion topologique de squelettisation pour favoriser le respect de certaines propriétés telles que la connectivité et le nombre de composantes connexes dans le contexte de segmentation de structures tubulaires comme des vaisseaux ou des neurones.

Les méthodes précédentes font donc le lien entre la nature discrète des images nu-

mériques et leur représentation continue en manipulant des outils topologiques parfois compliqués. Un autre axe de recherche s'articule autour de formulations purement continues pour introduire des prescriptions topologiques. À cet égard, les modèles conjoints de segmentation et recalage se montrent particulièrement intéressants. En effet, tout comme la segmentation, le recalage constitue un traitement essentiel en analyse d'images. Pour deux images qualifiées de Template et Référence, le recalage consiste à déterminer une déformation régulière optimale φ qui établit une correspondance du Template vers la Référence. Pour de plus amples informations, le lecteur peut se référer à [124] dans lequel Sotiras *et al.* répertorient et classifient de nombreuses méthodes de recalage. Puisque des critères tels que la comparaison de la distribution d'intensité lumineuse ou la correspondance entre les caractéristiques géométriques régissent le recalage, il semble pertinent de connecter les tâches de segmentation et de recalage dans un cadre unique. Plusieurs approches variationnelles traitent de ce problème conjoint ([77, 135, 145]) et s'appuient sur le modèle de segmentation de Chan-Vese [25] pour formuler la fonctionnelle à minimiser. Dans [37], les auteurs introduisent un modèle de recalage guidé par une tâche de segmentation qui préserve la topologie de manière non locale. Récemment, un modèle conjoint 2D de segmentation et recalage, capable de respecter la topologie prescrite, est introduit dans [23]. La recherche de la déformation est dictée par l'équation de Beltrami. Dans [147], Zhang *et al.* s'inspirent de ce modèle pour l'étendre au cas 3D et utilisent en plus la régularisation hyperélastique instaurée dans [16]. Cette idée de tâche simultanée émerge aussi dans la communauté DL. Dans l'approche DeepAtlas [143], il est question d'apprendre parallèlement les paramètres d'un réseau profond pour une tâche de recalage faiblement supervisée, et ceux d'un autre pour une tâche de segmentation semi-supervisée. En particulier, la segmentation profite d'une forme d'augmentation de données fournit par l'autre CNN. Cet apprentissage conjoint améliore les deux processus pour des images IRM du genou et du cerveau. Dans [81], un CNN hybride formé de deux branches distinctes, chacune assignée à l'un des traitements, est optimisé à l'aide d'une fonction de perte commune. De cette façon, le bénéfice est mutuel pour le recalage et la segmentation d'images IRM cérébrale. Un peu différemment, le réseau U-ReSNet [46] se scinde en deux branches après quelques couches profondes, les deux tâches partageant ainsi les paramètres des premières couches connectées. Les auteurs avancent des améliorations à la fois pour la segmentation et le recalage dans un contexte de détection de structures cérébrales.

À la lumière de ces travaux, les *a priori* topologiques témoignent de progrès dans le processus de segmentation automatique, notamment lorsqu'ils agrémentent la fonction de perte d'un CNN. D'un autre côté, les méthodes conjointes prouvent que le recalage peut aiguiller la segmentation. De ces constats naît l'idée d'appréhender l'étape de recalage sous la forme de connaissances préalables au niveau de la fonction objectif d'un réseau de neurones convolutif. L'intérêt consiste à guider la segmentation vers un résultat qui respecte certaines prescriptions topologiques et géométriques et qui, dans une application médicale, se conforme davantage à la réalité anatomique. Dans ce chapitre, nous proposons donc un cadre unifiant approches variationnelle et par apprentissage profond dans

lequel nous introduisons une nouvelle fonction de perte sujette à une régularisation et une contrainte topologiques et géométriques. En effet, celle-ci inclut une première composante assurant l'appariement standard des intensités lumineuses, classiquement modélisée par une fonction objectif de Dice ou d'entropie croisée, et assortie à une seconde composante formulée comme un modèle de recalage appariant la vérité terrain et l'image à étiqueter. Cette dernière composante se fonde sur des principes d'élasticité non linéaire et se décompose en (i) un terme d'attache aux données, (ii) une régularisation sur la déformation qui pénalise les changements de longueurs et (iii) une contrainte dure d'incompressibilité locale portant sur le jacobien de la déformation, et assurant la préservation du volume et de la topologie (relations contextuelles).

Le problème de minimisation de la fonction de perte élaboré se résout en le séparant en deux, relativement à chaque variable impliquée, puis à l'aide d'un schéma alterné. Par conséquent, cette opération donne lieu à deux sous-problèmes. Le premier, lisse et non convexe, traite de l'optimisation des paramètres du réseau et est résolu de façon habituelle par une méthode SGD. Le second se rapporte à la recherche de la déformation du modèle de recalage et, bien qu'il exhibe de bonnes propriétés théoriques, se montre complexe sur le plan numérique du fait de sa non-régularité (liée à la contrainte dure) et non-convexité (liée à la régularisation portant sur la déformation). L'introduction de variables auxiliaires en facilite alors la résolution au moyen d'une technique de *splitting* puisque chaque sous-problème présente désormais des solutions explicites.

Ce travail, en ligne avec notre volonté de créer une synergie entre les deux formalismes, a pour but d'introduire des informations préalables de nature topologique et géométrique pour la segmentation d'images médicales. En ce sens, il concilie le caractère discret de la segmentation avec la dimension continue des méthodes variationnelles. Nos contributions revêtent plusieurs formes (i) de nature méthodologique par la conception d'une fonction de perte inspirée des modèles de recalage d'images afin d'encoder des propriétés géométriques et topologiques ; (ii) de nature théorique par rapport au sous-problème relatif à la déformation à retrouver, avec l'existence de minimiseurs (en s'assurant d'abord de la propriété homéomorphe de la déformation) et une analyse asymptotique ; (iii) de nature numérique par le développement de solutions explicites et une mise en oeuvre optimisée ; (iv) et finalement de nature plus appliquée en évaluant la méthode dans un cas binaire d'images IRM cardiaques et un cas multi-classes de scanners thoraciques, montrant son caractère généralisable.

4.2 Modèle de recalage/segmentation conjoint proposé fondé sur l'apprentissage profond et les approches variationnelles

L'entraînement supervisé d'un CNN à des fins de segmentation nécessite un jeu de données constitué de K images ainsi que de leur vérité terrain associée et notée $\left\{y^k\right\}_{k=1,\cdots,K}$ dans ce chapitre. Plus spécifiquement, $y^k \in \{1, \cdots, L\}$, avec L le nombre de classes à segmenter. Le problème d'optimisation du réseau de neurones étant séparable par rapport à la variable k, nous omettons la dépendance à k à partir de maintenant. Pour chaque image, le CNN fournit en sortie une fonction de segmentation paramétrée par θ, notée $s(\theta)$, et souhaitée aussi proche que possible de y. Dans ce but, l'apprentissage des paramètres θ du réseau requiert la conception d'une fonction de perte $\mathcal{L}$ adaptée et différentiable pour rendre efficace la rétropropagation de l'erreur du gradient. Nous présentons dans la suite la modélisation de cette nouvelle fonction, avec pour objectif d'introduire des prescriptions topologiques et géométriques.

4.2.1 Description du modèle mathématique

Soit le domaine image Ω, un sous-ensemble ouvert, connexe et borné de $\mathbb{R}^2$ à frontière régulière (une façon commode de dire que dans une définition donnée, la régularité de la frontière est telle que tous les arguments ont un sens et qui nous permet d'utiliser des injections compactes de Sobolev, entre autres). Comme introduit en amont, $y : \bar{\Omega} \to \{1, \cdots, L\}$ désigne la vérité terrain supposée être un élément de $BV(\Omega, \{1, \cdots, L\})$ et $s(\theta) \in L^2(\Omega)$ représente la fonction de segmentation prédite par le réseau.

Le premier terme qui compose notre fonction de coût $\mathcal{L}$ encourage l'appariement standard des intensités lumineuses entre la carte de probabilités $s(\theta)$ et la vérité terrain y. Nous le notons $\mathcal{F}_{DL}(s(\theta), y)$ et choisissons une fonction objectif classique en DL pour la segmentation d'images.

Ensuite, nous introduisons $\varphi : \bar{\Omega} \to \mathbb{R}^2$ la déformation recherchée permettant de transformer y en $s(\theta)$. En pratique, φ doit être à valeurs dans $\bar{\Omega}$ mais du point de vue mathématique, si nous travaillons avec de tels espaces, nous perdons la structure d'espace vectoriel. Les résultats de Ball [8, Théorèmes 1 et 2] permettent de contourner cette difficulté en travaillant classiquement sur des espaces fonctionnels à valeurs dans $\mathbb{R}^2$, tout en établissant que les déformations générées par le modèle sont à valeurs dans $\bar{\Omega}$.
D'après la définition de Ciarlet ([29]), une déformation est une fonction régulière qui préserve l'orientation et est injective, sauf éventuellement sur $\partial\Omega$ où l'auto-contact est autorisé. Le gradient de la déformation est $\nabla\varphi : \bar{\Omega} \to M_2(\mathbb{R})$, $M_2(\mathbb{R})$ représentant l'en-

semble des matrices réelles carrées d'ordre 2. Cela se traduit mathématiquement par la condition $\det \nabla\varphi > 0$ presque partout. Le champ de déplacement associé est noté u avec $\varphi = \mathrm{Id} + u$ et $\nabla\varphi = I_2 + \nabla u$, Id étant la fonction identité, et I_2 la matrice identité 2×2. Pour définir le modèle, nous avons aussi besoin des notations suivantes : $A : B = \mathrm{tr}\, A^T B$, définissant le produit scalaire usuel, et $\|A\| = \sqrt{A : A}$ la norme associée (norme de Frobenius).
La conception de la fonctionnelle à minimiser est dictée par plusieurs considérations, et apparaît comme un compromis entre la régularité de la déformation à construire et la qualité de l'appariement entre la vérité terrain déformée et la prédiction. Ainsi, la fonctionnelle comprend d'abord un terme F_{id} quantifiant la proximité de la vérité terrain déformée $y \circ \varphi$ de la prédiction $s(\theta)$, soit le degré d'alignement. Nous prenons classiquement la somme des différences au carré :

$$\mathrm{F}_{id}(\varphi) = \frac{1}{2} \|s(\theta) - y \circ \varphi\|^2_{L^2(\Omega)}. \tag{4.1}$$

Puis, la fonctionnelle inclut une régularisation agissant comme un modèle de déformation et fondée sur des concepts mécaniques [1]. Nous adoptons le point de vue de l'élasticité incompressible ([56]), ce qui motive l'introduction d'une première condition concernant le déterminant $\det \nabla\varphi$. En effet, l'hypothèse d'incompressibilité signifie que le matériau ne subit pas de modification de volume lorsqu'une déformation est appliquée. Du point de vue du problème de segmentation, la condition d'incompressibilité encode deux propriétés intéressantes : (i) une première de nature topologique, puisque la prédiction $s(\theta)$ doit exhiber le même nombre de composantes connexes que y, et (ii) une seconde qui assure la préservation de l'aire, ce qui mène indéniablement à des segmentations plus précises.
L'exigence d'incompressibilité peut alors être énoncée sous forme algébrique relativement à la déformation comme $\det \nabla\varphi = 1$ presque partout (*p.p.*) sur Ω. Cette condition locale d'incompressibilité est complétée par une régularisation plus classique qui peut être vue, du point de vue mécanique, comme une pénalisation de l'énergie élastique interne $\pi(u)$ du corps, typiquement formulée comme :

$$\pi(u) = \int_\Omega \sigma(x, \nabla u(x))\, dx. \tag{4.2}$$

La fonction σ correspond à la fonction de densité d'énergie, dépendant de la position x et du gradient de déplacement ∇u. Sa forme caractérise les propriétés du matériau constituant le corps et dicte le choix de l'espace fonctionnel sur lequel π est définie. Dans notre cas, nous proposons d'introduire la densité d'énergie suivante

$$W(\nabla\varphi) = W(I_2 + \nabla u) = \sigma(\nabla u) = \mu \left(\|\nabla\varphi\|^4 - 4 \right), \tag{4.3}$$

μ étant un paramètre de réglage, conduisant au terme de régularisation suivant

$$\mathrm{Reg}(\varphi) = \mu \int_\Omega \left(\|\nabla\varphi\|^4 - 4 \right) dx. \tag{4.4}$$

1. À noter que dans le contexte de la mécanique, l'objet à déformer (l'image de la vérité terrain dans notre contexte) est appelé corps et est soumis à des forces.

Le premier terme contrôle la régularité et pénalise les changements de longueur des contours. La constante 4 est ajoutée pour se conformer à la propriété $W(I_2) = 0$, I_2 étant le jacobien de la déformation identité. Comme nous le verrons, ce choix de régularisation conduit à travailler sur l'espace fonctionnel $W^{1,p}(\Omega, \mathbb{R}^2)$ avec $p = 4 > 2$ et permet, in fine, de démontrer le caractère homéomorphe de la déformation engendrée. Combinée avec la condition d'incompressibilité locale, la composante (4.4) exhibe de bonnes propriétés théoriques, utiles pour l'analyse mathématique conduite ci-après.

Le modèle de recalage proposé, dans un cadre variationnel, se décline finalement comme suit :

$$\inf_{\varphi \in \mathcal{W}} \Bigg\{ \mathcal{F}_{Rec}(\varphi) = \frac{\nu}{2} \mathrm{F}_{id}(\varphi) + \mathrm{Reg}(\varphi), \qquad (\mathcal{P}) \\ = \frac{\nu}{2} \|s(\theta) - y \circ \varphi\|^2_{L^2(\Omega)} + \int_\Omega W(\nabla \varphi)\, dx \Bigg\},$$

sous la contrainte $\det \nabla\varphi = 1$ *p.p.*, avec $\mathcal{W} = \{\psi \in \mathrm{Id} + W_0^{1,4}(\Omega, \mathbb{R}^2) \mid \det \nabla\psi = 1 \text{ } p.p.\}$.

Enfin, le modèle conjoint proposé combine les deux fonctionnelles $\mathcal{F}_{DL}$ et $\mathcal{F}_{Rec}$. Par conséquent, le problème de minimisation de la fonction de perte qui en résulte, se formule de la manière suivante :

$$\inf_{\theta, \varphi \in \mathcal{W}} \mathcal{L}(\theta, \varphi) = \mathcal{F}_{DL}(s(\theta), y) + \frac{\nu}{2} \|s(\theta) - y \circ \varphi\|^2_{L^2(\Omega)} + \int_\Omega W(\nabla \varphi)\, dx,$$

sous la contrainte $\det \nabla\varphi = 1$ *p.p.*.

Avant de détailler la résolution de ce problème d'optimisation, nous présentons des résultats théoriques associés au problème de recalage $(\mathcal{P})$ témoignant de son caractère bien posé.

4.2.2 Résultats théoriques

Dans cette section, nous présentons un premier résultat d'abord d'existence de minimiseurs pour le problème $(\mathcal{P})$ relatif à la transformation φ, puis un résultat asymptotique. Dans cette partie théorique, nous désignerons la fonctionnelle $\mathcal{F}_{Rec}$ simplement par $\mathcal{F}$ pour des raisons de lisibilité.

Nous énonçons d'abord le résultat d'existence de minimiseurs.

Théorème 4.2.1. *Le problème $(\mathcal{P})$ admet au moins un minimiseur dans $\mathcal{W}$.*

Démonstration. La preuve suit les arguments de la méthode classique directe du calcul des variations.

D'abord, nous avons clairement

$$\begin{aligned}\mathcal{F}(\varphi) &= \frac{\nu}{2}\left\|s(\theta) - y\circ\varphi\right\|^2_{L^2(\Omega)} + \mu\int_\Omega \left(\|\nabla\varphi\|^4 - 4\right)\,dx \\ &\geq \mu\left\|\nabla\varphi\right\|^4_{L^4(\Omega,M_2(\mathbb{R}))} - 4\mu\,\mathrm{meas}(\Omega),\end{aligned}$$

meas désignant la mesure de Lebesgue. La quantité $\mathcal{F}(\varphi)$ est donc bornée inférieurement par $-4\mu\,\mathrm{meas}(\Omega)$. De plus, la fonctionnelle $\mathcal{F}$ est propre. En effet, pour $\varphi = \mathrm{Id}$, $\mathcal{F}(\varphi) = \frac{\nu}{2}\left\|s(\theta) - y\right\|^2_{L^2(\Omega)}$ est finie en raison de l'injection $BV(\Omega) \subset L^2(\Omega)$, ou simplement par le fait que $L^\infty(\Omega) \subset L^2(\Omega)$.
Par conséquent, l'infimum de $\mathcal{F}$ est fini.

Considérons désormais une suite minimisante $(\varphi_k)_k \in \mathcal{W}$, *i.e.*, $\lim\limits_{k\to+\infty} \mathcal{F}(\varphi_k) = \inf\limits_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi)$. Il existe $K \in \mathbb{N}$ tel que $\forall k \in \mathbb{N}$,

$$\left(k \geq K \;\Rightarrow\; \mathcal{F}(\varphi_k) \leq \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + 1\right).$$

À partir de maintenant, nous supposons que $k \geq K$. D'après l'inégalité de coercivité, il vient que $(\varphi_k)_k$ est uniformément bornée par rapport à k dans $W^{1,4}(\Omega,\mathbb{R}^2)$, utilisant l'inégalité de Poincaré ([38, pp. 106-107] et le fait que $\varphi_k = \mathrm{Id}$ sur $\partial\Omega$. De plus, $\det\nabla\varphi_k = 1$ *p.p.* sur Ω de telle sorte que $\det\nabla\varphi_k$ est uniformément borné dans $L^\infty(\Omega)$. Aussi, il existe une sous-suite, toujours désignée par $(\varphi_k)_k$, et $\bar\varphi \in W^{1,4}(\Omega,\mathbb{R}^2)$ telles que

$$\varphi_k \underset{k\to+\infty}{\rightharpoonup} \bar\varphi \text{ dans } W^{1,4}(\Omega,\mathbb{R}^2).$$

En outre, nous savons que $\varphi_k \underset{k\to+\infty}{\to} \bar\varphi$ dans $\mathcal{C}^{0,w}(\bar\Omega,\mathbb{R}^2)$ en raison de l'injection compacte $W^{1,4}(\Omega,\mathbb{R}^2) \subset \mathcal{C}^{0,w}(\bar\Omega,\mathbb{R}^2)$ ([35, Théorème de Sobolev, Théorème 12.12]) pour tout $w < \frac{1}{2}$.

De plus, il existe une sous-suite (commune avec celle précédemment extraite, ce qui est toujours possible), toujours notée $(\det\nabla\varphi_k)_k$, et $\delta \in L^\infty(\Omega)$ telles que

$$\det\nabla\varphi_k \underset{k\to+\infty}{\overset{*}{\rightharpoonup}} \delta \text{ dans } L^\infty(\Omega),$$

signifiant que $\forall\phi \in L^1(\Omega)$,

$$\int_\Omega \det\nabla\varphi_k\,\phi\,dx \underset{k\to+\infty}{\to} \int_\Omega \delta\,\phi\,dx.$$

Comme Ω est borné, $L^2(\Omega) \subset L^1(\Omega)$ donc

$$\det\nabla\varphi_k \underset{k\to+\infty}{\rightharpoonup} \delta \text{ dans } L^2(\Omega).$$

En appliquant [35, Théorème 8.20] et par unicité de la limite faible dans $L^2(\Omega)$, nous déduisons que $\det \nabla\bar{\varphi} = \delta$ et $\det \nabla\varphi_k \underset{k\to+\infty}{\overset{*}{\rightharpoonup}} \det \nabla\bar{\varphi}$. Puis, $\det \nabla\bar{\varphi} = 1$ *p.p.*. Par continuité de l'opérateur de trace ([15, Théorème III.9]), nous obtenons que $\bar{\varphi} \in \mathrm{Id} + W_0^{1,4}(\Omega, \mathbb{R}^2)$.

Maintenant, pour tout $k \geq K$,

$$\int_\Omega \| \left(\nabla\varphi_k\right)^{-1} \|^4 \det \nabla\varphi_k \, dx = \int_\Omega \frac{1}{\left(\det \nabla\varphi_k\right)^3} \|\nabla\varphi_k\|^4 \, dx \leq C,$$

C étant une constante positive qui dépend seulement de Ω. Les hypothèses des théorèmes de Ball ([8, Théorèmes 1 et 2]) sont donc valides, ce qui implique que φ_k est un homéomorphisme de $\bar{\Omega}$ dans $\bar{\Omega}$ et $\varphi_k^{-1} \in W^{1,4}(\Omega, \mathbb{R}^2)$. La fonction φ_k fait correspondre des ensembles mesurables dans $\bar{\Omega}$ à des ensembles mesurables dans $\bar{\Omega}$, et la formule de changement de variable

$$\int_A f(u(x)) \det \nabla u(x) \, dx = \int_{u(A)} f(v) \, dv$$

est valable pour tout ensemble mesurable $A \subset \bar{\Omega}$ et toute fonction mesurable $f : \mathbb{R}^2 \to \mathbb{R}$, à condition que l'une des intégrales existe dans la relation précédente. La matrice des dérivées faibles de φ_k^{-1} est donnée par $\nabla\varphi_k^{-1} = \left(\nabla\varphi_k\right)^{-1} \circ \varphi_k^{-1}$ presque partout dans Ω. Le même raisonnement s'applique à $\bar{\varphi}$ qui hérite des mêmes propriétés lisses : c'est un homéomorphisme de $\bar{\Omega}$ dans $\bar{\Omega}$ et $\bar{\varphi}^{-1} \in W^{1,4}(\Omega, \mathbb{R}^2)$.

Il reste maintenant à établir la semi-continuité inférieure de $\mathcal{F}$. D'une part, la fonction de la densité d'énergie W est continue et convexe. Nous savons que si $\psi_k \underset{k\to+\infty}{\longrightarrow} \bar{\psi}$ dans $W^{1,4}(\Omega, \mathbb{R}^2)$, alors $\nabla\psi_k \underset{k\to+\infty}{\longrightarrow} \nabla\bar{\psi}$ dans $L^4(\Omega, M_2(\mathbb{R}))$ et nous pouvons extraire une suite telle que $\nabla\psi_k \underset{k\to+\infty}{\longrightarrow} \nabla\bar{\psi}$ presque partout dans Ω. Par continuité de W, il vient donc que $W(\nabla\psi_k) \underset{k\to+\infty}{\longrightarrow} W(\nabla\bar{\psi})$ presque partout dans Ω. L'application du lemme de Fatou donne alors

$$\int_\Omega W(\nabla\bar{\psi}) \, dx \leq \liminf_{k\to+\infty} \int_\Omega W(\nabla\psi_k) \, dx.$$

Comme W est convexe, nous pouvons appliquer [15, Corollaire III.8] de sorte que

$$\int_\Omega W(\nabla\bar{\varphi}) \, dx \leq \liminf_{k\to+\infty} \int_\Omega W(\nabla\varphi_k) \, dx.$$

Pour conclure, il nous faut étudier $\liminf_{k\to+\infty} \|s(\theta) - y \circ \varphi_k\|^2_{L^2(\Omega)}$ (en fait, nous travaillons avec $\lim_{k\to+\infty} \|s(\theta) - y \circ \varphi_k\|^2_{L^2(\Omega)}$) en nous inspirant des travaux antérieurs de Wirth ([140]). Dans un premier temps, nous prouvons que $\varphi_k \circ \bar{\varphi}^{-1} \underset{k\to+\infty}{\longrightarrow} \mathrm{Id}$ dans $\mathcal{C}^{0,\alpha}(\bar{\Omega}, \mathbb{R}^2)$ avec $\alpha < \frac{1}{2}$.

Rappelons que, $W^{1,4}(\Omega,\mathbb{R}^2) \subset \mathcal{C}^{0,\lambda}(\bar{\Omega},\mathbb{R}^2)$ ([35, Théorème de Sobolev, Théorème 12.11]) pour tout $\lambda \in [0,\frac{1}{2}]$ et l'injection est compacte pour tout $0 \leq \lambda < \frac{1}{2}$ ([35, Théorème de Rellich-Kondrachov, Théorème 12.12]). Prenons alors (α,λ) deux nombres réels positifs tels que $\left\{ \begin{array}{ccc} \lambda & < & \frac{1}{2} \\ \frac{\alpha}{\lambda} & < & \frac{1}{2} \end{array} \right.$, ce qui est toujours possible. Des calculs simples permettent d'obtenir que

$$
\begin{aligned}
&\|\varphi_k \circ \bar{\varphi}^{-1} - \bar{\varphi} \circ \bar{\varphi}^{-1}\|_{\mathcal{C}^{0,\alpha}(\bar{\Omega},\mathbb{R}^2)} \leq \sup_{x\in\bar{\Omega}} |\varphi_k(x) - \bar{\varphi}(x)| \\
&+ \sup_{\substack{(x,y)\in\bar{\Omega}\times\bar{\Omega} \\ x\neq y}} \frac{|\varphi_k \circ \bar{\varphi}^{-1}(x) - \varphi_k \circ \bar{\varphi}^{-1}(y) - \bar{\varphi} \circ \bar{\varphi}^{-1}(x) + \bar{\varphi} \circ \bar{\varphi}^{-1}(y)|}{|\bar{\varphi}^{-1}(x) - \bar{\varphi}^{-1}(y)|^{\lambda}} \frac{|\bar{\varphi}^{-1}(x) - \bar{\varphi}^{-1}(y)|^{\lambda}}{|x-y|^{\alpha}}, \\
&\leq \sup_{x\in\bar{\Omega}} |\varphi_k(x) - \bar{\varphi}(x)| + \sup_{\substack{(x,y)\in\bar{\Omega}\times\bar{\Omega} \\ x\neq y}} \frac{|\varphi_k(x) - \varphi_k(y) - \bar{\varphi}(x) + \bar{\varphi}(y)|}{|x-y|^{\lambda}} \sup_{\substack{(x,y)\in\bar{\Omega}\times\bar{\Omega} \\ x\neq y}} \frac{|\bar{\varphi}^{-1}(x) - \bar{\varphi}^{-1}(y)|^{\lambda}}{|x-y|^{\alpha}}, \\
&\leq \sup_{x\in\bar{\Omega}} |\varphi_k(x) - \bar{\varphi}(x)| + \sup_{\substack{(x,y)\in\bar{\Omega}\times\bar{\Omega} \\ x\neq y}} \frac{|\varphi_k(x) - \varphi_k(y) - \bar{\varphi}(x) + \bar{\varphi}(y)|}{|x-y|^{\lambda}} \sup_{\substack{(x,y)\in\bar{\Omega}\times\bar{\Omega} \\ x\neq y}} \left| \frac{\bar{\varphi}^{-1}(x) - \bar{\varphi}^{-1}(y)}{|x-y|^{\frac{\alpha}{\lambda}}} \right|^{\lambda}, \\
&\leq (1 + \|\bar{\varphi}^{-1}\|^{\lambda}_{\mathcal{C}^{0,\frac{\alpha}{\lambda}}(\bar{\Omega},\mathbb{R}^2)}) \|\varphi_k - \bar{\varphi}\|_{\mathcal{C}^{0,\lambda}(\bar{\Omega},\mathbb{R}^2)} \underset{k\to+\infty}{\longrightarrow} 0.
\end{aligned}
$$

Pour la suite, remarquons d'abord que $y \circ \varphi_k \in L^2(\Omega)$, et de même pour $y \circ \bar{\varphi}$. En effet, d'après la formule du changement de variable, valable eu égard aux résultats de Ball, et de l'inégalité de Hölder avec le paramètre $s \geq 2$,

$$
\begin{aligned}
\int_{\Omega} |y \circ \varphi_k|^2 \, dx &= \int_{\Omega} |y|^2 \, \frac{1}{\det \nabla\varphi_k \circ (\varphi_k)^{-1}} \, dx, \\
&\leq L^2 \int_{\Omega} 1 \times \frac{1}{\det \nabla\varphi_k \circ (\varphi_k)^{-1}} \, dx, \\
&\leq L^2 \left(\int_{\Omega} dx \right)^{\frac{s-1}{s}} \left(\int_{\Omega} \left(\frac{1}{\det \nabla\varphi_k \circ (\varphi_k)^{-1}} \right)^{s} dx \right)^{\frac{1}{s}}, \\
&\leq L^2 \, (\mathrm{meas}(\Omega))^{\frac{s-1}{s}} \left(\int_{\Omega} \left(\frac{1}{\det \nabla\varphi_k} \right)^{s-1} dx \right)^{\frac{1}{s}} < +\infty,
\end{aligned}
$$

la quantité à droite étant uniformément bornée en k, ayant $y \in L^{\infty}(\Omega)$ et $\det \nabla\varphi_k = 1$ presque partout. De plus, soit $\mathcal{N}_k \subset \Omega$ un ensemble tel que $\mathrm{meas}(\mathcal{N}_k) = 0$ et $\forall x \in \Omega \setminus \mathcal{N}_k$, $\det \nabla\varphi_k = 1$. Posons $\mathcal{N}'_k = \varphi_k(\mathcal{N}_k)$. Puisque $\varphi_k \in W^{1,p}(\Omega,\mathbb{R}^2)$ avec $p = 4 > 2$, dimension de Ω, $\mathrm{meas}(\mathcal{N}'_k) = 0$ d'après le corollaire 1 de [91, Théorème 1].
$\forall y \notin \mathcal{N}'_k$,

$$\det \nabla\varphi_k \circ \varphi_k^{-1}(y) = \det \nabla\varphi_k(x),$$

avec $x \in \Omega \setminus \mathcal{N}_k$, ce qui conduit à $\det \nabla\varphi_k \circ \varphi_k^{-1} = 1$ *p.p.* et finalement, $\det \nabla\varphi_k^{-1} = 1$ *p.p.*. Cette propriété permet en particulier d'appliquer le Théorème 1 de [8] à φ_k^{-1}. Un

raisonnement similaire s'applique pour $y \circ \bar{\varphi}$.

Soit maintenant l'ensemble $\mathcal{O}_i = \{x \in \Omega \,|\, y(x) = i\}$ avec i un indice quelconque dans $\{1, \cdots, L\}$. Nous observons que $\Omega = \cup_{i=1}^{L} \mathcal{O}_i \cup \mathcal{N}$ avec $\mathcal{N}$ tel que $\text{meas}(\mathcal{N}) = 0$.
Considérons alors un ensemble mesurable $\mathcal{S} \subset \Omega$ de mesure de Lebesgue satisfaisant $\text{meas}(\mathcal{S}) \leq \eta$.

Nous établissons, en procédant comme précédemment, que

$$\begin{aligned}
\int_{\bar{\varphi}^{-1}(\mathcal{S})} dx &= \int_{\mathcal{S}} 1 \times \frac{1}{\det \nabla\bar{\varphi} \circ (\bar{\varphi})^{-1}}\, dx, \\
&\leq \left(\int_{\mathcal{S}} dx\right)^{\frac{s-1}{s}} \left(\int_{\mathcal{S}} \left(\frac{1}{\det \nabla\bar{\varphi} \circ (\bar{\varphi})^{-1}}\right)^{s} dx\right)^{\frac{1}{s}}, \\
&\leq (\text{meas}(\mathcal{S}))^{\frac{s-1}{s}} \left(\int_{\Omega} \left(\frac{1}{\det \nabla\bar{\varphi}}\right)^{s-1} dx\right)^{\frac{1}{s}}, \\
&\leq (\text{meas}(\Omega))^{\frac{1}{s}} \, (\text{meas}(\mathcal{S}))^{\frac{s-1}{s}}, \qquad (4.5)
\end{aligned}$$

où nous avons exploité les propriétés de $\bar{\varphi}$ et en particulier le fait que $\det \nabla\bar{\varphi} = 1$ presque partout. Encore une fois, en raison des propriétés de régularité de $\bar{\varphi}$, $\Omega = \bar{\varphi}^{-1}(\Omega) = \cup_{i=1}^{L} \bar{\varphi}^{-1}(\mathcal{O}_i) \cup \bar{\varphi}^{-1}(\mathcal{N})$, ce qui combiné avec l'inégalité (4.5), montre que pour achever la preuve, il suffit que démontrer que $y \circ \varphi_k \underset{k \to +\infty}{\longrightarrow} y \circ \bar{\varphi}$ dans $L^2(\bar{\varphi}^{-1}(\mathcal{O}_i))$.

Aussi,

$$\begin{aligned}
\|y \circ \varphi_k - y \circ \bar{\varphi}\|^2_{L^2(\bar{\varphi}^{-1}(\mathcal{O}_i))} &= \int_{\bar{\varphi}^{-1}(\mathcal{O}_i)} |y \circ \varphi_k - y \circ \bar{\varphi}|^2 \, dx = \int_{\bar{\varphi}^{-1}(\mathcal{O}_i)} |y \circ \varphi_k - i|^2 \, dx, \\
&= \int_{\varphi_k \circ \bar{\varphi}^{-1}(\mathcal{O}_i)} |y - i|^2 \, \det \nabla \left(\varphi_k^{-1}\right) dx, \\
&\leq \int_{\varphi_k \circ \bar{\varphi}^{-1}(\mathcal{O}_i) \Delta \mathcal{O}_i} |y - i|^2 \, \det \nabla \left(\varphi_k^{-1}\right) dx + \int_{\mathcal{O}_i} |y - i|^2 \, \det \nabla \left(\varphi_k^{-1}\right) dx, \\
&\leq (L-1)^2 \int_{\varphi_k \circ \bar{\varphi}^{-1}(\mathcal{O}_i) \Delta \mathcal{O}_i} \det \nabla \left(\varphi_k^{-1}\right) dx,
\end{aligned}$$

où Δ représente la différence symétrique. D'après la formule du changement de variable, valable à partir des résultats de Ball, et de l'inégalité de Hölder avec le paramètre $s \geq 2$,

$$\begin{aligned}
\|y \circ \varphi_k - y \circ \bar{\varphi}\|^2_{L^2(\bar{\varphi}^{-1}(\mathcal{O}_i))} &\leq (L-1)^2 \left(\int_{\varphi_k \circ \bar{\varphi}^{-1}(\mathcal{O}_i) \Delta \mathcal{O}_i} dx\right)^{\frac{s-1}{s}} \left(\int_{\varphi_k \circ \bar{\varphi}^{-1}(\mathcal{O}_i) \Delta \mathcal{O}_i} \left(\frac{1}{\det \nabla\varphi_k}\right)^{s-1} dx\right)^{\frac{1}{s}}, \\
&\leq (L-1)^2 \left(\text{meas}\left(\varphi_k \circ \bar{\varphi}^{-1}(\mathcal{O}_i) \, \Delta \mathcal{O}_i\right)\right)^{\frac{s-1}{s}} \left(\int_{\Omega} \left(\frac{1}{\det \nabla\varphi_k}\right)^{s-1} dx\right)^{\frac{1}{s}}.
\end{aligned}$$

Le terme le plus à droite est uniformément borné par rapport à k puisque $\det \nabla\varphi_k = 1$ *p.p.*. Par ailleurs, comme $\varphi_k \circ \bar{\varphi}^{-1} \underset{k\to+\infty}{\longrightarrow} \mathrm{Id}$ dans $\mathcal{C}^{0,\lambda}(\bar{\Omega}, \mathbb{R}^2)$, il peut être démontré (voir [140, Footnote 1., p.455]) que $\mathrm{meas}\left(\varphi_k \circ \bar{\varphi}^{-1}\left(\mathcal{O}_i\right) \Delta \mathcal{O}_i\right) \underset{k\to+\infty}{\longrightarrow} 0$, menant à $y \circ \varphi_k \underset{k\to+\infty}{\longrightarrow} y \circ \bar{\varphi}$ dans $L^2(\bar{\varphi}^{-1}(\mathcal{O}_i))$.

Cela permet de conclure que le problème $(\mathcal{P})$ admet au moins une solution $\bar{\varphi} \in \mathcal{W}$. □

Avant d'énoncer notre résultat asymptotique, nous donnons un résultat préliminaire dédié à l'approximation locale par une fonction régulière de notre y spécifique.

Théorème 4.2.2. *Approximation par des fonctions régulières*

Nous supposons que $y \in BV(\Omega, \{0, \cdots, L-1\}) \subset L^\infty(\Omega)$ (un décalage s'opère dans les valeurs prises par y) —0 représentant le fond —de telle sorte que son support essentiel $\mathrm{ess\,supp}\,(y)$, qui est le plus petit ensemble fermé K de Ω tel que $y = 0$ presque partout, est inclus dans $\Omega' \subset\subset \Omega$, Ω' étant un ensemble ouvert borné de Ω. Soit $p > 2$. Alors il existe des fonctions $\{y_k\}_{k=1}^\infty \subset BV(\Omega) \cap \mathcal{C}^\infty(\Omega)$ telles que

(i) $y_k \underset{k\to+\infty}{\longrightarrow} y$ dans $L^p(\Omega)$,
(ii) $\|Dy_k\|(\Omega) \underset{k\to+\infty}{\longrightarrow} \|Dy\|(\Omega)$.

De plus, il existe un ensemble compact $\widehat{K} \subset \Omega$ tel que $\forall k \in \mathbb{N}^*$, $supp(y_k) \subset \widehat{K}$ et les fonctions y_k sont uniformément bornées en norme L^∞.

Démonstration. La preuve est une adaptation de celle de [47, Théorème 2, Chapitre 5]. Nous en suivons donc les étapes et soulignons les différences principales sans toutefois nous attarder sur les éléments communs. En particulier, nous renvoyons le lecteur à [47, Théorème 2, Chapitre 5] pour la preuve de (ii).

Fixons $\varepsilon > 0$. Soit m un entier positif. Nous définissons les ensembles ouverts

$$\Omega_k = \left\{ x \in \Omega \,|\, \mathrm{dist}(x, \partial\Omega) > \frac{1}{m+k} \right\} \cap B(0, k+m),\ k = 1, \cdots,$$

$B(x, r)$ désignant la boule ouvert de centre x et de rayon r, et m étant choisi assez grand pour que $\|Dy\|(\Omega - \Omega_1) < \varepsilon$.
Nous posons $\Omega_0 \equiv \emptyset$ et définissons $V_k = \Omega_{k+1} \setminus \overline{\Omega_{k-1}}$, $k = 1, \cdots$. Soit $\{\zeta_k\}_{k=1}^\infty$ une suite de fonctions régulières telles que

$$\begin{cases} \zeta_k \in \mathcal{C}_c^\infty(V_k) \quad 0 \le \zeta_k \le 1 \quad (k = 1, \cdots) \\ \sum_{k=1}^\infty \zeta_k = 1 \qquad\qquad\qquad\qquad \text{sur } \Omega. \end{cases}$$

Soit $\rho \in \mathcal{C}_c^\infty(\mathbb{R}^2)$ avec $\operatorname{supp}(\rho) \subset B(0,1)$, $\rho \geq 0$ et $\int_{\mathbb{R}^2} \rho(x)\,dx = 1$, et nous définissons ensuite $\rho_\varepsilon(x) = \frac{1}{\varepsilon^2}\,\rho\left(\frac{x}{\varepsilon}\right)$. Pour chaque k, nous sélectionnons $\varepsilon_k > 0$ petit de sorte que

$$\begin{cases} \operatorname{supp}\left(\rho_{\varepsilon_k} * (y\zeta_k)\right) \subset V_k, \\ \left(\int_\Omega \left|\rho_{\varepsilon_k} * (y\zeta_k) - y\zeta_k\right|^p dx\right)^{\frac{1}{p}} < \dfrac{\varepsilon}{2^k}, \\ \int_\Omega \left|\rho_{\varepsilon_k} * (yD\zeta_k) - yD\zeta_k\right| dx < \dfrac{\varepsilon}{2^k}. \end{cases}$$

Nous définissons maintenant $y_\varepsilon = \sum_{k=1}^{\infty} \rho_{\varepsilon_k} * (y\zeta_k)$. Par construction des ensembles $\{\Omega_k\}_k$, il existe n_0 tel que $\operatorname{ess\,supp}(y) \subset \Omega' \subset\subset \Omega_{n_0}$. La quantité $\sum_{k=1}^{\infty} \rho_{\varepsilon_k} * (y\zeta_k)$ peut donc se restreindre aux indices $\sum_{k=1}^{n_0} \rho_{\varepsilon_k} * (y\zeta_k)$, aboutissant à $y_\varepsilon = \sum_{k=1}^{n_0} \rho_{\varepsilon_k} * (y\zeta_k) \in \mathcal{C}^\infty(\Omega)$ avec $\operatorname{supp}(y_\varepsilon) \subset \Omega_{n_0+1}$, ensemble qui dépend de ε. À noter que prendre $\tilde{\varepsilon} \leq \varepsilon$ générerait un autre partitionnement $\tilde{\Omega}_k$ que l'on peut obtenir en réétiquetant le partitionnement Ω_k. Puisque $y = \sum_{k=1}^{n_0} y\zeta_k$, en utilisant l'inégalité de Minkowski, il vient

$$\begin{aligned} \|y_\varepsilon - y\|_{L^p(\Omega)} &= \left(\int_\Omega \left|\sum_{k=1}^{n_0} \rho_{\varepsilon_k} * (y\zeta_k) - y\zeta_k\right|^p dx\right)^{\frac{1}{p}}, \\ &\leq \sum_{k=1}^{n_0} \left(\int_\Omega \left|\rho_{\varepsilon_k} * (y\zeta_k) - y\zeta_k\right|^p dx\right)^{\frac{1}{p}} < \varepsilon. \end{aligned}$$

Par conséquent, $y_\varepsilon \longrightarrow y$ dans $L^p(\Omega)$ quand ε tend vers 0 (et en particulier, dans $L^1(\Omega)$). Comme $\|Dy\|(\Omega) \leq \liminf_{\varepsilon\to 0} \|Dy_\varepsilon(\Omega)\|$, il suffit de prouver que $\limsup_{\varepsilon\to 0} \|Dy_\varepsilon(\Omega)\| \leq \|Dy\|(\Omega)$, nous renvoyons le lecteur à suivre [47, Théorème 2, Chapitre 5].

Maintenant, si $x \in \Omega_1 \subset\subset \Omega_2$ alors

$$\begin{aligned} y_\varepsilon(x) &= \rho_{\varepsilon_1} * (y\zeta_1), \\ &= \int_{V_1=\Omega_2} \rho_{\varepsilon_1}(x-s)\, y(s)\zeta_1(s)\, ds, \end{aligned}$$

menant à $|y_\varepsilon(x)| \leq (L-1) \int \rho_{\varepsilon_1}(x-s)\,ds = L-1$.

Si $x \in \Omega_l \setminus \overline{\Omega_{l-1}}$ avec $l > 1$, alors

$$\begin{aligned} y_\varepsilon(x) &= \rho_{\varepsilon_{l-1}} * (y\zeta_{l-1}) + \rho_{\varepsilon_l} * (y\zeta_l), \\ &= \int_{V_{l-1}} \rho_{\varepsilon_{l-1}}(x-s)\, y(s)\zeta_{l-1}(s)\, ds + \int_{V_l} \rho_{\varepsilon_l}(x-s)\, y(s)\zeta_l(s)\, ds, \\ &\leq 2(L-1), \end{aligned}$$

montrant que y_k est uniformément bornée en norme L^∞ (la borne dépend uniquement de L qui est fixé). □

Pour des raisons pratiques, nous considérons l'extension par 0 de y, désignée $Ey = \left| \begin{array}{ll} y & \text{sur} \quad \Omega \\ 0 & \text{sur} \quad \mathbb{R}^2 \setminus \bar{\Omega} \end{array} \right.$, afin que Ey soit appréhendée comme un élément de $BV(\mathbb{R}^2, \{0, \cdots, L-1\})$. À noter que $\|D(Ey)\|(\mathbb{R}^2) = \|Dy\|(\Omega)$. Le même processus d'extension est appliqué aux $\{y_k\}_k$ qui sont donc des éléments de $\mathcal{C}_c^\infty(\mathbb{R}^2)$ à support inclus dans un compact $\widehat{K}$ fixé.

Rappelons que le problème original $(\mathcal{P})$, pour lequel nous avons fourni un résultat d'existence de minimiseurs, est défini par

$$\inf_{\varphi \in \mathcal{W}} \Big\{ \mathcal{F}(\varphi) = \frac{\nu}{2} \operatorname{Fid}(\varphi) + \operatorname{Reg}(\varphi), \qquad (\mathcal{P})$$
$$= \frac{\nu}{2} \|s(\theta) - y \circ \varphi\|^2_{L^2(\Omega)} + \int_\Omega W(\nabla \varphi)\, dx \Big\},$$

sous la contrainte $\det \nabla \varphi = 1$ *p.p.*, avec $\mathcal{W} = \{\psi \in \mathrm{Id} + W_0^{1,4}(\Omega, \mathbb{R}^2) \mid \det \nabla \psi = 1 \text{ p.p.}\}$.

Le problème $(\mathcal{P})$ exhibe des difficultés numériques relatives à la non-linéarité et la non-convexité. Communément, en élasticité non linéaire, l'idée consiste à introduire des variables auxiliaires dans le but de fragmenter le problème initial en une séquence de problèmes plus facilement résolubles. Dans cet esprit, inspirés par les travaux pionniers de Negrón Marrero ([100]), nous introduisons deux variables auxiliaires V et W simulant toutes les deux $\nabla \varphi$ et le couplage s'opère par le biais d'une pénalisation L^2. L'idée sous-jacente consiste à transférer la non-linéarité de la régularisation sur V tandis que la contrainte dure est reportée sur W de sorte que $\det W = 1$ *p.p.*. Le processus décompose donc le problème en séparant la question de la régularité de celle de la préservation d'aire, ce qui donne, dans le cadre discret, une série de problèmes avec, pour la plupart, des solutions *closed-form*.

Maintenant, soit $l \in \mathbb{N}^*$. Par définition de l'infimum, il existe $\varphi_l \in \mathcal{W}$ tel que

$$\mathcal{F}(\varphi_l) \leq \inf_{\varphi \in \mathcal{W}} \mathcal{F}(\varphi) + \frac{1}{l}.$$

De plus, d'après le Théorème 4.2.2, il existe un rang $N_l \in \mathbb{N}^*$ tel que $\forall k \in \mathbb{N}^*$,

$$\left(k \geq N_l \implies \|y_k - y\|_{L^p(\Omega)} \leq \frac{1}{l} \right),$$

$p \in \mathbb{N}^*$ étant strictement supérieur à 2. À partir de maintenant, nous posons $k = N_l$ et pour une meilleure lisibilité, nous notons $y_l := y_{N_l}$. Grâce aux théorèmes de Ball

([8, Théorèmes 1 et 2]), φ_l est un homéomorphisme de $\bar{\Omega}$ dans $\bar{\Omega}$ et $\varphi_l^{-1} \in W^{1,4}(\Omega, \mathbb{R}^2)$. La fonction φ_l fait correspondre des ensembles mesurables dans $\bar{\Omega}$ à des ensembles mesurables dans $\bar{\Omega}$, et le changement de variable classique se vérifie. Ainsi (—en remarquant que $y_l \circ \varphi_l$, y_l sont régulières et bornés, et que $y \circ \varphi_l \in L^2(\Omega)$ d'après ce qui précède —), on a :

$$\begin{aligned}\left| \|y_l \circ \varphi_l - s(\theta)\|^2_{L^2(\Omega)} - \|y \circ \varphi_l - s(\theta)\|^2_{L^2(\Omega)} \right| &\leq \left| \int_\Omega (y_l \circ \varphi_l - y \circ \varphi_l)(y_l \circ \varphi_l + y \circ \varphi_l - 2s(\theta))\, dx \right|, \\ &\leq \|y_l \circ \varphi_l - y \circ \varphi_l\|_{L^2(\Omega)} \, \|y_l \circ \varphi_l + y \circ \varphi_l - 2s(\theta)\|_{L^2(\Omega)}\end{aligned}$$

eu égard à l'inégalité de Cauchy-Schwarz. La composante la plus à droite est bornée uniformément par rapport à l. De plus, en appliquant le changement de variable classique, valable d'après les résultats de Ball, et en utilisant l'inégalité de Hölder

$$\begin{aligned}\|y_l \circ \varphi_l - y \circ \varphi_l\|^2_{L^2(\Omega)} &= \int_\Omega |y_l - y|^2 \, \frac{1}{\det \nabla \varphi_l \circ (\varphi_l)^{-1}} \, dx, \\ &\leq \|y_l - y\|^2_{L^p(\Omega)} \left(\int_\Omega \left(\frac{1}{\det \nabla \varphi_l \circ (\varphi_l)^{-1}} \right)^{\frac{p}{p-2}} dx \right)^{\frac{p-2}{p}}.\end{aligned}$$

En vertu du théorème de Ball, une fois de plus, il s'ensuit que

$$\|y_l \circ \varphi_l - y \circ \varphi_l\|^2_{L^2(\Omega)} \leq \|y_l - y\|^2_{L^p(\Omega)} \left(\int_\Omega \left(\frac{1}{\det \nabla \varphi_l} \right)^{\frac{2}{p-2}} dx \right)^{\frac{p-2}{p}},$$

la partie la plus à droite étant uniformément bornée puisque $\det \nabla \varphi_l = 1$ *p.p.*. En rassemblant les résultats précédents, il vient finalement

$$\|y_l \circ \varphi_l - s(\theta)\|^2_{L^2(\Omega)} \leq \|y \circ \varphi_l - s(\theta)\|^2_{L^2(\Omega)} + \frac{C}{l},$$

avec $C > 0$ une constante dépendant seulement de Ω, L et p.

Munis de ces éléments, notre approche de *splitting* est introduite, fondée sur la famille de fonctionnelles $\{E_{l,j}\}_{l,j}$ définie ci-dessous, $(\gamma_j)_j$ étant une suite croissante de nombres réels positifs tels que $\lim_{j \to +\infty} \gamma_j = +\infty$. Précisément,

$$\begin{aligned}E_{l,j}(\varphi, V, W) = &\mu \|V\|^4_{L^4(\Omega, M_2(\mathbb{R}))} + \frac{\nu}{2} \|s(\theta) - (Ey_l) \circ \varphi\|^2_{L^2(\Omega)} \\ &+ \frac{\mu\alpha}{2} \|\det V - 1\|^2_{L^2(\Omega)} + \frac{\gamma_j}{2} \|V - \nabla\varphi\|^2 \qquad (DP_{l,j}) \\ &+ \frac{\gamma_j}{2} \|W - \nabla\varphi\|^2,\end{aligned}$$

avec $(\varphi, V, W) \in \left(\overline{\mathcal{W}} = \mathrm{Id} + W^{1,2}_0(\Omega, \mathbb{R}^2)\right) \times L^4(\Omega, M_2(\mathbb{R})) \times \left(\overline{\overline{\mathcal{W}}} = \{X \in L^2(\Omega, M_2(\mathbb{R})) \mid \det X = 1 \text{ p.p.}\}\right)$.

Le résultat qui suit revêt un intérêt théorique. Il permet en particulier de garantir un sens à l'expression $(Ey_l) \circ \varphi$ alors même que φ n'est que $W^{1,2}(\Omega, \mathbb{R}^2)$ ici. En pratique, nous avons implémenté le modèle avec le terme d'attache aux données $y \circ \varphi$. Pour $y \in BV(\Omega, \{0, \dots, L-1\})$ et $\varphi \in \mathrm{Id} + W^{1,2}(\Omega, \mathbb{R}^2)$, il n'est pas clair que ce terme soit correctement défini d'un point de vue théorique.

Théorème 4.2.3. *Résultat asymptotique*

Soit (y_l) la suite définie précédemment.
Soit $(\varphi_{l,j,k}, V_{l,j,k}, W_{l,j,k}) \in \overline{\mathcal{W}} \times L^4(\Omega, M_2(\mathbb{R})) \times \overline{\overline{\mathcal{W}}}$ une suite minimisante du problème $(DP_{l,j})$.
Alors il existe une sous-suite désignée par $\left(\varphi_{l,\Psi(j),N(l,\Psi(j))}, V_{l,\Psi(j),N(l,\Psi(j))}, W_{l,\Psi(j),N(l,\Psi(j))}\right)$ telle que

$$\lim_{l\to+\infty} \lim_{j\to+\infty} E_{l,\Psi(j)}\left(\varphi_{l,\Psi(j),N(l,\Psi(j))}, V_{l,\Psi(j),N(l,\Psi(j))}, W_{l,\Psi(j),N(l,\Psi(j))}\right) = \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi)$$

Démonstration. En premier lieu, nous observons qu'avec φ_l définie ci-dessus,

$$\inf_{\substack{(\varphi,V,W)\in\overline{\mathcal{W}}\times L^4(\Omega,M_2(\mathbb{R}))\\ \times\overline{\overline{\mathcal{W}}}}} E_{l,j}(\varphi, V, W) \leq E_{l,j}(\varphi_l, \nabla\varphi_l, \nabla\varphi_l) \leq \mathcal{F}(\varphi_l) + \frac{C}{l} \leq \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + \frac{C+1}{l}.$$

$E_{l,j}(\varphi, V, W) \geq 0$ et en prenant $\varphi = \mathrm{Id}$, $V = I_2$ et $W = I_2$ nous montrons que la fonctionnelle est propre puisque $s(\theta) \in L^2(\Omega)$ et $y_l \in \mathcal{C}^\infty(\Omega)$ (uniformément borné par rapport à l), et par conséquent, que l'infimum est fini.

Désignons par $(\varphi_{l,j,k}, V_{l,j,k}, W_{l,j,k})$ une suite minimisante du problème découplé $(DP)_{l,j}$, c'est-à-dire,

$$\lim_{k\to+\infty} E_{l,j}(\varphi_{l,j,k}, V_{l,j,k}, W_{l,j,k}) = \inf_{\substack{(\varphi,V,W)\in\overline{\mathcal{W}}\times L^4(\Omega,M_2(\mathbb{R}))\\ \times\overline{\overline{\mathcal{W}}}}} E_{l,j}(\varphi, V, W).$$

En particulier,

$$\forall \varepsilon > 0,\ \exists N_{\varepsilon,l,j} \in \mathbb{N},\ \forall k \in \mathbb{N},$$
$$\left(k \geq N_{\varepsilon,l,j} \implies E_{l,j}(\varphi_{l,j,k}, V_{l,j,k}, W_{l,j,k}) \leq \inf_{\substack{(\varphi,V,W)\in\overline{\mathcal{W}}\times L^4(\Omega,M_2(\mathbb{R}))\\ \times\overline{\overline{\mathcal{W}}}}} E_{l,j}(\varphi, V, W) + \varepsilon \right).$$

Prenons en particulier $\varepsilon = \dfrac{1}{\gamma_j}$. Alors il existe $N_{l,j} \in \mathbb{N}$ tel que $\forall k \in \mathbb{N}$,

$$\begin{aligned}\Bigg(k \geq N_{l,j} \implies E_{l,j}(\varphi_{l,j,k}, V_{l,j,k}, W_{l,j,k}) &\leq \inf_{\substack{(\varphi,V,W)\in \overline{\mathcal{W}}\times L^4(\Omega,M_2(\mathbb{R}))\\ \times\overline{\overline{\mathcal{W}}}}} E_{l,j}(\varphi, V, W) + \frac{1}{\gamma_j},\\ &\leq \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + \frac{C+1}{l} + \frac{1}{\gamma_j}, \qquad (4.6)\\ &\leq \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + \frac{C+1}{l} + \frac{1}{\gamma_0}\Bigg).\end{aligned}$$

À présent, nous définissons $k := N_{l,j}$ et à des fins de lisibilité, nous désignons par $\varphi_{l,j} := \varphi_{l,j,N_{l,j}}$, similairement pour $V_{l,j}$ et $W_{l,j}$. En utilisant l'inégalité $(a-b)^2 \geq \frac{1}{2}a^2 - b^2$, une première inégalité de type coercivité peut être obtenue, de la forme

$$\mu \left\|V_{l,j}\right\|^4_{L^4(\Omega,M_2(\mathbb{R}))} + \frac{\mu\alpha}{4} \left\|\det V_{l,j}\right\|^2_{L^2(\Omega)} - \frac{\mu\alpha}{2}\operatorname{meas}(\Omega) \leq \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + \frac{C+1}{l} + \frac{1}{\gamma_0},$$

montrant que

$$\left|\begin{array}{cc} & (V_{l,j}) \text{ est uniformément bornée dans } L^4(\Omega, M_2(\mathbb{R})) \text{ et donc dans } L^2(\Omega, M_2(\mathbb{R})),\\ \text{et} & (\det V_{l,j}) \text{ est uniformément borné dans } L^2(\Omega).\end{array}\right.$$

Il existe donc une sous-suite commune, obtenue en appliquant une procédure d'extraction diagonale, désignée par $(V_{l,\Psi(j)})$ (respectivement $(\det V_{l,\Psi(j)})$) ainsi que $\bar{V}_l \in L^4(\Omega, M_2(\mathbb{R}))$ et $\bar{\delta}_l \in L^2(\Omega)$ tels que

$$\begin{aligned} V_{l,\Psi(j)} &\underset{j\to+\infty}{\rightharpoonup} \bar{V}_l \text{ dans } L^4(\Omega, M_2(\mathbb{R})),\\ \det V_{l,\Psi(j)} &\underset{j\to+\infty}{\rightharpoonup} \bar{\delta}_l \text{ dans } L^2(\Omega).\end{aligned}$$

De plus,

$$\frac{\gamma_{\Psi(j)}}{2} \left\|V_{l,\Psi(j)} - \nabla\varphi_{l,\Psi(j)}\right\|^2_{L^2(\Omega,M_2(\mathbb{R}))} \leq \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + \frac{C+1}{l} + \frac{1}{\gamma_0},$$

conduisant à

$$\begin{aligned}\Big|\left\|\nabla\varphi_{l,\Psi(j)}\right\|_{L^2(\Omega,M_2(\mathbb{R}))} - \left\|V_{l,\Psi(j)}\right\|\Big|_{L^2(\Omega,M_2(\mathbb{R}))} &\leq \left\|V_{l,\Psi(j)} - \nabla\varphi_{l,\Psi(j)}\right\|_{L^2(\Omega,M_2(\mathbb{R}))},\\ &\leq \left(\frac{2}{\gamma_0}\left(\inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + \frac{C+1}{l} + \frac{1}{\gamma_0}\right)\right)^{\frac{1}{2}}.\end{aligned}$$

La suite $(\varphi_{l,\Psi(j)})$ est donc uniformément bornée dans $W^{1,2}(\Omega, \mathbb{R}^2)$ d'après l'inégalité généralisée de Poincaré, et il existe une sous-suite obtenue en appliquant une procédure d'extraction diagonale et toujours désignée par $(\varphi_{l,\Psi(j)})$ et $\bar{\varphi}_l \in \mathrm{Id} + W^{1,2}_0(\Omega, \mathbb{R}^2)$ telles que

$$\varphi_{l,\Psi(j)} \underset{j\to+\infty}{\rightharpoonup} \bar{\varphi}_l \text{ dans } W^{1,2}(\Omega, \mathbb{R}^2).$$

Le même type d'argument appliqué à $(W_{l,\Psi(j)})$ mène à

$$W_{l,\Psi(j)} \underset{j\to+\infty}{\rightharpoonup} \bar{W}_l \text{ dans } L^2(\Omega, M_2(\mathbb{R})).$$

En outre, $(\det W_{l,\Psi(j)})$ est uniformément borné en norme L^∞ et donc il existe $\bar{\bar{\delta}}_l \in L^\infty(\Omega)$ tel que

$$\det W_{l,\Psi(j)} \underset{j\to+\infty}{\overset{*}{\rightharpoonup}} \bar{\bar{\delta}}_l \text{ dans } L^\infty(\Omega),$$

et enfin, $\bar{\bar{\delta}}_l = 1$ *p.p.*.

Nous définissons maintenant $x_{l,\Psi(j)} = V_{l,\Psi(j)} - \nabla\varphi_{l,\Psi(j)}$. Puisque $\|V_{l,\Psi(j)} - \nabla\varphi_{l,\Psi(j)}\|_{L^2(\Omega,M_2(\mathbb{R}))} \leq \frac{2}{\gamma_{\Psi(j)}}\left[\inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + \frac{C+1}{l} + \frac{1}{\gamma_0}\right]$, cela implique que $x_{l,\Psi(j)} \underset{j\to+\infty}{\longrightarrow} 0$ dans $L^2(\Omega, M_2(\mathbb{R}))$. En conséquence, $V_{l,\Psi(j)} - \nabla\varphi_{l,\Psi(j)} \underset{j\to+\infty}{\rightharpoonup} 0$ dans $L^2(\Omega, M_2(\mathbb{R}))$, entraînant $\nabla\varphi_{l,\Psi(j)} \underset{j\to+\infty}{\rightharpoonup} \bar{V}_l$ dans $L^2(\Omega, M_2(\mathbb{R}))$. En effet, $\forall\phi \in L^2(\Omega, M_2(\mathbb{R}))$, $\int_\Omega x_{l,\Psi(j)} : \phi\, dx \underset{j\to+\infty}{\to} 0$, c'est-à-dire, $\int_\Omega V_{l,\Psi(j)} - \nabla\varphi_{l,\Psi(j)} : \phi\, dx \underset{j\to+\infty}{\to} 0$. Mais $V_{l,\Psi(j)} \underset{j\to+\infty}{\rightharpoonup} \bar{V}_l$ dans $L^4(\Omega, M_2(\mathbb{R}))$, donc dans $L^2(\Omega, M_2(\mathbb{R}))$, ce qui montre que, $\forall\phi \in L^2(\Omega, M_2(\mathbb{R}))$,

$$\begin{aligned}\int_\Omega \nabla\varphi_{l,\Psi(j)} : \phi\, dx &= \int_\Omega \nabla\varphi_{l,\Psi(j)} - V_{l,\Psi(j)} : \phi\, dx + \int_\Omega V_{l,\Psi(j)} : \phi\, dx,\\ &\underset{j\to+\infty}{\longrightarrow} \int_\Omega \bar{V}_l : \phi\, dx.\end{aligned}$$

Alors $\nabla\varphi_{l,\Psi(j)} \underset{j\to+\infty}{\rightharpoonup} \bar{V}_l$ dans $L^2(\Omega, M_2(\mathbb{R}))$. D'un autre côté, $\nabla\varphi_{l,\Psi(j)} \underset{j\to+\infty}{\rightharpoonup} \nabla\bar{\varphi}_l$ dans $L^2(\Omega, M_2(\mathbb{R}))$, ce qui, par unicité de la limite faible dans $L^2(\Omega, M_2(\mathbb{R}))$, permet de conclure que $\nabla\bar{\varphi}_l = \bar{V}_l \in L^4(\Omega, M_2(\mathbb{R}))$.
Concentrons-nous désormais sur $(\det V_{l,\Psi(j)})$. Nous avons $\det V_{l,\Psi(j)} = \det\nabla\varphi_{l,\Psi(j)} + d_{l,\Psi(j)}$ avec

$$\begin{aligned}d_{l,\Psi(j)} = \det x_{l,\Psi(j)} &+ \left(x_{l,\Psi(j)}\right)_{1,1}\frac{\partial\varphi^2_{l,\Psi(j)}}{\partial x_2} + \left(x_{l,\Psi(j)}\right)_{2,2}\frac{\partial\varphi^1_{l,\Psi(j)}}{\partial x_1}\\ &- \left(x_{l,\Psi(j)}\right)_{1,2}\frac{\partial\varphi^2_{l,\Psi(j)}}{\partial x_1} - \left(x_{l,\Psi(j)}\right)_{2,1}\frac{\partial\varphi^1_{l,\Psi(j)}}{\partial x_2},\end{aligned}$$

$\left(x_{l,\Psi(j)}\right)_{k,n}$ étant un élément de la $k^{\text{ième}}$ ligne et $n^{\text{ième}}$ colonne de la matrice $x_{l,\Psi(j)}$ et avec $\varphi_{l,\Psi(j)} = \left(\varphi^1_{l,\Psi(j)}, \varphi^2_{l,\Psi(j)}\right)$. Alors il s'ensuit, avec l'inégalité $ab \leq \frac{a^2+b^2}{2}$, que

$$\|d_{l,\Psi(j)}\|_{L^1(\Omega)} \leq \frac{1}{2}\|x_{l,\Psi(j)}\|^2_{L^2(\Omega,M_2(\mathbb{R}))} + \|x_{l,\Psi(j)}\|_{L^2(\Omega,M_2(\mathbb{R}))}\|\nabla\varphi_{l,\Psi(j)}\|_{L^2(\Omega,M_2(\mathbb{R}))}.$$

La quantité $\|\nabla\varphi_{l,\Psi(j)}\|_{L^2(\Omega,M_2(\mathbb{R}))}$ est bornée indépendamment de j et l, et comme $x_{l,\Psi(j)} \underset{j\to+\infty}{\longrightarrow} 0$ dans $L^2(\Omega, M_2(\mathbb{R}))$, il en découle que $d_{l,\Psi(j)} \underset{j\to+\infty}{\longrightarrow} 0$ dans $L^1(\Omega)$.

D'après [35, Théorème 1.14], si $\Psi_j \underset{j\to+\infty}{\rightharpoonup} \bar{\Psi}$ dans $W^{1,2}(\Omega,\mathbb{R}^2)$, alors $\det\nabla\Psi_j \underset{j\to+\infty}{\longrightarrow} \det\nabla\bar{\Psi}$ au sens des distributions. D'une part, $\forall\phi\in\mathcal{D}(\Omega)$,

$$\int_\Omega \det V_{l,\Psi(j)}\,\phi\,dx \underset{j\to+\infty}{\longrightarrow} \bar{\delta}_l\,\phi\,dx,$$

car $(\det V_{l,\Psi(j)})$ converge faiblement vers $\bar{\delta}_l$ dans $L^2(\Omega)$ lorsque j tend vers $+\infty$. D'autre part,

$$\int_\Omega \det V_{l,\Psi(j)}\,\phi\,dx = \int_\Omega \det\nabla\varphi_{l,\Psi(j)}\,\phi\,dx + \int_\Omega d_{l,\Psi(j)}\,\phi\,dx,$$

avec $\int_\Omega \det\nabla\varphi_{l,\Psi(j)}\,\phi\,dx \underset{j\to+\infty}{\longrightarrow} \int_\Omega \det\nabla\bar{\varphi}_l\,\phi\,dx$, comme $\det\nabla\varphi_{l,\Psi(j)}$ converge vers $\det\nabla\bar{\varphi}_l$ au sens des distributions quand j tend vers $+\infty$ et $\left|\int_\Omega d_{l,\Psi(j)}\,\phi\,dx\right| \leq \|d_{l,\Psi(j)}\|_{L^1(\Omega)}\|\phi\|_{\mathcal{C}^0(\bar{\Omega})} \underset{j\to+\infty}{\to} 0$. En conséquence, $L^2(\Omega)\ni\det\nabla\bar{\varphi}_l=\bar{\delta}_l\in L^2(\Omega)$ au sens des distributions et enfin $\det\nabla\bar{\varphi}_l=\bar{\delta}_l$ *p.p.*.
Un raisonnement similaire s'applique pour le traitement de la composante $\|W_{l,\Psi(j)} - \nabla\varphi_{l,\Psi(j)}\|^2_{L^2(\Omega,M_2(\mathbb{R}))}$ et aboutit à

$$\left|\begin{array}{c}\nabla\bar{\varphi}_l=\bar{V}_l=\bar{W}_l\in L^4(\Omega,M_2(\mathbb{R}))),\\ \det\nabla\bar{\varphi}_l=1\ \ p.p..\end{array}\right.$$

En invoquant à nouveau l'inégalité de Poincaré généralisée, il en résulte que $\bar{\varphi}_l\in \mathrm{Id}+W^{1,4}_0(\Omega,\mathbb{R}^2)$ avec $\det\nabla\bar{\varphi}_l=1$ *p.p.*. Subséquemment, toujours en appliquant les résultats de Ball, $\bar{\varphi}_l$ est un homéomorphisme de $\bar{\Omega}$ dans $\bar{\Omega}$ et $\bar{\varphi}_l^{-1}\in W^{1,4}(\Omega,\mathbb{R}^2)$. La fonction $\bar{\varphi}_l$ fait correspondre des ensembles mesurables dans $\bar{\Omega}$ à des ensembles mesurables dans $\bar{\Omega}$, et la formule classique du changement de variable s'applique.

L'utilisation de la propriété de Lipschitz de Ey_l ainsi que du théorème de Rellich-Kondrachov ([15, Théorème IX.16]), qui donne l'injection compacte de $W^{1,2}(\Omega,\mathbb{R}^2)\subset L^2(\Omega)$, permet d'observer que $\|s(\theta)-(Ey_l)\circ\varphi_{l,\Psi(j)}\|^2_{L^2(\Omega)} \underset{j\to+\infty}{\longrightarrow} \|s(\theta)-(Ey_l)\circ\bar{\varphi}_l\|^2_{L^2(\Omega)}$, donc en revenant à (4.6), nous obtenons

$$\begin{aligned}\liminf_{j\to+\infty}\Big[\mu\,\|V_{l,\Psi(j)}\|^4_{L^4(\Omega,M_2(\mathbb{R}))} + \frac{\nu}{2}\,\|s(\theta)-(Ey_l)\circ\varphi_{l,\Psi(j)}\|^2_{L^2(\Omega)} + \frac{\mu\alpha}{2}\,\|\det V_{l,\Psi(j)}-1\|^2_{L^2(\Omega)}\Big]\\ \leq \liminf_{j\to+\infty} E_{l,\Psi(j)}(\varphi_{l,\Psi(j)},V_{l,\Psi(j)},W_{l,\Psi(j)}) \leq \inf_{\varphi\in\mathcal{W}}\mathcal{F}(\varphi)+\frac{C+1}{l}.\end{aligned}$$

Ainsi,

$$\begin{aligned}\mu \left\|\nabla \bar{\varphi}_l\right\|^4_{L^4(\Omega,M_2(\mathbb{R}))} + \frac{\nu}{2}\left\|s(\theta) - (Ey_l)\circ\bar{\varphi}_l\right\|^2_{L^2(\Omega)} + \frac{\mu\alpha}{2}\left\|\det\nabla\bar{\varphi}_l - 1\right\|^2_{L^2(\Omega)} \\ \leq \liminf_{j\to+\infty} E_{l,\Psi(j)}(\varphi_{l,\Psi(j)}, V_{l,\Psi(j)}, W_{l,\Psi(j)}) \\ \leq \limsup_{j\to+\infty} E_{l,\Psi(j)}(\varphi_{l,\Psi(j)}, V_{l,\Psi(j)}, W_{l,\Psi(j)}) \\ \leq \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + \frac{C+1}{l},\end{aligned}$$

et enfin, comme $\det\nabla\bar{\varphi}_l = 1$ *p.p.*,

$$\begin{aligned}\mu \left\|\nabla \bar{\varphi}_l\right\|^4_{L^4(\Omega,M_2(\mathbb{R}))} + \frac{\nu}{2}\left\|s(\theta) - y_l\circ\bar{\varphi}_l\right\|^2_{L^2(\Omega)} &\leq \liminf_{j\to+\infty} E_{l,\Psi(j)}(\varphi_{l,\Psi(j)}, V_{l,\Psi(j)}, W_{l,\Psi(j)}) \\ &\leq \limsup_{j\to+\infty} E_{l,\Psi(j)}(\varphi_{l,\Psi(j)}, V_{l,\Psi(j)}, W_{l,\Psi(j)}) \\ &\leq \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) + \frac{C+1}{l}.\end{aligned}$$

Nous remarquons que $\forall\varphi \in \mathrm{Id} + W^{1,4}_0(\Omega,\mathbb{R}^2)$ avec $\det\nabla\varphi = 1$ *p.p.*,

$$E_{l,\Psi(j)}(\varphi_{l,\Psi(j)}, V_{l,\Psi(j)}, W_{l,\Psi(j)}) \leq E_{l,\Psi(j)}(\varphi, \nabla\varphi, \nabla\varphi) + \frac{1}{\gamma_{\Psi(j)}},$$

ce qui permet de conclure que $\bar{\varphi}_l$ est un minimiseur de $\mu\left\|\nabla\cdot\right\|^4_{L^4(\Omega,M_2(\mathbb{R}))} + \frac{\nu}{2}\left\|s(\theta) - y_l\circ\cdot\right\|^2_{L^2(\Omega)}$. En raisonnant comme avant,

$$\left|\left\|s(\theta) - y_l\circ\bar{\varphi}_l\right\|^2_{L^2(\Omega)} - \left\|s(\theta) - y\circ\bar{\varphi}_l\right\|^2_{L^2(\Omega)}\right| \leq C\left\|y_l - y\right\|_{L^p(\Omega)} \leq \frac{C}{l},$$

donc

$$\frac{\nu}{2}\left\|s(\theta) - y\circ\bar{\varphi}_l\right\|^2_{L^2(\Omega)} - \frac{C}{l} \leq \frac{\nu}{2}\left\|s(\theta) - y_l\circ\bar{\varphi}_l\right\|^2_{L^2(\Omega)},$$

avec $C := \frac{C\nu}{2}$ et

$$\inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) - \frac{C}{l} \leq \mathcal{F}(\bar{\varphi}_l) - \frac{C}{l} \leq \mu\left\|\nabla\bar{\varphi}_l\right\|^4_{L^4(\Omega,M_2(\mathbb{R}))} + \frac{\nu}{2}\left\|s(\theta) - y_l\circ\bar{\varphi}_l\right\|^2_{L^2(\Omega)}.$$

Le passage à la limite quand l tend vers $+\infty$ conduit à

$$\begin{aligned}\inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi) &\leq \liminf_{l\to+\infty}\liminf_{j\to+\infty} E_{l,\Psi(j)}(\varphi_{l,\Psi(j)}, V_{l,\Psi(j)}, W_{l,\Psi(j)}) \\ &\leq \limsup_{l\to+\infty}\limsup_{j\to+\infty} E_{l,\Psi(j)}(\varphi_{l,\Psi(j)}, V_{l,\Psi(j)}, W_{l,\Psi(j)}) \leq \inf_{\varphi\in\mathcal{W}} \mathcal{F}(\varphi),\end{aligned}$$

et le résultat suit. □

Forts de ces éléments théoriques, assurant le caractère bien posé du problème $(\mathcal{P})$, nous abordons maintenant la résolution du problème de minimisation de la fonction de perte $\mathcal{L}$.

4.3 Résolution numérique et implémentation

Dans cette section, nous exposons la procédure de résolution numérique du problème de minimisation de la fonction de perte et en donnons des précisions concernant l'implémentation.

4.3.1 Résolution numérique

Tout d'abord, nous rappelons que le problème de minimisation de la fonction de perte étudiée se formule comme :

$$\inf_{\theta,\,\varphi\in\mathcal{W}}\Bigg\{\mathcal{L}(\theta,\,\varphi)=\mathcal{F}_{DL}(s(\theta),y)+\mathcal{F}_{Rec}(\varphi)$$
$$=\mathcal{F}_{DL}(s(\theta),y)+\frac{\nu}{2}\left\|s(\theta)-y\circ\varphi\right\|^2_{L^2(\Omega)}+\int_\Omega W(\nabla\varphi)\,dx\Bigg\},\tag{4.7}$$

sous la contrainte $\det\nabla\varphi=1\ p.p.$, avec $\mathcal{W}=\{\psi\in\mathrm{Id}+W_0^{1,4}(\Omega,\mathbb{R}^2)\mid\det\nabla\psi=1\ p.p.\}$ et $\mathcal{F}_{DL}$ une fonction de perte standard en segmentation d'images, comme par exemple le score de Dice [127] ou l'entropie croisée.
Ce problème se résout par le biais d'une approche de *splitting* et d'un schéma alterné, détaillés dans l'Algorithme 5, et dont le principe consiste à actualiser de manière successive chaque variable en considérant l'autre fixe. Dans la suite, nous explicitons les méthodes de résolution des sous-problèmes relatifs aux deux variables θ et φ impliquées dans le problème (4.7).

Algorithme 5 Algorithme pour résoudre le problème (4.7)

Initialiser θ_0 aléatoirement
Fixer $\mu,\nu>0$
pour n$=1,2,\cdots$ **faire**
  Calculer $s(\theta_n)$
  $\varphi_{n+1}=\arg\min_{\varphi\in\mathcal{W}}\mathcal{F}_{Rec}(\varphi_n)$
  $\bar{\mathcal{L}}(\theta_n)=\mathcal{F}(s(\theta_n),y)+\frac{\nu}{2}\left\|s(\theta_n)-y\circ\varphi_{n+1}\right\|^2$
  $\theta_{n+1}=\theta_n-\eta\nabla_\theta\bar{\mathcal{L}}(\theta_n)$
fin pour

Actualisation de θ Pour mettre à jour les paramètres du réseau θ, nous considérons la variable φ fixe et utilisons simplement une technique de descente de gradient par lots sur la fonction perte (la fonctionnelle $\bar{\mathcal{L}}$ à minimiser par rapport à θ étant lisse (non-convexe) ici) qui s'exprime comme

$$\bar{\mathcal{L}}(\theta)=\mathcal{F}_{DL}(s(\theta),y)+\frac{\nu}{2}\left\|s(\theta)-y\circ\varphi\right\|^2.\tag{4.8}$$

Le gradient de cette fonction perte est donné par

$$\nabla_\theta \bar{\mathcal{L}}(\theta) = \nabla_\theta \, \mathcal{F}_{DL}(s(\theta), y) + \nu \; (s(\theta) - y \circ \varphi) \; \nabla_\theta s(\theta).$$

Les paramètres sont alors actualisés en prenant l'opposé du gradient *i.e.*, $\theta_{n+1} = \theta_n - \eta \, \nabla_\theta \bar{\mathcal{L}}(\theta_n)$, où η est le pas d'apprentissage et n représente l'ensemble de toutes les étapes successives de descente de gradient réalisées durant une époque.

Actualisation de φ Pour obtenir la déformation φ, nous considérons la variable θ fixe. Comme amorcé dans la partie théorique, le problème relatif à φ exhibe des difficultés numériques en raison de la non-linéarité et de la non-convexité. Pour parer à ce problème, l'idée consiste à introduire des variables auxiliaires pour séparer le problème initial en sous-problèmes plus faciles à résoudre. Nous introduisons donc deux variables auxiliaires V et W (sans confusion possible avec la fonction de la densité d'énergie W introduite plus tôt) telles que $V = \nabla\varphi$ et $W = \nabla\varphi$, le couplage s'opérant par le biais d'une pénalisation L^2. La contrainte dure est maintenue sur W, en revanche nous la relaxons sur V. Cela nous permet de reformuler le problème $(\mathcal{P})$, relativement aux variables φ, V, et W, de la manière suivante

$$\begin{aligned} \min_{\substack{(\varphi,V,W)\in \overline{\mathcal{W}}\times L^4(\Omega,M_2(\mathbb{R}))\\ \times\overline{\overline{\mathcal{W}}}}} \; & \mu\|V\|^4_{L^4(\Omega,M_2(\mathbb{R}))} + \frac{\nu}{2}\|s(\theta) - y\circ\varphi\|^2_{L^2(\Omega)} \\ & + \frac{\mu\alpha}{2}\|\det V - 1\|^2_{L^2(\Omega)} + \frac{\gamma_1}{2}\|W - \nabla\varphi\|^2 + \frac{\gamma_2}{2}\|V - \nabla\varphi\|^2, \end{aligned} \tag{$\mathcal{DP}$}$$

où α, γ_1 et γ_2 sont des paramètres positifs et avec $(\varphi, V, W) \in \left(\overline{\mathcal{W}} = \mathrm{Id} + W_0^{1,2}(\Omega, \mathbb{R}^2)\right) \times L^4(\Omega, M_2(\mathbb{R})) \times \left(\overline{\overline{\mathcal{W}}} = \{X \in L^2(\Omega, M_2(\mathbb{R})) \,|\, \det X = 1 \; p.p.\}\right)$.
Pour résoudre ce problème d'optimisation, nous utilisons une technique de *splitting*. À nouveau, l'optimisation de chacune des variables φ, V et W s'opère successivement en considérant les deux autres fixes. Par conséquent, le problème $(\mathcal{DP})$ est divisé en trois sous-problèmes plus facilement résolubles.

Nous nous concentrons d'abord sur le sous-problème de minimisation relatif à V formulé dans le domaine discret :

$$\min_V \mu\|V\|^4 + \frac{\mu\alpha}{2}\|\det V - 1\|^2 + \frac{\gamma_2}{2}\|V - \nabla\varphi\|^2 \tag{4.9}$$

Posons $V = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}$. Pour simplifier la résolution de ce problème en ne faisant intervenir que des termes quadratiques, nous introduisons la variable $q = (q_1, q_2, q_3, q_4)^T \in \mathbb{R}^4$

afin d'établir la relation $q = S.\begin{pmatrix} v_{11} \\ v_{22} \\ v_{12} \\ v_{21} \end{pmatrix}$ avec $S = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix}$, une matrice orthogonale.
Ainsi, nous obtenons

$$\begin{cases} v_{11} = \frac{q_1+q_2}{\sqrt{2}}, v_{12} = \frac{q_3+q_4}{\sqrt{2}}, \\ v_{21} = \frac{q_3-q_4}{\sqrt{2}}, v_{22} = \frac{q_1-q_2}{\sqrt{2}}, \end{cases}$$

$\det V = v_{11}v_{22} - v_{12}v_{21} = \frac{1}{2}(q_1^2 - q_2^2 - q_3^2 + q_4^2)$ et $\|V\|^4 = (v_{11}^2 + v_{12}^2 + v_{21}^2 + v_{22}^2)^2 = (q_1^2 + q_2^2 + q_3^2 + q_4^2)^2$.

Ce changement de variable nous amène donc à reformuler le problème (4.9) de manière équivalente, relativement à la nouvelle variable q :

$$\begin{aligned} &\min_q \mu\|q\|^4 + \frac{\mu\alpha}{2}(\frac{q_1^2 - q_2^2 - q_3^2 + q_4^2}{2} - 1)^2 + \frac{\gamma_2}{2}\|q\|^2 - \gamma_2(q : S\nabla\varphi), \\ =&\min_q \mu\|q\|^4 + \frac{\mu\alpha}{2}(\frac{q_1^2 - q_2^2 - q_3^2 + q_4^2}{2} - 1)^2 + \frac{\gamma_2}{2}\|q\|^2 - q : b, \end{aligned}$$

avec $b = \gamma_2 S\nabla\varphi \in \mathbb{R}^4$, ce qui conduit au système d'optimalité suivant

$$\begin{cases} 4\mu q_1\|q\|^2 + \mu\alpha q_1(\frac{q_1^2 - q_2^2 - q_3^2 + q_4^2}{2} - 1) + \gamma_2 q_1 = b_1, \\ 4\mu q_2\|q\|^2 - \mu\alpha q_2(\frac{q_1^2 - q_2^2 - q_3^2 + q_4^2}{2} - 1) + \gamma_2 q_2 = b_2, \\ 4\mu q_3\|q\|^2 - \mu\alpha q_3(\frac{q_1^2 - q_2^2 - q_3^2 + q_4^2}{2} - 1) + \gamma_2 q_3 = b_3, \\ 4\mu q_4\|q\|^2 + \mu\alpha q_4(\frac{q_1^2 - q_2^2 - q_3^2 + q_4^2}{2} - 1) + \gamma_2 q_4 = b_4. \end{cases}$$

En définissant les quatre nouvelles variables ci-dessous

$$\bar{b} = \begin{pmatrix} b_1 \\ b_4 \end{pmatrix} \text{ et } \quad \bar{q} = \begin{pmatrix} q_1 \\ q_4 \end{pmatrix},$$
$$\hat{b} = \begin{pmatrix} b_2 \\ b_3 \end{pmatrix} \text{ et } \quad \hat{q} = \begin{pmatrix} q_2 \\ q_3 \end{pmatrix},$$

le système d'optimalité précédent devient

$$\begin{cases} \bar{q}\left(4\mu\|q\|^2 + \mu\alpha(\frac{q_1^2 - q_2^2 - q_3^2 + q_4^2}{2} - 1) + \gamma_2\right) = \bar{b} \\ \hat{q}\left(4\mu\|q\|^2 - \mu\alpha(\frac{q_1^2 - q_2^2 - q_3^2 + q_4^2}{2} - 1) + \gamma_2\right) = \hat{b}. \end{cases}$$

En particulier, si $\alpha = 8$, un découplage s'opère et le système se réduit à

$$\begin{cases} \bar{q}\,(8\mu\|\bar{q}\|^2 - 8\mu + \gamma_2) = \bar{b} \\ \hat{q}\,(8\mu\|\hat{q}\|^2 - 8\mu + \gamma_2) = \hat{b}. \end{cases}$$

C'est l'hypothèse que nous faisons dans la suite : α est fixé à 8.
Il en résulte que $\bar{q}$ et $\bar{b}$ ainsi que $\hat{q}$ et $\hat{b}$ sont colinéaires, ce qui nous permet de les exprimer respectivement sous la forme $\bar{q} = d\bar{b}$ et $\hat{q} = c\hat{b}$. En remplaçant dans le système, nous obtenons alors

$$\begin{cases} d^3 + \dfrac{d(\gamma_2 - 8\mu)}{8\mu\|\bar{b}\|^2} - \dfrac{1}{8\mu\|\bar{b}\|^2} = 0 \\ c^3 + \dfrac{c(\gamma_2 + 8\mu)}{8\mu\|\hat{b}\|^2} - \dfrac{1}{8\mu\|\hat{b}\|^2} = 0, \end{cases}$$

dans le cas général $\|\bar{b}\|^2 \neq 0$ et $\|\hat{b}\|^2 \neq 0$. Nous traitons les cas particuliers plus loin dans le développement. L'objectif consiste donc à résoudre deux équations cubiques de la forme

$$P = \alpha^3 + A\alpha - B = 0 \qquad (\varepsilon)$$

avec d'une part, $\bar{A} = \dfrac{(\gamma_2 - 8\mu)}{8\mu\|\bar{b}\|^2}$, $\bar{B} = \dfrac{1}{8\mu\|\bar{b}\|^2}$ et d'autre part, $\hat{A} = \dfrac{(\gamma_2 + 8\mu)}{8\mu\|\hat{b}\|^2}$, $\hat{B} = \dfrac{1}{8\mu\|\hat{b}\|^2}$.
En introduisant une nouvelle variable auxiliaire $\alpha = w - \dfrac{A}{3w}$, l'équation générique cubique (ε) se ramène à :

$$w^3 - \frac{A^3}{27w^3} - B = 0.$$

En multipliant par w^3, puis en posant $W = w^3$ (sans confusion possible avec la variable auxiliaire W introduite plus tôt), elle se transforme en une équation quadratique qui s'écrit

$$W^2 - BW - \frac{A^3}{27} = 0. \qquad (*)$$

Nous fixons en particulier $\gamma_2 > 8\mu \geq 0$ pour avoir $A > 0$. Par conséquent le polynôme P admet une unique racine réelle qui est de surcroît positive. En effet, $P'(\alpha) = 3\alpha^2 + A > 0$, ce qui implique que P est strictement croissant et de plus $P(0) = -B \leq 0$.
En outre, nous calculons les racines de l'équation $(*)$ données par $W = \dfrac{1}{2}\left(B \pm \sqrt{B^2 + \dfrac{4A^3}{27}}\right)$,
donc l'équation $w^3 = W$ admet deux racines réelles distinctes :

$$w_1 = \left(\frac{B + \sqrt{B^2 + \dfrac{4A^3}{27}}}{2}\right)^{\frac{1}{3}} \quad \text{et} \quad w_2 = -\left(\frac{-B + \sqrt{B^2 + \dfrac{4A^3}{27}}}{2}\right)^{\frac{1}{3}}.$$

Rappelons que $\alpha = w - \dfrac{A}{3w} = \dfrac{3w^2 - A}{3w}$. Nous recherchons une racine positive, ce qui implique que $w \geq \sqrt{\dfrac{A}{3}}$ ou $w \leq -\sqrt{\dfrac{A}{3}}$. Nous avons bien $w_1 \geq \sqrt{\dfrac{A}{3}}$, en revanche $-\sqrt{\dfrac{A}{3}} \leq w_2 < \sqrt{\dfrac{A}{3}}$. En définitive, l'unique racine réelle et positive de P est

$$\alpha^* = w_1 - \frac{A}{3w_1}.$$

Pour résumer, si $(b_1^2 + b_4^2) > 0$ et $(b_2^2 + b_3^2) > 0$, il faut d'abord calculer les termes $\bar{A}$, $\bar{B}$, $\hat{A}$, $\hat{B}$ afin d'obtenir

$$\hat{w} = \left(\frac{\hat{B} + \sqrt{\hat{B}^2 + \dfrac{4\hat{A}^3}{27}}}{2} \right)^{\frac{1}{3}} \quad \text{et} \quad \bar{w} = \left(\frac{\bar{B} + \sqrt{\bar{B}^2 + \dfrac{4\bar{A}^3}{27}}}{2} \right)^{\frac{1}{3}}.$$

Nous déterminons ensuite la racine α^* dans les deux cas, et comme $q_1 = db_1 = \bar{\alpha}^* b_1$, $q_2 = cb_2 = \hat{\alpha}^* b_2$, $q_3 = cb_3 = \hat{\alpha}^* b_3$ et $q_4 = db_4 = \bar{\alpha}^* b_4$, nous sommes finalement en mesure de retrouver V.

Pour finir, intéressons-nous aux cas particuliers. Si $(b_1^2 + b_4^2) = 0$, il vient simplement $d(\gamma_2 - 8\mu) = 1$ et donc $d = \dfrac{1}{\gamma_2 - 8\mu}$. De la même façon, si $(b_2^2 + b_3^2) = 0$, il vient $c(\gamma_2 + 8\mu) = 1$ et donc $c = \dfrac{1}{\gamma_2 + 8\mu}$.

Maintenant, pour résoudre, dans le cas discret, le problème relatif à W :

$$\min_W \frac{\gamma_1}{2} \|W - \nabla\varphi\|^2, \tag{4.10}$$

sous la contrainte $\det W = 1$, nous nous appuyons sur les travaux de Glowinski [56]. Nous posons d'abord $W = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$. Ensuite, de la même manière que précédemment, nous introduisons la variable $z = (z_1, z_2, z_3, z_4)^T \in \mathbb{R}^4$ telle que $z = S. \begin{pmatrix} w_{11} \\ w_{22} \\ w_{12} \\ w_{21} \end{pmatrix}$.
Le problème est découplé par rapport à chaque pixel, nous permettant de reformuler le

problème (4.10) dans $\mathbb{R}^4$ comme

$$\begin{aligned} &\min_{z\in\mathcal{B}} \frac{\gamma_1}{2}\|z\|^2 - z : \gamma_1 S\nabla\varphi, \\ =&\min_{z\in\mathcal{B}} \frac{\gamma_1}{2}\|z\|^2 - z : b, \end{aligned} \tag{$\mathcal{D}$}$$

avec $\mathcal{B} = \{z \in \mathbb{R}^4 \,|\, z_1^2 - z_2^2 - z_3^2 + z_4^2 = 2\}$ et $b = \gamma_1 S\nabla\varphi \in \mathbb{R}^4$. Soit y solution du problème de minimisation précédent ($\mathcal{D}$). Nous introduisons le multiplicateur de Lagrange λ associé à la contrainte d'égalité $h(z) = \frac{1}{2}(z_1^2 - z_2^2 - z_3^2 + z_4^2 - 2) = 0$. Nous montrons d'abord que la contrainte est qualifiée en tout point. Soit $z \in \mathcal{B}$, nous avons

$$\nabla h(z) = \begin{pmatrix} z_1 \\ -z_2 \\ -z_3 \\ z_4 \end{pmatrix},$$

alors $\nabla h(z) = 0$ si et seulement si $z = 0_{\mathbb{R}^4}$. Or $0_{\mathbb{R}^4}$ ne vérifie pas la contrainte, par conséquent nous pouvons conclure qu'elle est qualifiée en tout point. Les conditions de Karush-Kuhn-Tucker sont satisfaites et nous obtenons le système d'optimalité suivant :

$$\begin{cases} \gamma_1 y_1 = b_1 - \lambda y_1 \\ \gamma_1 y_2 = b_2 + \lambda y_2 \\ \gamma_1 y_3 = b_3 + \lambda y_3 \\ \gamma_1 y_4 = b_4 - \lambda y_4 \\ y_1^2 - y_2^2 - y_3^2 + y_4^2 = 2, \end{cases} \tag{4.11}$$

ce qui implique alors que λ est solution de

$$\frac{b_1^2 + b_4^2}{(\gamma_1 + \lambda)^2} = \frac{b_2^2 + b_3^2}{(\gamma_1 - \lambda)^2} + 2. \tag{4.12}$$

Pour conclure cette analyse concernant la variable auxiliaire W, nous détaillons la démarche pour déterminer le multiplicateur de Lagrange λ, d'abord dans le cas général $b_1^2 + b_4^2 \neq 0$ et $b_2^2 + b_3^2 \neq 0$, pour lequel un résultat d'existence et d'unicité de la solution est donné, puis dans les cas particuliers.

Détermination du multiplicateur : En fait, les équations (4.11) et (4.12) caractérisent chaque extremum de la fonctionnelle qui intervient dans le problème ($\mathcal{D}$) et définie par

$$\mathcal{J}(z) = \frac{\gamma_1}{2}\|z\|^2 - z : b,$$

pour $z \in \mathcal{B}$. Puisque nous recherchons seulement les minima, il faut imposer quelques restrictions aux solutions λ de (4.12).

Nous traitons d'abord le cas général $b_1^2 + b_4^2 \neq 0$ et $b_2^2 + b_3^2 \neq 0$. Sans perte de généralité, nous supposons alors que $b_1 \neq 0$ et $b_3 \neq 0$ impliquant, d'après (4.11) que $y_1 \neq 0$ et $y_3 \neq 0$. Pour trouver une caractérisation des minima de $\mathcal{J}$, nous utilisons une condition nécessaire d'optimalité du deuxième ordre ([2, Proposition 10.2.11]) stipulant que tout minimum local y de $\mathcal{J}$ sur $\mathcal{B}$ vérifie

$$\left(D^2\mathcal{J}(y) + \lambda D^2 h(y)\right)(w, w) \geq 0, \qquad \forall w \in K(y) = \{w \in \mathbb{R}^4 | < \nabla h(y), w >= 0\},$$

où $<,>$ désigne le produit scalaire dans $\mathbb{R}^4$.

Cela se traduit dans notre cas par, $\forall w \in K(y)$,

$$w^T \begin{pmatrix} \gamma_1 + \lambda & 0 & 0 & 0 \\ 0 & \gamma_1 - \lambda & 0 & 0 \\ 0 & 0 & \gamma_1 - \lambda & 0 \\ 0 & 0 & 0 & \gamma_1 + \lambda \end{pmatrix} w \geq 0.$$

Puisque $\nabla h(y) = (y_1, -y_2, -y_3, y_4)^T$, nous choisissons en particulier $w = (-y_4, 0, 0, y_1)^T \in K(y)$, menant à

$$(y_1^2 + y_4^2)(\gamma_1 + \lambda) \geq 0.$$

En prenant maintenant $w = (0, -y_3, y_2, 0)^T$, cela conduit à

$$(y_2^2 + y_3^2)(\gamma_1 - \lambda) \geq 0.$$

Par conséquent, puisque $y_1 \neq 0$ et $y_3 \neq 0$ dans le cas général, pour avoir un minimum il faut nécessairement que

$$\begin{cases} \gamma_1 + \lambda \geq 0 \\ \gamma_1 - \lambda \geq 0. \end{cases}$$

Le multiplicateur de Lagrange λ doit donc satisfaire $|\lambda| \leq \gamma_1$. Comme le montre la Figure (4.2), il y a une seule solution à l'équation (4.12) qui appartient à $[-\gamma_1, +\gamma_1]$ (elle appartient en fait à $]-\gamma_1, +\gamma_1[$).

En résumé, dans ce cas où $b_1^2 + b_4^2 \neq 0$ et $b_2^2 + b_3^2 \neq 0$, il est d'abord nécessaire de trouver la valeur de λ comprise dans l'intervalle $]-\gamma_1, +\gamma_1[$. Nous y parvenons par le biais de l'algorithme de Newton dont l'initialisation est réalisée par une étape préliminaire de dichotomie. Ensuite, les solutions y du problème sont données par $y_i = \dfrac{b_i}{\gamma_1 \pm \lambda}$ avec $i = \{1, 2, 3, 4\}$.

Nous énonçons maintenant un théorème d'existence et d'unicité de la solution dans ce cas général.

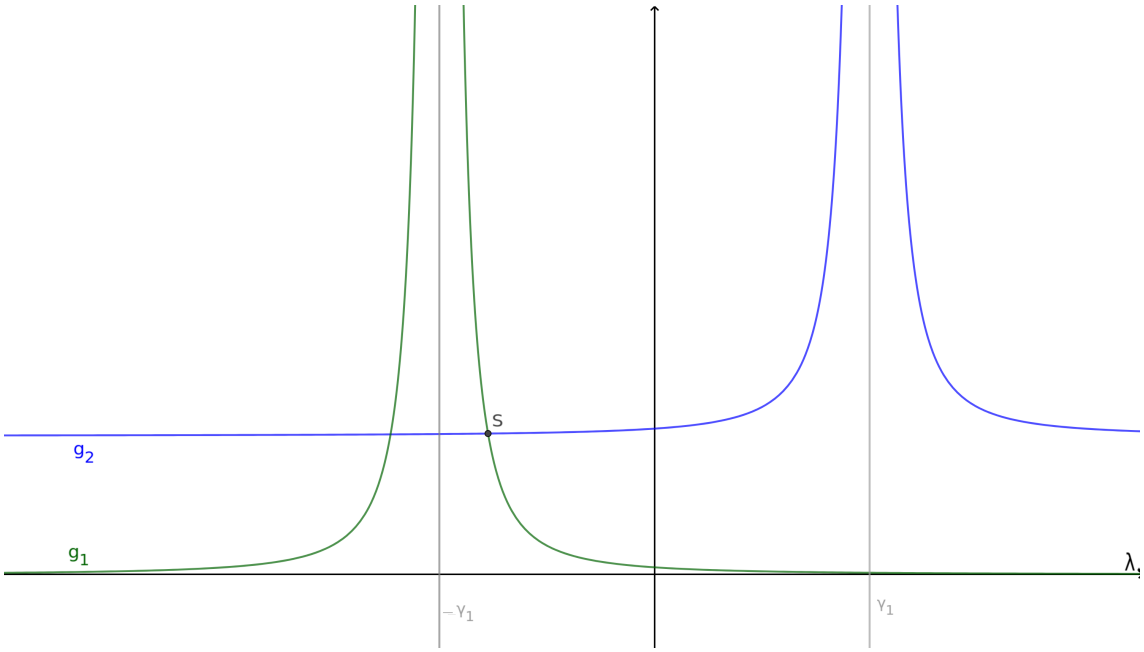


FIGURE 4.2 – Solution de (4.12), *i.e.* $g_1 = g_2$ avec $g_1 = \frac{b_1^2+b_4^2}{(\gamma_1+\lambda)^2}$ et $g_2 = 2 + \frac{b_2^2+b_3^2}{(\gamma_1-\lambda)^2}$, dans le cas général où $b_1^2 + b_4^2 \neq 0$ et $b_2^2 + b_3^2 \neq 0$. S représente la solution unique. (Extrait de [56]).

Théorème 4.3.1. *Dans le cas où $b_1^2 + b_4^2 \neq 0$ et $b_2^2 + b_3^2 \neq 0$, le problème $(\mathcal{D})$ admet une solution unique.*

Démonstration. La partie existence est évidente. Soit y une solution caractérisée précédemment et soit $\mathcal{C}$ la quantité définie par

$$\mathcal{C} = \mathcal{J}(y + \delta y) - \mathcal{J}(y),$$

avec $(y + \delta y) \in \mathcal{B}$. Alors

$$\begin{aligned}
\mathcal{C} &= \frac{\gamma_1}{2} \|y + \delta y\|^2 - (y + \delta y) : b - \frac{\gamma_1}{2} \|y\|^2 + y : b, \\
&= \gamma_1\, y : \delta y + \frac{\gamma_1}{2} \|\delta y\|^2 - \delta y : b, \\
&= (\gamma_1 y - b) : \delta y + \frac{\gamma_1}{2} \|\delta y\|^2.
\end{aligned}$$

$$\text{Mais} \left\{ \begin{array}{rcl} \gamma_1 y_1 - b_1 + \lambda\, y_1 & = & 0 \\ \gamma_1 y_2 - b_2 - \lambda\, y_2 & = & 0 \\ \gamma_1 y_3 - b_3 - \lambda\, y_3 & = & 0 \\ \gamma_1 y_4 - b_4 + \lambda\, y_4 & = & 0 \\ y_1^2 - y_2^2 - y_3^2 + y_4^2 & = & 2 \end{array} \right. \text{, avec } \lambda \in (-\gamma_1, \gamma_1), \text{ menant à}$$

$$\mathcal{C} = \begin{pmatrix} -\lambda\, y_1 \\ \lambda\, y_2 \\ \lambda\, y_3 \\ -\lambda y_4 \end{pmatrix} : \delta y + \frac{\gamma_1}{2} \|\delta y\|^2.$$

Puisque les variables y et $(y + \delta y)$ appartiennent toutes deux à $\mathcal{B}$, il s'ensuit que

$$\left\{ \begin{array}{rcl} y_1^2 - y_2^2 - y_3^2 + y_4^2 & = & 2 \\ (y_1 + \delta y_1)^2 - (y_2 + \delta y_2)^2 - (y_3 + \delta y_3)^2 + (y_4 + \delta y_4)^2 & = & 2 \end{array} \right. .$$

Il en résulte que

$$\begin{aligned} & 2y_1\delta y_1 - 2y_2\delta y_2 - 2y_3\delta y_3 + 2y_4\delta y_4 + \delta y_1^2 - \delta y_2^2 - \delta y_3^2 + \delta y_4^2 = 0, \\ \Longleftrightarrow & \frac{\lambda}{2} \left[\delta y_1^2 - \delta y_2^2 - \delta y_3^2 + \delta y_4^2\right] = -\lambda y_1\delta y_1 + \lambda y_2\delta y_2 + \lambda y_3\delta y_3 - \lambda y_4\delta y_4. \end{aligned}$$

Par conséquent,

$$\mathcal{C} = \frac{\gamma_1 + \lambda}{2}\, \delta y_1^2 + \frac{\gamma_1 - \lambda}{2}\, \delta y_2^2 + \frac{\gamma_1 - \lambda}{2}\, \delta y_3^2 + \frac{\gamma_1 + \lambda}{2}\, \delta y_4^2,$$

cette quantité étant positive puisque $\lambda \in (-\gamma_1, \gamma_1)$, égale à zéro si et seulement si $\delta y = 0_{\mathbb{R}^4}$. □

Nous discutons maintenant des cas particuliers. Nous traitons le premier cas particulier $b_1 = b_4 = 0$. À partir du système d'optimalité (4.11), nous trouvons comme solution

$$y_2 = \frac{b_2}{2\gamma_1}, \quad y_3 = \frac{b_3}{2\gamma_1}, \quad \{y_1, y_4\} \in \mathcal{C}_1$$

où $\mathcal{C}_1 = \left\{\{\xi, \eta\} \in \mathbb{R}^2, \xi^2 + \eta^2 = 2 + \frac{b_2^2 + b_3^2}{4\gamma_1^2}\right\}$. Par conséquent, dans ce cas l'équation (4.12) admet $\lambda = -\gamma_1$ comme solution unique dans [-γ_1,+γ_1] (voir Fig. 4.3) et nous observons que y_2 et y_3 sont toujours données par (4.11) (avec $\lambda = -\gamma_1$). De plus, nous choisissons arbitrairement de prendre $y_1 = y_4 = \frac{\sqrt{2}}{2}\sqrt{2 + \frac{b_2^2 + b_3^2}{4\gamma_1^2}}$.

Enfin, nous abordons le cas particulier $b_2 = b_3 = 0$. À partir de (4.11) nous trouvons que :

$$\left\{ \begin{array}{l} (\gamma_1 - \lambda) y_2 = 0, \\ (\gamma_1 - \lambda) y_3 = 0. \end{array} \right.$$

Deux cas se distinguent à nouveau :

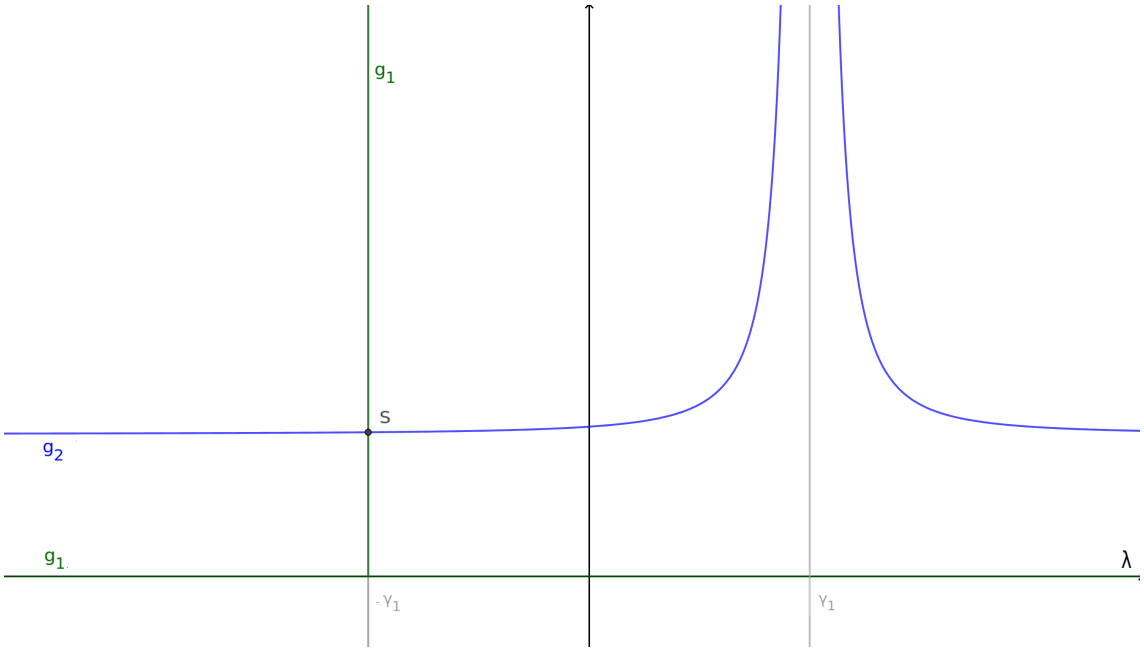


FIGURE 4.3 – Solution de (4.12), *i.e.* $g_1 = g_2$ avec $g_1 = 0$, si $\lambda \neq -\gamma_1, [0, +\infty[$ sinon, et $g_2 = 2 + \frac{b_2^2+b_3^2}{(\gamma_1-\lambda)^2}$, dans le cas où $b_1 = b_4 = 0$. S représente la solution unique. (Extrait de [56]).

(i) si $\lambda = \gamma_1$ alors

$$y_1 = \frac{b_1}{2\gamma_1}, \quad y_4 = \frac{b_4}{2\gamma_1}.$$

L'équation (4.12) devient

$$\frac{b_1^2 + b_4^2}{4\gamma_1^2} = 2 + y_2^2 + y_3^2.$$

Si $b_1^2+b_4^2 > 8\gamma_1^2$, alors nécessairement $\{y_2, y_3\} \in \mathcal{C}_2$ où $\mathcal{C}_2 = \left\{\{\xi, \eta\} \in \mathbb{R}^2, \xi^2 + \eta^2 = \frac{b_1^2+b_4^2}{4\gamma_1^2} - 2\right\}$. Comme précédemment, nous choisissons de prendre arbitrairement $y_2 = y_3 = \frac{\sqrt{2}}{2}\sqrt{-2 + \frac{b_1^2+b_4^2}{4\gamma_1^2}}$. La condition $b_1^2 + b_4^2 \leq 8\gamma_1^2$ correspond au cas (ii) suivant.

(ii) si $y_2 = y_3 = 0$, alors la contrainte devient $y_1^2 + y_4^2 = 2$ et avec (4.11) il vient que

$$\frac{b_1^2 + b_4^2}{(\gamma_1 + \lambda)^2} = 2,$$

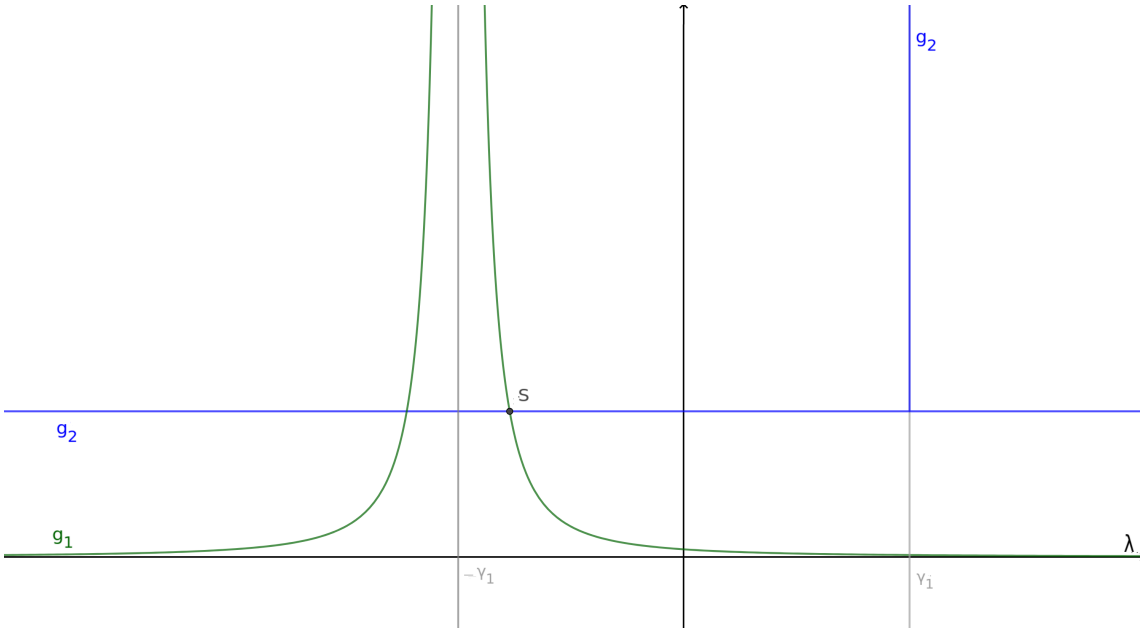


FIGURE 4.4 – Solution de (4.12), *i.e.* $g_1 = g_2$ avec $g_1 = \frac{b_1^2+b_4^2}{(\gamma_1+\lambda)^2}$ et $g_2 = 2$, si $\lambda \neq \gamma_1$, $[2, +\infty[$ sinon, dans le cas où $b_2 = b_3 = 0$. S représente la solution unique. (Extrait de [56]).

ce qui implique $\lambda = -\gamma_1 + \left(\frac{b_1^2+b_4^2}{2}\right)^{\frac{1}{2}}$. Finalement, la solution est donnée par

$$y_1 = \frac{\sqrt{2}b_1}{\sqrt{b_1^2+b_4^2}}, \quad y_4 = \frac{\sqrt{2}b_4}{\sqrt{b_1^2+b_4^2}}, \quad y_2 = y_3 = 0.$$

Alors l'équation (4.12) admet $\lambda = \gamma_1$ (resp. $\lambda = -\gamma_1 + \left(\frac{b_1^2+b_4^2}{2}\right)^{\frac{1}{2}}$) comme unique solution dans l'intervalle [-γ_1,+γ_1] dans le cas (i) (resp. (ii)) (voir Fig 4.4) et nous observons que y_1, y_4 sont toujours obtenues par (4.11) avec λ comme précisé.

Nous avons donc développé les démarches à suivre pour calculer le multiplicateur de Lagrange λ pour obtenir ensuite les solutions y du problème $(\mathcal{D})$, ce qui nous permet finalement de retrouver la valeur W qui minimise le sous-problème initial (4.10).

Il reste maintenant à étudier, toujours dans un cadre discret, le sous-problème de minimisation relatif à φ

$$\min_{\varphi} \frac{\nu}{2}\|s(\theta) - y \circ \varphi\|^2 + \frac{\gamma_1}{2}\|W - \nabla\varphi\|^2 + \frac{\gamma_2}{2}\|V - \nabla\varphi\|^2. \tag{4.13}$$

Numériquement, nous résolvons les équations d'Euler-Lagrange en utilisant une méthode de descente de gradient, où la direction de descente est paramétrée par un temps artificiel

$t \geq 0$. Les équations d'Euler-Lagrange pour ce problème sont données par

$$-\nu(s(\theta) - y \circ \varphi)\nabla y(\varphi) + \gamma_1 \begin{pmatrix} \operatorname{div} W_1 \\ \operatorname{div} W_2 \end{pmatrix} + \gamma_2 \begin{pmatrix} \operatorname{div} V_1 \\ \operatorname{div} V_2 \end{pmatrix} - (\gamma_1 + \gamma_2)\Delta\varphi = 0,$$

où V_1 et V_2 sont respectivement la première ligne et la seconde ligne de V, et de même pour W_1 et W_2. Ces équations sont discrétisées en utilisant un schéma implicite de type différences finies avec un schéma à 5 points pour le Laplacien. En pratique, nous cherchons u, le champ de déplacement associé à φ, plutôt que φ elle-même. Il faut donc résoudre un système linéaire de la forme

$$AU_l^{n+1} = B(U_l^n)$$

avec A $\in M_{(M-2)\times(N-2)}(\mathbb{R})$, U_l^{n+1}, U_l^n, $B \in \mathbb{R}^{(M-2)\times(N-2)}$ et l = 1,2. Ici M et N désignent les dimensions de la grille qui correspondent à la taille de l'image. Le déplacement est supposé nul sur les bords donc il n'est pas nécessaire de résoudre le système pour $i = \{1, M\}$ ou $j = \{1, N\}$. Nous avons donc, en considérant une concaténation par ligne

$$U_l^n = \begin{pmatrix} u_{l,2,2}^n \\ u_{l,2,3}^n \\ \vdots \\ u_{l,2,j}^n \\ \vdots \\ u_{l,i,j-1}^n \\ u_{l,i,j}^n \\ u_{l,i,j+1}^n \\ \vdots \\ u_{l,M-1,N-1}^n \end{pmatrix}, U_l^{n+1} = \begin{pmatrix} u_{l,2,2}^{n+1} \\ u_{l,2,3}^{n+1} \\ \vdots \\ u_{l,2,j}^{n+1} \\ \vdots \\ u_{l,i,j-1}^{n+1} \\ u_{l,i,j}^{n+1} \\ u_{l,i,j+1}^{n+1} \\ \vdots \\ u_{l,M-1,N-1}^{n+1} \end{pmatrix}$$

et

$$B(U_l^n) = U_l^n + dt \left(\nu \left(s(\theta) - y \circ \varphi^n \right) \frac{\partial y}{\partial x_l}(\varphi^n) - \gamma_1 \operatorname{div} W_l - \gamma_2 \operatorname{div} V_l \right).$$

La matrice A est symétrique tridiagonale par blocs :

$$A = \begin{pmatrix} D & -\gamma I_{N-2} & \mathbb{0}_{N-2} & \dots & \mathbb{0}_{N-2} \\ -\gamma I_{N-2} & D & -\gamma I_{N-2} & \mathbb{0}_{N-2} & \vdots \\ \mathbb{0}_{N-2} & -\gamma I_{N-2} & \ddots & \ddots & \mathbb{0}_{N-2} \\ \vdots & \mathbb{0}_{N-2} & \ddots & D & -\gamma I_{N-2} \\ \mathbb{0}_{N-2} & \dots & \mathbb{0}_{N-2} & -\gamma I_{N-2} & D \end{pmatrix}$$

avec

$$D = \begin{pmatrix} d & -\gamma & 0 & 0 \\ -\gamma & d & \ddots & 0 \\ 0 & \ddots & \ddots & -\gamma \\ 0 & 0 & -\gamma & d \end{pmatrix} \in M_{(N-2)}(\mathbb{R}),$$

où $d = 1 + \dfrac{4(\gamma_1+\gamma_2)dt}{h^2}$, $\gamma = \dfrac{(\gamma_1+\gamma_2)dt}{h^2}$, dt est le pas temporel, h est le pas spatial, I_{N-2} la matrice identité et $\mathbb{0}_{N-2}$ la matrice nulle d'ordre $N-2$.

Pour terminer, l'Algorithme 6 récapitule les étapes successives de résolution du problème de minimisation $(\mathcal{P})$. Des détails d'implémentation pour l'Algorithme 5 et l'Algorithme 6 suivent dans la prochaine section.

Algorithme 6 Approche de *splitting* pour la résolution du problème $(\mathcal{P})$ relatif à φ

Initialisation $\varphi^0 = \mathrm{Id}$, $W^0 = V^0 = I_2$,
Fixer $\gamma_1, \gamma_2, \mu, \nu > 0$
pour $n = 0, 1, \cdots$ **faire**
$V^{n+1} = \arg\min_V \ \mu\|V^n\|^4 + \dfrac{\mu\alpha}{2}\|\det V^n - 1\|^+ \dfrac{\gamma_2}{2}\|V^n - \nabla\varphi^n\|^2$
$W^{n+1} = \arg\min_W \ \dfrac{\gamma_1}{2}\|W^n - \nabla\varphi^n\|^2 \ \ \text{s.c} \ \ \det W^n = 1$
$\varphi^{n+1} = \arg\min_\varphi \ \dfrac{\nu}{2}\|s(\theta) - y \circ \varphi^n\|^2 + \dfrac{\gamma_1}{2}\|W^{n+1} - \nabla\varphi^n\|^2 + \dfrac{\gamma_2}{2}\|V^{n+1} - \nabla\varphi^n\|^2$
fin pour

4.3.2 Implémentation

Notre modèle conjoint se divise en une première partie orientée DL et une seconde qui s'inscrit dans un cadre variationnel. Les ressources nécessaires de calcul et de mémoire peuvent représenter un frein important à l'application de ce modèle. L'optimisation de notre code de calcul constitue donc un point essentiel de notre travail. Concernant la partie DL, le réseau de neurones convolutif impliqué dans l'apprentissage des paramètres θ est développé en Python avec la librairie Pytorch et parallélisé sur plusieurs GPU. La partie variationnelle dédiée à l'optimisation de la déformation φ, c'est-à-dire à la résolution du problème de recalage, est, elle aussi, majoritairement développée en Python. Cependant, deux points critiques se dégagent de l'Algorithme 6.
Le premier concerne la procédure d'interpolation, dans notre cas par splines cubiques. En effet l'algorithme requiert l'évaluation de la vérité terrain y en $\varphi(x)$. Or les différentes librairies développées en Python pour l'interpolation sont gourmandes en temps et en coût de calculs. Pour pallier cet effet, nous avons développé une librairie en langage C qui permet de calculer d'une part les coefficients d'interpolation, et de procéder aux différentes

étapes d'interpolation d'autre part. Nous utilisons notamment la librairie BLAS pour les calculs et OpenMP pour la parallélisation des différentes opérations. L'appel en Python à notre librairie s'effectue grâce à la bibliothèque ctypes.
Le second point se rapporte à la résolution du système linéaire $AU = B$ (et donc lié à φ) qui peut également poser problème. D'une part, la matrice A est composée de $[(M-2) \times (N-2)]^2$ éléments rendant son stockage en mémoire compliqué, voire impossible pour de grandes images. D'autre part, la résolution de ce système intervient à de multiples reprises, nécessitant un choix optimal quant à la méthode de résolution. Nous remarquons qu'il est possible d'effectuer une décomposition spectrale de la matrice A par le biais d'une transformée en sinus discrète (notée DST, *Discrete Sine Transform*). Les conditions aux limites de A correspondent à une DST normalisée d'ordre 1 (notée DST-I). Ainsi, en appliquant la DST-I pour résoudre le système linéaire, nous obtenons

$$FD \cdot \mathrm{DST}(U_l^{n+1}) = \mathrm{DST}(B(U_l^n))$$

où "$\cdot$" désigne la multiplication vectorielle terme à terme, et pour $1 \leq i \leq M-2$, $1 \leq j \leq N-2$,

$$FD = 1 + \frac{4(\gamma_1+\gamma_2)dt}{h^2} - \frac{2(\gamma_1+\gamma_2)dt}{h^2}\left(cos(\frac{i}{M-1}\pi) + cos(\frac{j}{N-1}\pi)\right) \in \mathbb{R}^{(M-2)\times(N-2)}.$$

Finalement,

$$U_l^{n+1} = \mathrm{DST}^{-1}\Big(\mathrm{DST}(B(U_l^n)) \cdot / FD\Big).$$

Ici "$\cdot/$" signifie que la division des vecteurs s'effectue terme à terme. Non seulement le calcul du terme FD s'opère une seule fois au début de l'algorithme, mais son coût de stockage est largement inférieur à celui de la matrice A. Nous implémentons donc cette fonction DST-I en Python.

Grâce aux solutions explicites obtenues dans cette section, et aux efforts d'implémentation fournis, nous sommes en mesure d'intégrer le modèle conjoint dans une tâche de segmentation automatique. La section suivante examine le comportement de la méthode, avec dans un premier temps, des simulations numériques pour le modèle de recalage, suivies par des résultats pour le traitement global.

4.4 Simulations numériques préalables, éléments quantitatifs et résultats

Dans cette section, nous présentons d'abord des simulations numériques illustrant le comportement de la méthode variationnelle développée pour la résolution du problème de recalage ($\mathcal{P}$). Ensuite, nous évaluons l'ensemble de notre méthode conjointe d'une part sur un jeu de données issu du challenge Decathlon, où l'objectif réside dans la segmentation

de l'atrium au sein d'images IRM cardiaques, et d'autre part sur le jeu de données SegTHOR, présenté dans le Chapitre 1, et cela en vue de souligner entre autres le caractère généralisable du modèle.

4.4.1 Simulations pour le recalage

Classiquement, nous prenons le pas spatial h égal à 1. Le pas temporel dt est quant à lui réglé à 10^{-2}. Après une recherche conduite par quadrillage, nous fixons les hyperparamètres de la manière suivante : $\gamma_1 = 10000$, $\gamma_2 = 80000$, $\nu = 1$ et $\mu = 125$. Nous présentons à présent trois exemples, dont deux très simples, obtenus avec l'Algorithme 6.

1. Exemple n°1 : la Figure 4.5 illustre notre méthode variationnelle dans le cas où nous souhaitons effectuer le recalage d'une image Template T d'un rectangle (b) (identifiée par y dans l'Algorithme 6), sur une image Référence R simplement simulée par un disque (a) (assimilée cette fois à $s(\theta)$). Dans cet exemple, la grille discrète est de taille $N \times N$, avec $N = 100$. De plus, les deux images Template et Référence sont simulées de sorte à avoir une aire identique. Nous observons que la déformation φ obtenue (c) n'exhibe pas d'auto-intersection de matière et permet de transformer l'image Template en une version très proche de l'image Référence. En effet, nous voyons seulement une pénalisation au niveau des frontières du disque (e), résultat d'un effet de bord (probablement en lien avec la phase d'interpolation).
2. Exemple n°2 : sur la Figure 4.6, nous visualisons la méthode pour déformer la même image Template rectangulaire T (b) afin d'approcher celle de Référence R composée cette fois de deux composantes connexes circulaires (a), toujours avec la même grille discrète. La déformation φ calculée (c) transforme assez largement le rectangle de manière à esquisser deux disques, mais qui sont reliés l'un à l'autre. De cette manière, le Template déformé (d) reste composé d'une seule composante connexe. Par conséquent, la liaison entre les deux éléments circulaires représente la pénalisation la plus forte (e).
3. Exemple n°3 : pour le dernier exemple, présenté sur la Figure 4.7, nous simulons une prédiction de segmentation $s(\theta)$ de trois OAR (a) issus du jeu de données SegTHOR : l'aorte (en jaune), la trachée (en vert) et l'oesophage (en bleu), sur une grille discrète $M \times N$, avec $M = 320$ et $N = 240$. La segmentation prédite de l'aorte exhibe une seule composante connexe tandis que la vérité terrain (b) en montre deux. La déformation obtenue (c) permet de conserver non seulement les relations contextuelles entre les organes, mais également la topologie de y. Par conséquent, la pénalisation la plus forte se situe au point de contact entre les deux composantes souhaitées de l'aorte (e).

Ces éléments laissent présager que le modèle conjoint proposé dans la section 4.2 améliore la précision des segmentations d'images médicales d'un point de vue quantitatif, mais également d'un point de vue qualitatif, avec notamment des segmentations plus fidèles à

la réalité anatomique. Pour le vérifier, nous présentons maintenant les résultats obtenus avec ce modèle pour la segmentation de deux jeux de données.

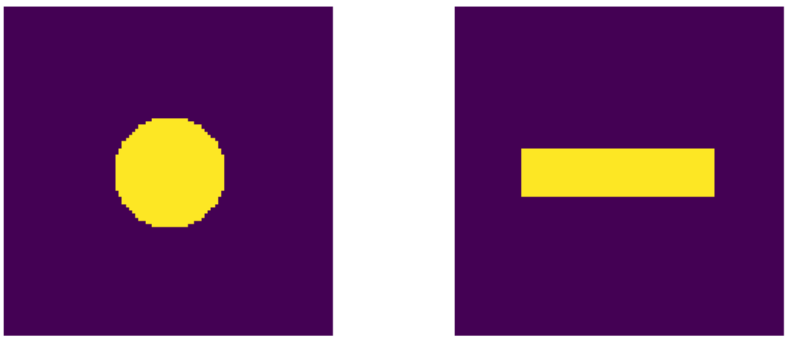
4.4.2 Cas binaire

La deuxième tâche du challenge Decathlon [2] propose de segmenter uniquement une petite partie du cœur. En effet, dans ce cas, la région d'intérêt est l'atrium gauche. L'ensemble de données est constitué de 20 images IRM 3D mono-modales du cœur entier acquises pendant une seule phase cardiaque. Chaque image est accompagnée de sa vérité terrain segmentée manuellement. Comme le souligne les auteurs dans [4], la segmentation de l'atrium se révèle particulièrement difficile de par le peu de données d'entraînement disponibles qui montrent en plus une grande variabilité anatomique, un nombre de composantes connexes qui diffère selon la coupe, allant de 1 à 3, et une absence de contours de l'organe à segmenter. Ces difficultés sont mises en exergue sur la Figure 4.8 qui affiche les coupes 43, 54 et 99 sur 120 d'un même patient, ainsi que les contours des segmentations manuelles de l'atrium pour chacune d'elles. En particulier, nous discernons la variabilité de forme, de volume et du nombre de composantes connexes de l'atrium, en plus de l'absence des contours. De plus, cette figure rend compte du déséquilibre de classes existant entre le fond et l'objet d'intérêt à segmenter. Pour toutes ces raisons, le processus de segmentation automatique aspire à bénéficier de notre modèle conjoint afin de mieux détecter l'atrium.

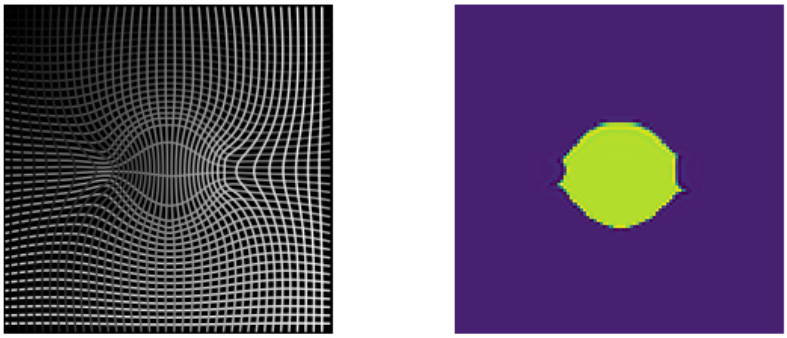
Pour évaluer ce potentiel bénéfice nous divisons aléatoirement cet ensemble de données en 14 patients pour l'entraînement et les 6 derniers sont conservés pour la phase d'inférence. Dans cette série d'expériences, aucune procédure d'augmentation de données n'est utilisée afin de mesurer l'impact du modèle dans ce contexte difficile de données limitées. Les images, de taille 320×320, sont simplement normalisées. Nous entraînons le réseau sU-Net (décrit au Chapitre 1) dont les paramètres θ sont aléatoirement initialisés grâce à la technique d'initialisation de Glorot [55] et optimisés par une technique de SGD. Les hyperparamètres du réseau sont définis en suivant une technique de quadrillage, en particulier nous fixons le taux d'apprentissage initial à 10^{-3} et la taille du lot à 24. Ce taux est divisé par un facteur dix lorsque l'entraînement ne progresse plus durant 10 époques successives. Les hyperparamètres relatifs au modèle de recalage fixés pour les simulations dans la section 4.4.1 sont conservés, à savoir $\gamma_1 = 10000$, $\gamma_2 = 80000$, $\nu = 1$ et $\mu = 125$. L'Algorithme 5 est itéré sur un maximum de 200 époques.

La Table 4.1 détaille les différents résultats quantitatifs de segmentation. Nous comparons les résultats obtenus en optimisant sU-Net d'une part seulement avec une fonction de perte $\mathcal{F}_{DL}$ de Dice multi-classes, et d'autre part en le combinant avec $\mathcal{F}_{Rec}$ et la contrainte d'incompressibilité sur le déterminant. En supplément du score de Dice et de la distance de Hausdorff, l'erreur absolue moyenne (MAE, *Mean Absolute Error*) du nombre de compo-

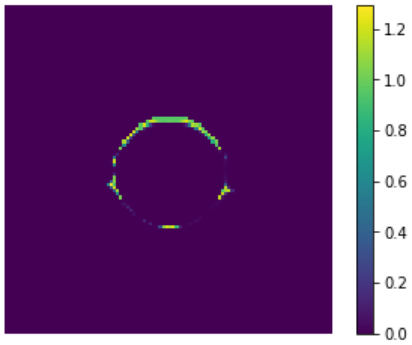
2. `https://decathlon-10.grand-challenge.org/`



(a) Référence R (b) Template T

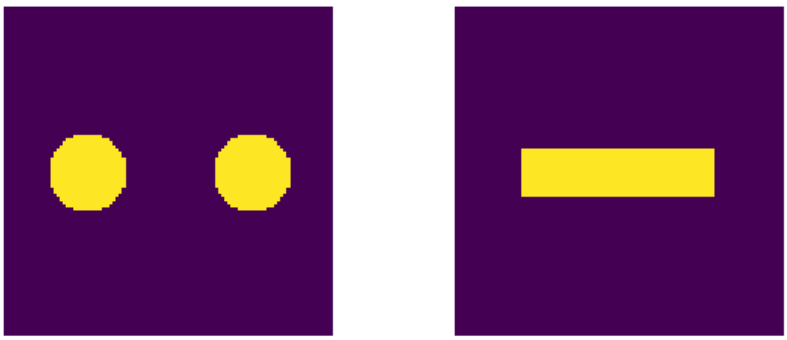


(c) Déformation φ (d) Template déformé $T \circ \varphi$

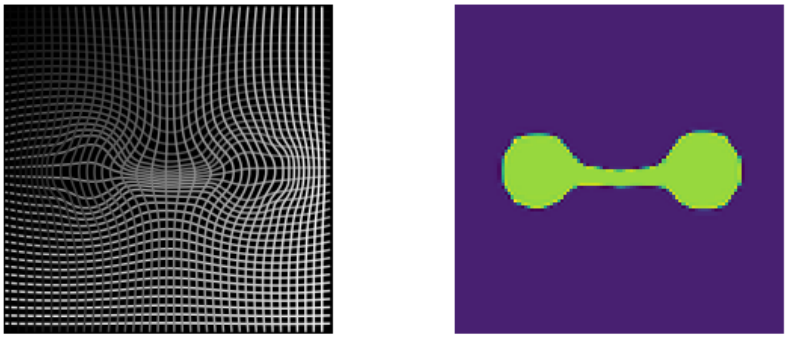


(e) $\|R - T \circ \varphi\|^2_{L^2(\Omega)}$

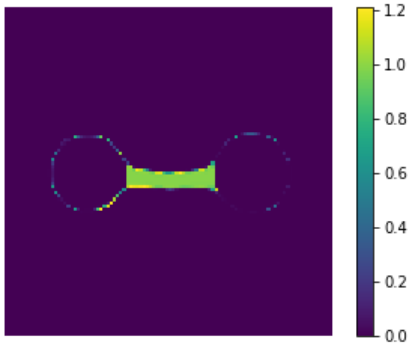
FIGURE 4.5 – Exemple n°1 : recalage d'un rectangle sur un disque.



(a) Référence R (b) Template T



(c) Déformation φ (d) Template déformé $T \circ \varphi$

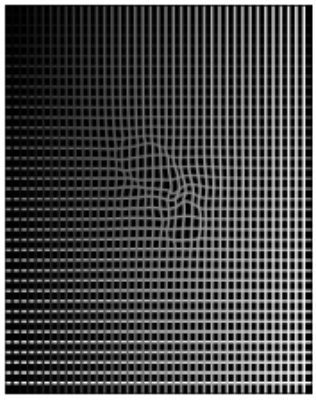


(e) $\|R - T \circ \varphi\|^2_{L^2(\Omega)}$

FIGURE 4.6 – Exemple n°2 : recalage d'un rectangle sur deux disques

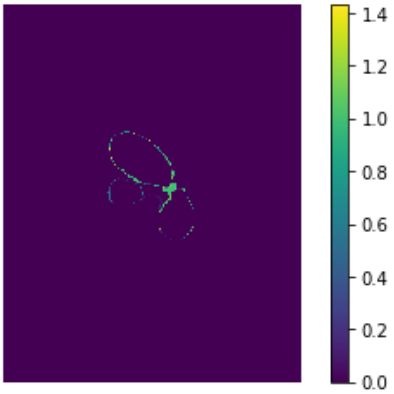
(a) Prédiction s(θ)

(b) Vérité terrain y



(c) Déformation φ

(d) Vérité terrain déformée $y \circ \varphi$



(e) $\|s(\theta) - y \circ \varphi\|^2_{L^2(\Omega)}$

FIGURE 4.7 – Exemple n°3 : recalage de la vérité terrain sur la prédiction de segmentation d'OAR du dataset SegTHOR.

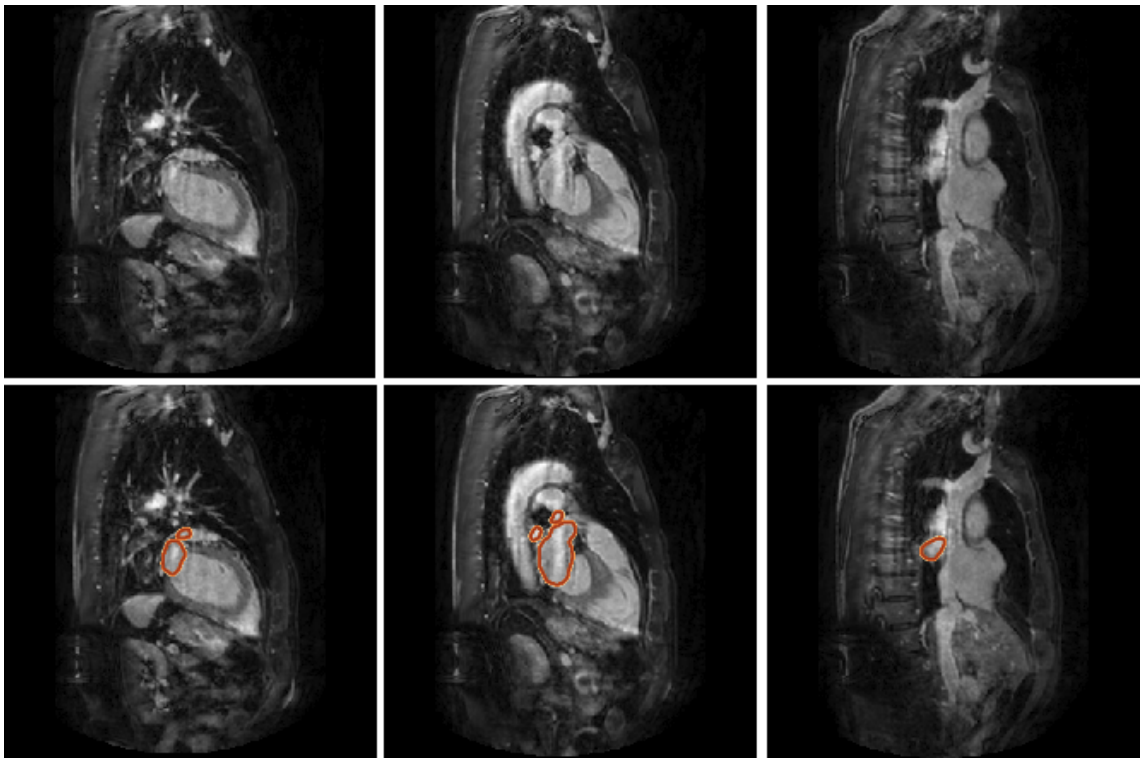


FIGURE 4.8 – Coupes axiales d'images IRM d'un même patient sans (en haut) et avec les contours de segmentations manuelles de l'atrium (en bas).

santes connexes (CC) est calculée pour mesurer le respect de la topologie. Notre approche hybride variationnelle-DL améliore de 3 pp le score de Dice, réduit de moitié la moyenne des HD et diminue également la MAE du nombre de CC. De plus, si l'importance du terme de pondération ν augmente, ces résultats s'améliorent un peu plus. Les résultats qualitatifs présentés sur la Figure 4.9 étayent cette analyse. En effet, les segmentations automatiques, dont les contours figurent en rouge, témoignent du bénéfice apporté par notre approche conjointe (b-c) par rapport à une méthode classique d'apprentissage profond (a). Le modèle proposé produit des contours davantage alignés avec ceux de la vérité terrain et corrige donc les erreurs de topologie et de pixels isolés. Par ailleurs, il n'est pas toujours aisé de visualiser nettement les différences entre les deux modèles. Ainsi le score de Dice et la distance de Hausdorff indiqués pour chaque segmentation 2D permettent de quantifier les améliorations perçues sur la Figure 4.9.

TABLE 4.1 – Résultats de segmentation (moyenne±écart-type) obtenus avec une fonction de perte de Dice $\mathcal{F}_{DL}$ et avec notre proposition. $\mathcal{F}$ est la fonction de perte. Métriques utilisées : score de Dice, distance de Hausdorff (HD) et erreur absolue moyenne (MAE) du nombre de composantes connexes (CC). Les meilleurs résultats sont en gras.

$\mathcal{F}$	Métriques	Atrium
Dice	Dice score %	83.77 ± 6.16
	HD (mm)	36.18 ± 9.85
	MAE nombre CC	0,27 ± 0,11
Dice + $\mathcal{F}_{Rec}$ avec cont. (ν=1)	Dice score %	86.71 ± 4.85
	HD (mm)	18.40 ± 5.42
	MAE nombre CC	0,21 ± 0,08
Dice +$\mathcal{F}_{Rec}$ avec cont. (ν=2)	Dice score %	**87.33 ± 4.64**
	HD (mm)	**15.40 ± 5.49**
	MAE nombre CC	**0,19 ± 0,18**

Le coût de calcul est évidemment plus élevé lorsque nous intégrons l'Algorithme 6 dans le processus de segmentation mais seulement lors de la phase d'entraînement. De plus, comme l'illustre la Figure 4.10, la vitesse de convergence compense ce coût et il serait alors possible de restreindre le nombre d'époques de l'Algorithme 5.

Finalement, notre fonction de perte bâtie sur une approche hybride fournit des résultats améliorés dans ce contexte de segmentation binaire : aussi bien quantitativement, pour l'ensemble des métriques utilisées, que qualitativement puisqu'il en résulte des segmentations plus fidèles à la réalité anatomique. Le prochain cas abordé est celui de la segmentation multi-classes des organes à risque du dataset SegTHOR.

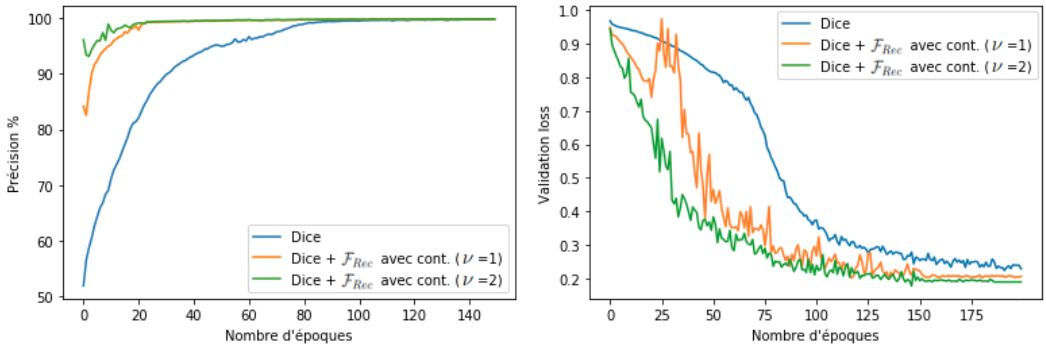


FIGURE 4.10 – Évolution de la précision et du score de Dice sur l'ensemble de validation au fil des époques, pour la fonction de perte composée du Dice seulement (lignes bleues) et combinée à $\mathcal{F}_{Rec}$ avec contrainte (lignes oranges et vertes)

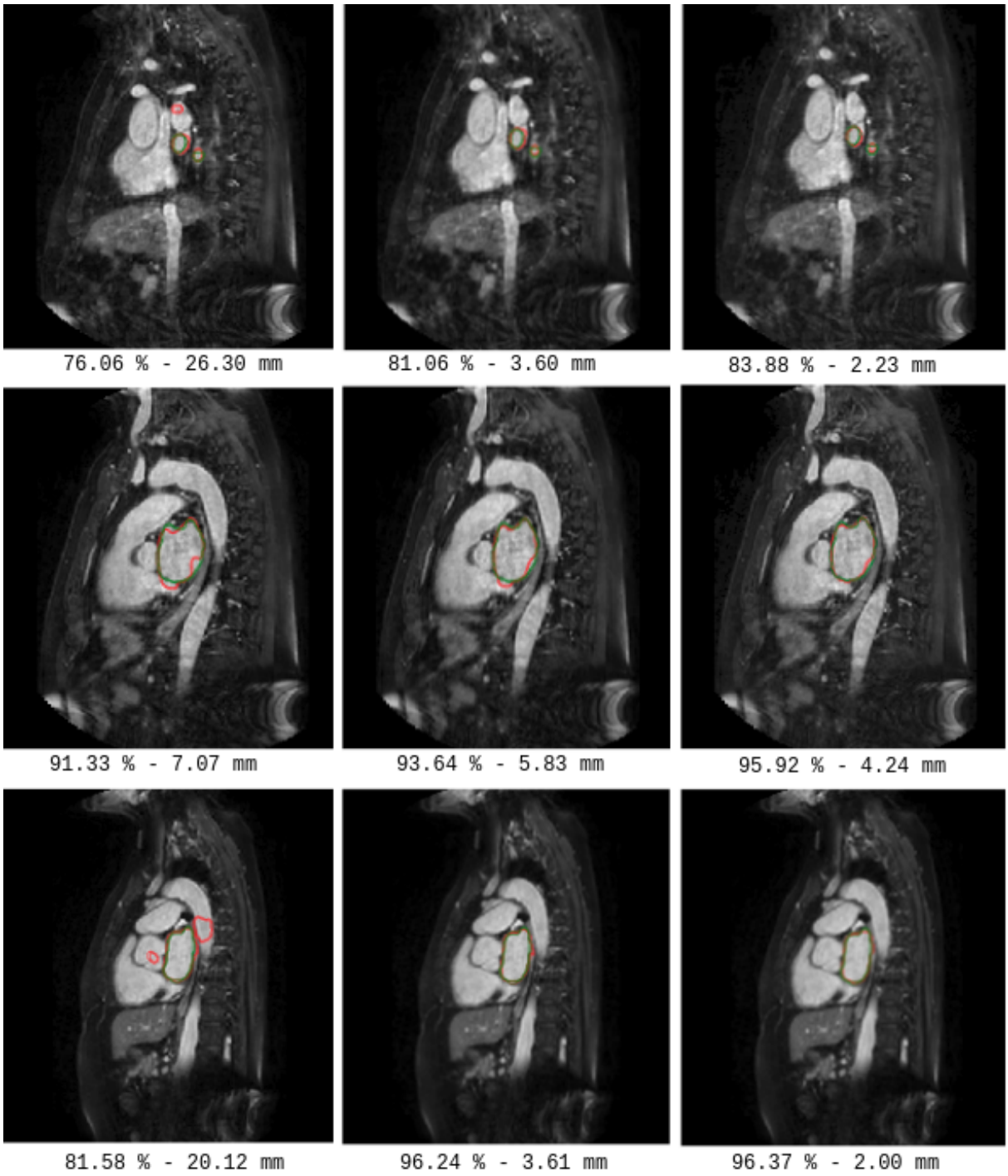


FIGURE 4.9 – Résultats de segmentation pour six patients de l'atrium (en rouge) obtenus avec une fonction de perte de Dice (a) et avec notre proposition (b-c). Les contours de la GT apparaissent en vert. Le score de Dice et la distance de Hausdorff (HD) sont indiqués pour chaque segmentation.

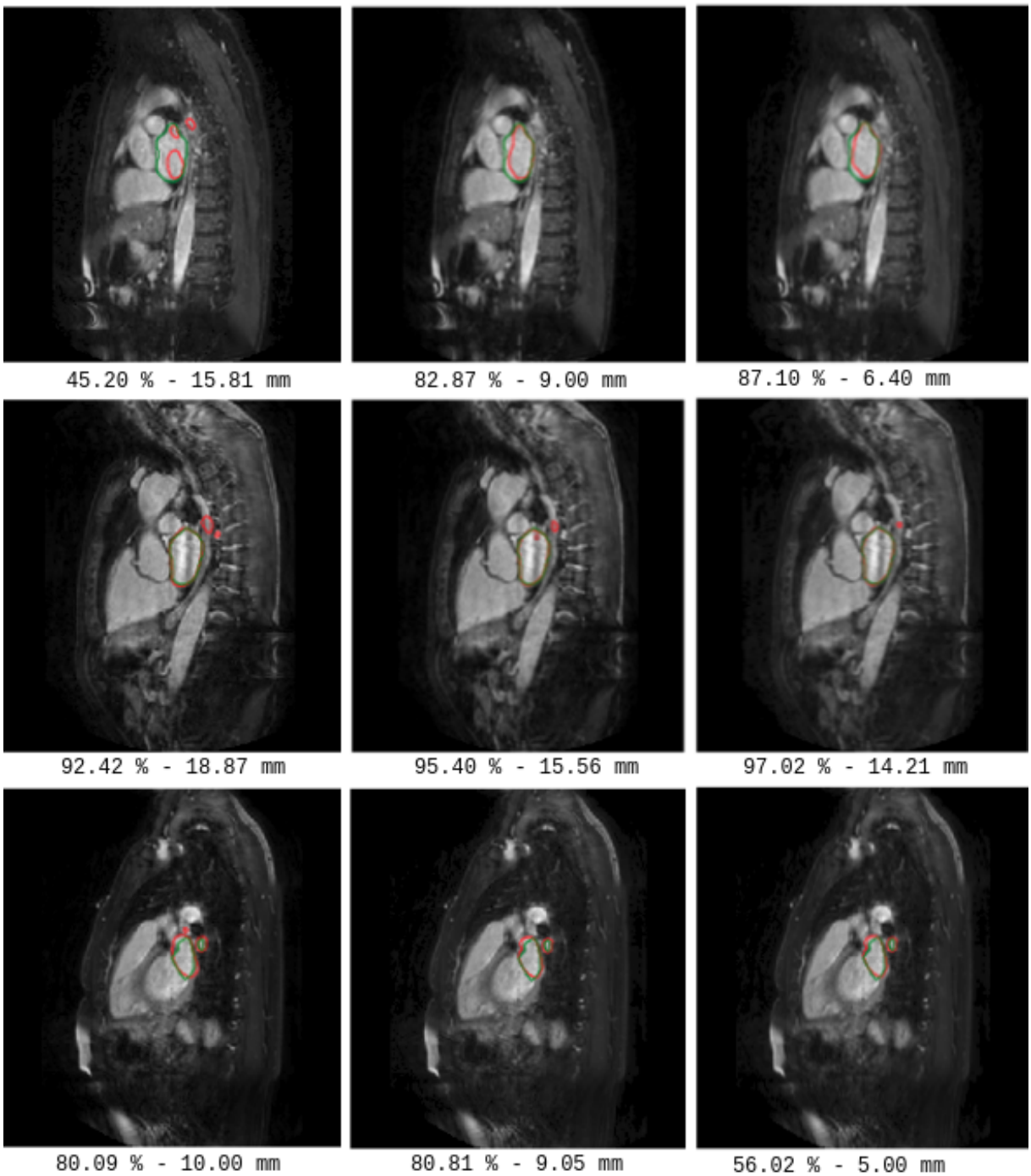


FIGURE 4.9 – Résultats de segmentation pour six patients de l'atrium (en rouge) obtenus avec une fonction de perte de Dice (a) et avec notre proposition (b-c). Les contours de la GT apparaissent en vert. Le score de Dice et la distance de Hausdorff (HD) sont indiqués pour chaque segmentation (suite).

4.4.3 Cas multi-classes

Notre nouvelle fonction de perte est maintenant évaluée dans un contexte de segmentation multi-classes. Pour tous les arguments évoqués dans les Chapitres 1 et 3, à savoir le déséquilibre des classes, la variabilité inter- et intra-patients, l'absence de contours ou encore la topologie dépendante de la coupe, notre jeu de données SegTHOR semble également revêtir tous les verrous motivant l'introduction de notre modèle. Pour le vérifier, le même protocole que celui établi lors du challenge s'applique, c'est-à-dire 40 patients conservés pour apprendre les paramètres du réseau sU-Net et 20 patients dédiés à la phase de test. Nous utilisons également la même procédure d'augmentation de données, décrite dans les chapitres précédents, pour augmenter artificiellement la taille de l'ensemble de données d'entraînement. Les paramètres θ du réseau sont à nouveau initialisés aléatoirement grâce à la technique de Glorot [55] et mis à jour par une descente stochastique de gradient (SGD). Les hyperparamètres du réseau sont optimisés en suivant une technique de quadrillage, en particulier nous fixons le taux d'apprentissage initial η à 10^{-2} et la taille du lot à 24. Ce taux est divisé par un facteur dix lorsque l'entraînement ne progresse plus durant 10 époques successives. Concernant les hyperparamètres du problème de recalage, nous gardons les mêmes valeurs que celles fixées lors des simulations numériques, hormis le terme de pondération ν pour lequel plusieurs valeurs sont expérimentées. Enfin, l'entraînement de l'approche hybride suivant l'Algorithme 5 se déroule sur un maximum de 100 époques, tandis que dans les résultats détaillés ensuite, l'entraînement de l'approche DL classique s'effectue avec 120 époques maximum. Dans toutes les expériences, la fonction $\mathcal{F}_{DL}$ choisie est le Dice multi-classes.

Les segmentations automatiques obtenues avec notre méthode, qui imbrique donc les deux formalismes, peuvent être analysées à la lumière de celles obtenues via l'approche classique par apprentissage profond. La Table 4.2 présente les résultats quantitatifs de segmentation des 4 organes à risque du dataset SegTHOR, et la Figure 4.11 les résultats qualitatifs. En particulier, nous indiquons ceux obtenus avec deux valeurs différentes de ν. Au regard de ces éléments, nous discutons de chaque OAR.
Comme explicité dans les premiers chapitres, la trachée (en vert clair sur la Figure 4.11) apparaît en noir sur l'imagerie scanner, ce qui la rend aisément distinguable. Les scores de Dice et les distances de Hausdorff sont très similaires pour les deux méthodologies. Cela se traduit visuellement par des formes pratiquement identiques. Pour l'aorte (en jaune sur la Figure 4.11), l'ajout de la fonctionnelle $\mathcal{F}_{Rec}$ et de la contrainte à la fonction objectif permet de gagner environ 1 pp de score de Dice. De surcroît, la distance de Hausdorff passe de 13.47 mm à 10.71 mm lorsque ν vaut 1, et 8.95 pour ν égal à 6. Cette diminution s'explique notamment par un meilleur respect de la topologie comme l'illustre par exemple la troisième ligne de la Figure 4.11. Concernant le score de Dice pour l'œsophage (en bleu sur la Figure 4.11) et le cœur (en vert foncé sur la Figure 4.11), le constat se révèle équivalent à celui émis dans le cas de l'aorte avec un gain d'environ 1 pp pour ces deux OAR. En revanche, notre proposition améliore considérablement la métrique HD en la

divisant approximativement par 2 pour l'œsophage et 2,5 pour le cœur. Pour ces deux organes, nous observons un effet de sur-segmentation avec la fonction de perte de Dice (c), dû aux valeurs aberrantes et à des groupes de pixels isolés, que notre méthode (d-e) tend à corriger. Finalement, si l'augmentation du terme de pondération ν donne globalement des scores de Dice similaires, elle participe néanmoins à diminuer un peu plus les distances de Hausdorff en rectifiant quelques anomalies subsistantes.

TABLE 4.2 – Résultats de segmentation (moyenne±écart-type) obtenus avec une fonction de perte de Dice $\mathcal{F}_{DL}$ et avec notre proposition. $\mathcal{F}$ est la fonction de perte. Métriques utilisées : score de Dice et distance de Hausdorff (HD). Les meilleurs résultats sont en gras.

$\mathcal{F}$	Métriques	Aorte	Trachée	Œsophage	Cœur
Dice	Dice score %	93.36 ± 1.62	90.60 ± 1.90	81.66 ± 5.45	92.30 ± 3.61
	HD (mm)	13.47 ± 8.55	18.75 ± 9.73	31.45 ± 19.02	45.23 ± 40.18
Dice + $\mathcal{F}_{Rec}$ avec cont. (ν=1)	Dice score %	**94.61 ± 1.62**	90.69 ± 2.64	82.26 ± 4.98	**93.14 ± 3.04**
	HD (mm)	10.71 ± 7.36	17.96 ± 9.32	18.89 ± 12.44	18.99± 8.51
Dice + $\mathcal{F}_{Rec}$ avec cont. (ν=6)	Dice score %	94.51 ± 1.88	91.26 ± 2.03	**82.56 ± 4.38**	93.05 ± 3.28
	HD (mm)	**8.95 ± 5.43**	17.85 ± 6.48	**15.83 ± 8.17**	**17.98 ± 7.96**

À noter que la remarque concernant le coût de calcul et la vitesse de convergence tient aussi dans ce cas multi-classes. Un reproche majeur dans l'apprentissage profond émane du manque d'interprétabilité des CNN. Ainsi, pour tenter de comprendre leur succès, Vinogradova *et al.* proposent Seg-Grad-Cam [139], une méthode fondée sur le gradient pour interpréter la segmentation sémantique. Appliquée localement, elle produit des cartes thermiques qui quantifient la pertinence de chaque pixel individuellement. Pour ce faire, il faut sélectionner les K cartes de caractéristiques d'intérêt $\{A^k\}_{k=1}^K$ d'une couche connectée. La méthode consiste alors à calculer la moyenne des gradients de $\sum_{i,j\in\mathcal{M}} s^l(\theta)_{i,j}$ par rapport aux P pixels (indexés par u, v) de chaque carte de caractéristiques A^k afin de produire une pondération α_k^l. Ainsi, cette pondération exprime l'importance de chaque carte de caractéristiques pour une classe l donnée. L'ensemble $\mathcal{M}$ peut représenter un seul pixel, un groupe de pixels ou tous les pixels de l'image. Finalement, une carte thermique H^l se calcule comme

$$H^l = ReLU\left(\sum_k \alpha_l^k A^k\right) \text{ avec } \alpha_l^k = \frac{1}{P}\sum_{u,v} \frac{\partial \sum_{i,j\in\mathcal{M}} s^l(\theta)_{i,j}}{\partial A_{u,v}^k}.$$

Par curiosité et pour essayer de comprendre d'où proviennent les différences entre les segmentations, nous avons expérimenté cette méthode pour le CNN optimisé uniquement avec la fonction de perte de Dice et celui optimisé avec notre proposition. En particulier, l'ensemble $\mathcal{M}$ compte tous les pixels de l'image dans cette expérience. L'exemple présenté sur la Figure 4.12 montre les résultats de la méthode dans le cas de l'œsophage (en bleu sur la GT (b)) au niveau quatrième bloc de convolution (Conv4), couramment appelé

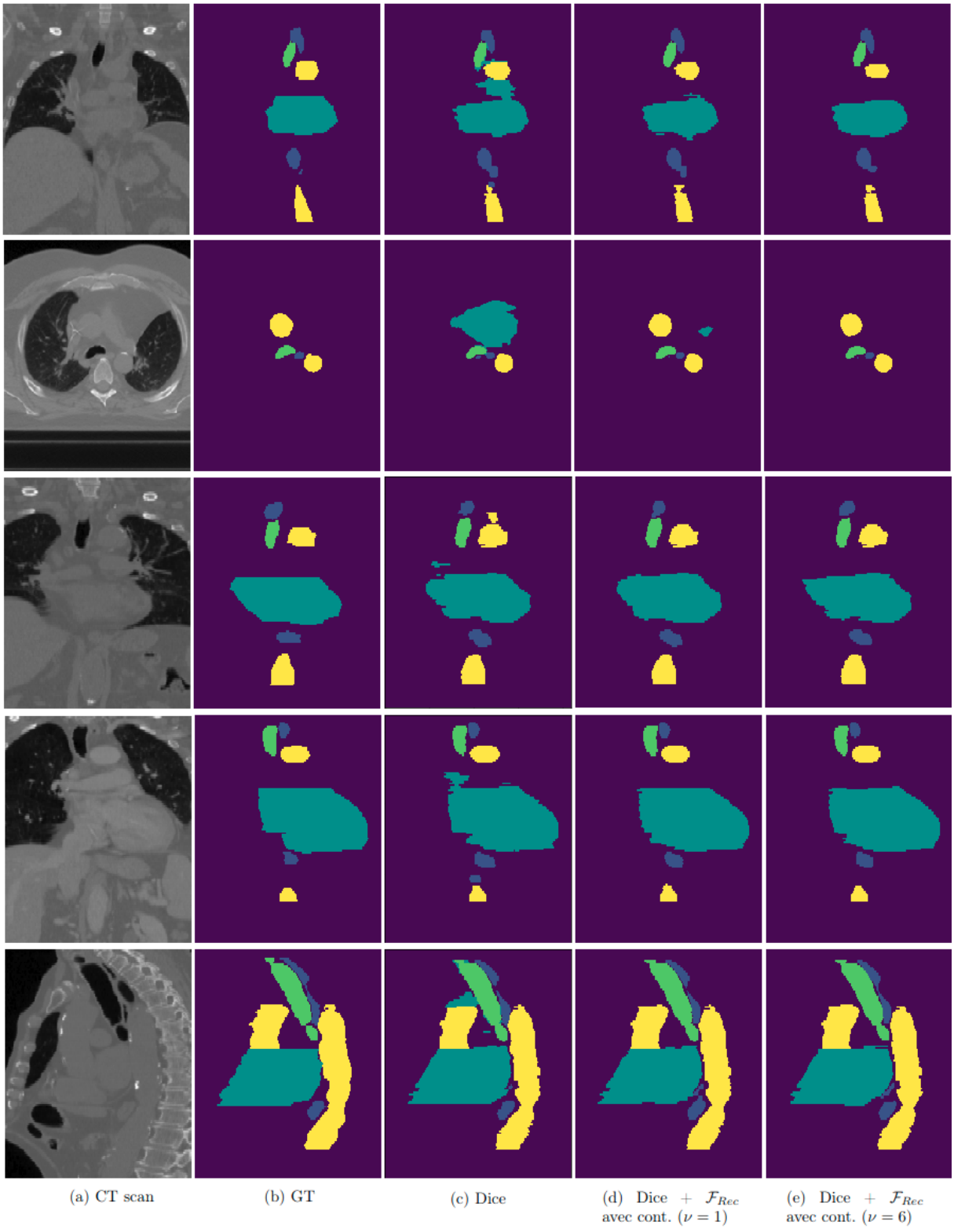
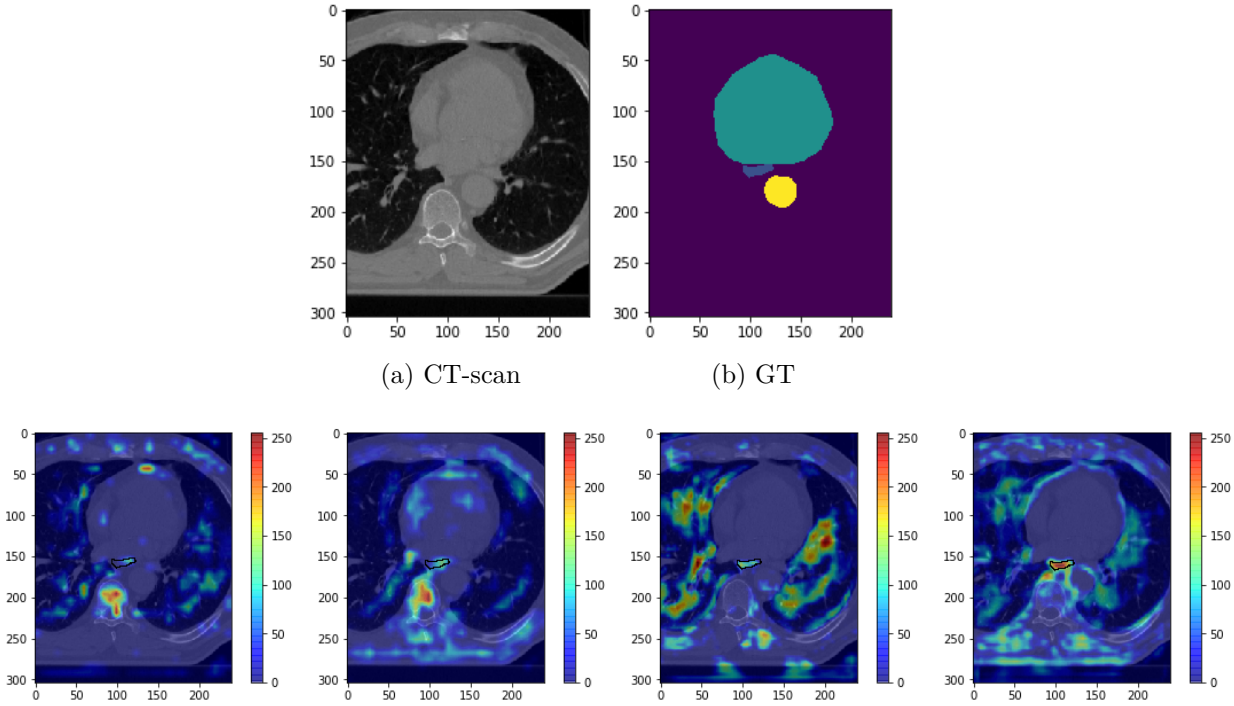


FIGURE 4.11 – Résultats de segmentation pour cinq patients de l'aorte (jaune), la trachée (vert clair), le cœur (vert foncé) et l'œsophage (bleu) obtenus avec une fonction de perte de Dice (c) et avec notre proposition (d-e).

bottleneck, et du cinquième bloc de convolution (Conv5). Dans cet exemple, les contours de l'organe apparaissent éminemment flous sur l'image de scanner (a), ce qui le rend d'autant plus difficile à délimiter. Pour le Conv4, notre proposition (d) donne un peu plus d'importance aux pixels localisés à proximité de la zone de l'œsophage en comparaison de la version standard (c). Concernant le bloc de convolution suivant, les pixels avec l'importance la plus élevée se situent exactement sur l'œsophage avec notre proposition (f), tandis que dans l'autre cas (e), l'importance donnée à l'œsophage est moindre, et ce sont même les pixels du fond auxquels la méthode accorde le plus de pertinence. Cette étude nous amène à penser que notre fonction de perte régularisée et contrainte se concentre davantage sur les pixels étiquetés comme objets d'intérêt que sur ceux du fond.



(a) CT-scan (b) GT

(c) Conv4 - sU-Net + Dice (d) Conv4 - sU-Net + notre proposition (e) Conv5 - sU-Net + Dice (f) Conv5 - sU-Net + notre proposition

FIGURE 4.12 – Résultats de la méthode Seg-Grad-Cam appliquée à sU-Net, optimisé avec la fonction de perte de Dice et avec notre proposition, pour interpréter la segmentation de l'œsophage (en bleu foncé sur la GT). Les cartes thermiques sont obtenues à la sortie des blocs Conv4 (c-d) et Conv5 (e-f). Les contours de l'œsophage sont superposés en noir.

4.5 Conclusion

Dans ce chapitre, nous avons introduit une nouvelle fonction de perte qui intègre des connaissances géométriques et topologiques dans l'entraînement d'un CNN 2D. Cette

fonction est bâtie sur l'association d'un terme de Dice et d'une fonctionnelle modélisant un problème de recalage. En effet, ici le recalage prend la forme d'un *a priori* pour guider la segmentation vers des résultats plus réalistes. Plus spécifiquement, il permet d'obtenir une version de la segmentation prédite homéomorphe à la vérité terrain. En plus, une contrainte d'égalité sur le déterminant de la matrice jacobienne assure la préservation de la topologie et du volume. De cette manière, les ensembles de points responsables des erreurs topologiques (nombre de composantes connexes, pixels isolés, trous) font l'objet d'une forte pénalisation durant le processus d'apprentissage via un couplage L^2. Un autre intérêt du recalage réside dans sa capacité à fournir de grandes déformations mais qui tiennent compte des relations contextuelles entre les objets d'intérêt présents dans l'image.

Le problème de minimisation, formulé dans un cadre hybride DL-variationnel, est fractionné en deux sous-problèmes de sorte à être résolu relativement à chaque variable de façon alternée. Le premier sous-problème lisse mais non convexe aborde l'optimisation des paramètres θ du réseau, accomplie par une descente de gradient. Le second se formule donc comme un problème de recalage sous contrainte. Des résultats théoriques en montrent le caractère bien posé. Cependant, il est non lisse et non convexe et de fait, exhibe des difficultés numériques. En s'appuyant sur des principes d'élasticité non linéaire, nous avons proposé une stratégie efficace de résolution grâce à l'introduction de variables auxiliaires et une technique de *splitting*. Par ailleurs, une implémentation adéquate est détaillée afin de rendre réalisable l'ensemble du processus imbriquant les deux formalismes.

Pour estimer les bénéfices apportés, nous évaluons notre méthode d'abord sur des images IRM cardiaques pour contourer l'atrium gauche, une petite partie du coeur. Plusieurs difficultés émergent de cet ensemble de données : très peu d'échantillons d'entraînement, frontières de l'atrium invisibles, existence d'une forte variabilité de forme et de taille ainsi qu'une topologie dépendante de la coupe. La seconde salve d'expériences concerne la délimitation des organes à risque des scanners du dataset SegTHOR qui présente des difficultés similaires. Pour ces raisons, ces deux tâches de segmentation prétendent à bénéficier de l'approche conjointe proposée, introduite dans l'entraînement d'un CNN basique. Les résultats quantitatifs et qualitatifs attestent de segmentations plus précises par rapport à une approche DL classique, aussi bien dans le cas binaire de l'atrium que dans le cas multi-classes des OAR thoraciques. En effet, si les scores de Dice gagnent quelques points, les distances de Hausdorff chutent considérablement. Cela démontre que notre proposition tend à respecter la topologie attendue en corrigeant les valeurs aberrantes. Ce constat se vérifie visuellement puisque les résultats obtenus sont plus fidèles à la réalité anatomique.

L'ensemble de ces analyses nous conforte dans l'intérêt d'un cadre hybride variationnel-DL pour opérer un contrôle sur la topologie, et notamment à l'aide d'un traitement par recalage, pour aboutir à des segmentations automatiques perfectionnées. De plus, cette méthodologie offre l'avantage de s'étendre directement à tous les réseaux de neurones et toutes les applications. Une suite envisagée de ces travaux s'orienterait vers l'adaptation du modèle à la phase d'inférence, par exemple à partir d'un Template moyen à déformer.

CHAPITRE

5 Conclusion et perspectives

Dans cette thèse, nous avons étudié le problème de l'inclusion de contraintes géométriques et topologiques pour la segmentation automatique d'images médicales. Celui-ci est abordé dans un cadre hybride mêlant approches variationnelles et par apprentissage profond. L'intérêt est de tirer profit des deux formalismes. D'une part, jouir des bons résultats quantitatifs réalisés par les réseaux de neurones convolutifs ainsi que de leur capacité de généralisation. D'autre part, bénéficier du cadre continu des méthodes variationnelles, qui de surcroît s'appuient sur des concepts mathématiques clairs, et surtout permettent l'incorporation de connaissances préalables de nature géométrique et topologique sur les solutions à reconstruire, comme en témoigne la littérature prolifique à ce sujet.
Initialement, notre idée consistait à intégrer ces connaissances pour la détection automatique d'organes à risque thoraciques en prévision d'un traitement par radiothérapie, et en ce sens, le fil rouge de nos travaux est l'application SegTHOR. Dans cette optique, notre travail a en premier lieu porté sur l'architecture employée et nous avons développé le réseau de neurones basique sU-Net, une variante simplifiée du réseau réputé U-Net. Le but étant d'établir des résultats de segmentation de référence pour évaluer de façon objective la significativité des contributions. Puis, pour contraindre ce réseau à satisfaire certains critères géométriques ou topologiques, et ainsi obtenir des segmentations plus conformes à la réalité anatomique, nous avons conçu deux nouvelles fonctions de perte modélisées dans ce cadre hybride et dont l'optimisation s'opère par le biais d'un schéma alternatif.

Ainsi, nous avons proposé un premier modèle dont la fonctionnelle à optimiser est formulée comme un terme d'attache aux données qui prend la forme du modèle constant par morceaux de Mumford-Shah pour assurer l'homogénéité des pixels, une régularisation géométrique bâtie sur la variation totale pondérée qui encourage l'alignement des contours, et une pénalisation de la surface, modélisée par une contrainte d'égalité dure. Des résultats théoriques garantissant le caractère bien posé du modèle et des expériences soulignant l'apport de ces contraintes sont fournis. De nouvelles perspectives sont également envisagées pour étendre ce travail.
Une première porte sur la généralisation de la méthode à des cadres faiblement et semi-supervisés. Dans ce contexte, la contrainte d'égalité portant sur la surface des objets

d'intérêt à segmenter ne tient plus et il faudrait plutôt envisager des contraintes d'inégalité établies à partir d'une étude statistique menée préalablement sur ces objets.
Une seconde piste consisterait à traiter des *a priori* de formes, comme par exemple la convexité dans une approche hybride similaire à ce qui a été exposé dans le Chapitre 3. En effet, ce type de contrainte apparaît pertinent du fait de la nature généralement convexe des organes à segmenter. À la manière de [89], une variable auxiliaire binaire u^l représenterait l'enveloppe convexe de la segmentation prédite $s^l(\theta)$. De cette façon, la méthode pénaliserait les valeurs aberrantes, les trous et les ensembles de points qui engendrent la non-convexité (des excroissances par exemple), menant ainsi à des résultats de segmentation corrigés.

Ensuite, nous avons développé un modèle conjoint recalage/segmentation pour introduire des prescriptions topologiques et géométriques. Dans ce contexte, la tâche de recalage est appréhendée comme une information préalable introduite dans le processus de segmentation où nous cherchons à construire une solution homéomorphe à un *a priori* connu. Ainsi, le modèle est constitué d'un terme d'attache aux données prenant la forme d'une pénalisation L^2 entre la segmentation prédite et cette solution. Ensuite, une régularisation sur la transformation, bâtie sur des principes d'élasticité non linéaire, s'apparente à une fonction de densité d'énergie et pénalise les changements de longueurs. Enfin, une condition d'incompressibilité qui se traduit par le fait de contraindre le déterminant de la matrice jacobienne de la transformation pour garantir la préservation du volume et de la topologie, sans auto-intersection de matière. Ce problème exhibe de bonnes propriétés soulignées par des résultats théoriques tels que l'existence de minimiseurs. De plus, des expériences valident l'efficacité du modèle, et en particulier la conformité à la topologie ciblée dans la plupart des cas.
Comme précédemment évoqué, il serait intéressant d'utiliser le modèle également en phase d'inférence, et par extension dans un cadre faiblement/semi-supervisé. Pour ce faire, nous pourrions construire un Template moyen à recaler sur la segmentation prédite pour obtenir le résultat final. Dans ce cas, puisque nous cherchons uniquement à préserver la relation contextuelle entre les objets, la contrainte d'égalité sur le déterminant doit être relâchée en la contrainte d'inégalité $\det \nabla\varphi > 0$, préservant toujours la topologie. Cependant, une difficulté provient du fait que la topologie dépend de la coupe, par exemple l'aorte exhibe une ou deux composantes connexes en raison de sa forme de canne, et il faudrait alors élaborer une stratégie afin d'associer le bon Template à chaque image 2D.
Pour contourner cette difficulté, le modèle pourrait être étendu au cadre 3D. Le défi scientifique tient principalement à l'implémentation numérique qui devient plus délicate en raison d'une expression plus complexe de la densité d'énergie faisant apparaître la matrice des cofacteurs de la matrice jacobienne $\nabla\varphi$ (contrôle des variations d'aire).
Une voie différente à explorer consisterait à insérer la tâche de recalage directement dans le réseau profond, notamment en s'inspirant des travaux [36], et [142] qui utilisent des couches connectées dites difféomorphiques, leur permettant de garantir systématiquement que la topologie ciblée est préservée. Dans ces travaux, les auteurs optimisent les para-

mètres de la déformation par le biais d'un réseau de transformation spatiale ([65,80]) qu'il serait intéressant d'exploiter.

Dans un autre registre, une tendance qui émerge et prend de plus en plus d'ampleur depuis quelques années s'articule autour des réseaux de neurones de type Transformers. Ils attisent la curiosité en imagerie de par leur capacité à apprendre les relations globales entre toutes les régions au sein d'une image grâce à leur module d'Attention [134]. Ainsi, la différence avec un CNN provient principalement du fait qu'un bloc convolutionnel a un champ réceptif limité (par la taille du filtre) alors que dans un Transformer, le champ réceptif est l'image entière.
Par exemple, pour la classification des images du dataset ImageNet, le réseau ViT [40] présente des scores dépassant ceux des CNN. Concernant la segmentation, une promesse vient des réseaux hybridant U-Net et Transformer, entre autres TransUnet [27], CATS [82] ou UNETR [59] obtiennent des résultats de segmentation d'images médicales particulièrement précis.
Ainsi, une piste envisagée serait de mettre en place un réseau de ce type avec l'objectif, d'une part de comparer les résultats avec ceux de sU-Net optimisé via nos modèles, et d'autre part, les intégrer pour entraîner ce nouveau réseau.

Finalement, dans une perspective un peu plus générale, une nouvelle tendance qui émerge au sein de la communauté Deep-Learning, et en particulier dans le contexte de l'analyse d'images médicales, consiste à mieux maîtriser les diverses ressources disponibles. En effet, un axe de recherche s'articule autour du développement de modèles d'apprentissage profond efficients dont l'entraînement s'effectue, par exemple, à partir d'un nombre limité de données d'entraînement, d'un nombre restreint d'images labellisées ou en réduisant les ressources matérielles engagées.
L'introduction de connaissances préalables semble représenter un bon moyen de limiter la taille du dataset d'entraînement. En effet, dans le Chapitre 4, nous avons vu que la méthode proposée permet d'obtenir de bons résultats de segmentation pour l'atrium malgré un ensemble de données limité. Dans de futures expériences, il serait opportun de valider avec d'autres datasets, présentant cette difficulté, que les différents *a priori* étudiés permettent d'aboutir à des segmentations précises.
Concernant la réduction des images labellisées, d'autant plus intéressante que leur collecte s'avère compliquée dans le domaine de la santé, nous avons avancé plusieurs pistes dans les paragraphes précédents pour procéder à un apprentissage profond semi-supervisé (relaxation de contraintes, utilisation d'un Template moyen à recaler etc.).
Enfin, pour maîtriser les ressources matérielles, différents sujets sont explorés par les chercheurs tels que la recherche de nouvelles architectures neuronales plus légères, la compression des architectures etc.. Cette volonté d'une architecture simple était déjà présente lors de l'implémentation du réseau sU-Net, mais nous pourrions essayer de le simplifier encore et évaluer nos fonctions de perte dans cette configuration.

Bibliographie

[1] O. Alexandrov and F. Santosa. A topology-preserving level set method for shape optimization. J. Comput. Phys., 204(1) :121 – 130, 2005. 32

[2] G. Allaire. Analyse numérique et optimisation : une introduction à la modélisation mathématique et à la simulation numérique. Editions Ecole Polytechnique, 2005. 130

[3] L. Ambrosio and V. M. Tortorelli. Approximation of functional depending on jumps by elliptic functional via Γ-convergence. Commun. Pure Appl. Math., 43(8) :999–1036, 1990. 22

[4] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, et al. The medical segmentation decathlon. arXiv preprint arXiv :2106.05735, 2021. 35, 139

[5] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh. Deep semantic segmentation of natural and medical images : a review. Artificial Intelligence Review, 54(1) :137–178, 2021. 23

[6] D. Azé. Éléments d'analyse convexe et variationnelle. Mathématiques pour le 2ème cycle. Ellipses, 1997. 57, 73

[7] A. Baldi. Weighted BV functions. Houston J. Math., 27(3) :683–705, 2001. 70, 71, 72

[8] J. M. Ball. Global invertibility of Sobolev functions and the interpenetration of matter. P. Roy. Soc. Edin. A, 88(3-4) :315–328, 1981. 56, 57, 108, 112, 113, 118

[9] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, Pheng-Ann Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis : is the problem solved ? IEEE Trans. Med. Imaging, 37(11) :2514–2525, 2018. 35, 63

[10] G. Bertrand. Simple points, topological numbers and geodesic neighborhoods in cubic grids. Pattern Recognit. Lett., 15(10) :1003–1011, 1994. 104

[11] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser, et al. The liver tumor segmentation benchmark (LiTS). CoRR, abs/1901.04056, 2019. 35, 63

[12] S. Bohlender, I. Oksuz, and A. Mukhopadhyay. A Survey on Shape-Constraint Deep Learning for Medical Image Segmentation. arXiv preprint arXiv :2101.07721, 2021. 32, 63, 66

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Found. Trends Mach. Learn, 3 :1–122, 2011. 74

[14] X. Bresson, S. Esedoḡlu, P. Vandergheynst, J.-P. Thiran, and S. Osher. Fast global minimization of the active contour/snake model. J. Math. Imaging Vision, 28(2) :151–167, 2007. 23, 69, 70

[15] H. Brezis. Analyse fonctionnelle. Dunod Paris, 2005. 47, 49, 112, 122

[16] M. Burger, J. Modersitzki, and L. Ruthotto. A hyperelastic regularization energy for image registration. SIAM J. Sci. Comput., 35(1) :B132–B148, 2013. 106

[17] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. Int. J. Comput. Vision, 22(1) :61–79, 1997. 21, 69

[18] A. Chambolle. An Algorithm for Total Variation Minimization and Applications. J. Math. Imaging Vis., 20(1) :89–97, 2004. 67, 72

[19] A. Chambolle, D. Cremers, and T. Pock. A convex approach to minimal partitions. SIAM J. Imag. Sci., 5(4) :1113–1158, 2012. 22

[20] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. J. Math. Imaging Vis., 40(1) :120–145, 2011. 75, 76

[21] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal-dual algorithm. Math. Program., 159(1) :253–287, 2016. 76, 86, 87, 88, 89, 90, 102

[22] A. Chambolle, P. Tan, and S. Vaiter. Accelerated Alternating Descent Methods for Dykstra-Like Problems. J. Math. Imaging Vis., 3(59) :481–497, 2017. 73

[23] H.-L. Chan, S. Yan, L.-M. Lui, and X.-C. Tai. Topology-preserving image segmentation by beltrami representation of shapes. J. Math. Imaging Vision, 60(3) :401–421, 2018. 106

[24] T. F. Chan, S. Esedoglu, and M. Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. SIAM J. Appl. Math., 66(5) :1632–1648, 2006. 23

[25] T. F. Chan and L. A. Vese. Active contours without edges. IEEE Trans. Image Process., 10(2) :266–277, 2001. 22, 23, 106

[26] T. F. Chan and L. A. Vese. Active contour and segmentation models using geometric PDE's for medical imaging. In Geometric methods in bio-medical image processing, pages 63–75. Springer, 2002. 23

[27] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, and Y. Zhou. Transunet : Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv :2102.04306, 2021. 155

[28] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams, and Y. Zheng. Learning Active Contour Models for Medical Image Segmentation. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11624–11632, 2019. 33

[29] P. G. Ciarlet. Elasticité Tridimensionnelle. Masson, 1985. 108

[30] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net : Learning Dense Volumetric Segmentation from Sparse Annotation. In MICCAI, page 424–432, 2016. 29

[31] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). arXiv preprint arXiv :1511.07289, 2015. 27

[32] J. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A Schnabel, and A. King. A Topological Loss Function for Deep-Learning based Image Segmentation using Persistent Homology. IEEE Trans. Pattern Anal. Mach. Intell., 2020. 33, 66, 105

[33] L. D. Cohen. On active contour models and balloons. CVGIP : Image Understanding, 53(2) :211–218, 1991. 21

[34] P. L. Combettes and J.-C. Pesquet. Proximal Splitting Methods in Signal Processing, pages 185–212. Springer New York, New York, NY, 2011. 57, 58, 59, 64, 75, 78

[35] B. Dacorogna. Direct Methods in the Calculus of Variations, Second Edition. Springer, 2008. 49, 51, 52, 54, 57, 111, 112, 113, 122

[36] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 729–738. Springer, 2018. 154

[37] N. Debroux, S. Ozeré, and C. Le Guyader. A non-local topology-preserving segmentation guided registration model. J. Math. Imaging Vision, pages 1–24, 2017. 106

[38] F. Demengel, G. Demengel, and R. Erné. Functional Spaces for the Theory of Elliptic Partial Differential Equations. Universitext. Springer London, 2012. 51, 52, 111

[39] J. Dolz, I. B. Ayed, and C. Desrosiers. Unbiased Shape Compactness for Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 755–763. Springer, 2017. 33, 66

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale. In International Conference on Learning Representations, 2021. 155

[41] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng. 3D Deeply Supervised Network for Automatic Liver Segmentation from CT volumes. CoRR, abs/1607.00582, 2016. 29

[42] H. Edelsbrunner and J. L. Harer. Computational topology : an introduction. American Mathematical Society, 2022. 104

[43] I. Ekeland and R. Temam. Convex Analysis and Variational Problems. Society for Industrial and Applied Mathematics, 1999. 82, 83, 86

[44] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah. Bb-unet : U-net with bounding box prior. IEEE J. Sel. Top. Signal Process., 14(6) :1189–1198, 2020. 33, 67

[45] R. El Jurdi, C. Petitjean, P. Honeine, V. Cheplygina, and F. Abdallah. High-level prior-based loss functions for medical image segmentation : A survey. Comput. Vision Image Understanding, 210 :103248, 2021. 32, 33, 66

[46] T. Estienne, M. Vakalopoulou, S. Christodoulidis, E. Battistela, M. Lerousseau, A. Carre, G. Klausner, R. Sun, C. Robert, S. Mougiakakou, N. Paragios, and E. Deutsch. U-ReSNet : Ultimate Coupling of Registration and Segmentation with Deep Nets. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pages 310–319, Cham, 2019. Springer International Publishing. 106

[47] L. C. Evans and R. F. Gariepy. Measure Theory and Fine Properties of Functions. Studies in Advanced Mathematics. Taylor & Francis, 1991. 52, 115, 116

[48] H. Fu, Y. Xu, S. Lin, D. W. K. Wong, and J. Liu. Deepvessel : Retinal Vessel Segmentation via Deep Learning and Conditional Random Field. In International Conference on MICCAI, pages 132–139. Springer, 2016. 66

[49] K.-S. Fu and J. K. Mui. A survey on image segmentation. Pattern Recognit., 13(1) :3–16, 1981. 19

[50] K. Fukushima and S. Miyake. Neocognitron : A self-organizing neural network model for a mechanism of visual pattern recognition. In Competition and cooperation in neural nets, pages 267–285. Springer, 1982. 25

[51] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. with Appl., 2(1) :17 – 40, 1976. 63

[52] P.-A. Ganaye, M. Sdika, and H. Benoit-Cattin. Towards integrating spatial localization in convolutional neural networks for brain image segmentation. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pages 621–625, 2018. 33, 67

[53] P.-A. Ganaye, M. Sdika, B. Triggs, and H. Benoit-Cattin. Removing Segmentation Inconsistencies with Semi-Supervised Non-adjacency Constraint. Med. Image Anal., 58 :101551, 2019. 33, 66

[54] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. Appl. Soft Comput., 70 :41–65, 2018. 23

[55] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 27, 41, 95, 139, 147

[56] R. Glowinski and P. Le Tallec. Numerical solution of problems in incompressible finite elasticity by augmented Lagrangian methods. I. Two-dimensional and axisymmetric problems. SIAM Journal on Applied Mathematics, 42(2) :400–429, 1982. 109, 128, 131, 133, 134

[57] M. Han, Y. Ma, J.and Li, M. Li, Y. Song, and Q. Li. Segmentation of organs at risk in CT volumes of head, thorax, abdomen, and pelvis. In SPIE Medical Imaging 2015 : Image Processing, volume 9413, page 94133J, 2015. 32

[58] X. Han, C. Xu, and J. L. Prince. A topology preserving level set method for geometric deformable models. IEEE Trans. Pattern Anal. Mach. Intell., 25(6) :755–768, 2003. 104

[59] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr : Transformers for 3D medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 574–584, 2022. 155

[60] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers : Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pages 1026–1034, 2015. 27

[61] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 28, 29

[62] M. H. Hesamian, W. Jia, X. He, and P. Kennedy. Deep learning techniques for medical image segmentation : achievements and challenges. Journal of digital imaging, 32(4) :582–596, 2019. 35

[63] X. Hu, F. Li, D. Samaras, and C. Chen. Topology-preserving deep image segmentation. Advances in neural information processing systems, 32, 2019. 104, 105

[64] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1) :106, 1962. 25

[65] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. Advances in neural information processing systems, 28, 2015. 155

[66] F. Jia, J. Liu, and X.-C. Tai. A regularized convolutional neural network for semantic image segmentation. Anal. Appl., 19(01) :147–165, 2021. 68, 100

[67] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient Multi-scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. Med. Image Anal., 36 :61–78, 2017. 33, 66

[68] M. Kass, A. P. Witkin, and D. Terzopoulos. Snakes : Active Contour Models. J. Comput. Vis., 1(4) :321–331, 1988. 20, 21, 33

[69] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed. Constrained-CNN Losses for Weakly Supervised Segmentation. Med. Image Anal., 54 :88–99, 2019. 33, 66

[70] B. Kim and J. C. Ye. Mumford–Shah Loss Functional for Image Segmentation with Deep Learning. IEEE Trans. Image Process., 29 :1856–1866, 2020. 33, 66, 71

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012. 28

[72] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. 33

[73] Z. Lambert, C. Le Guyader, and C. Petitjean. Analysis of the weighted Van der Waals-Cahn-Hilliard model for image segmentation. In 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–6, 2020. 15, 169

[74] Z. Lambert, C. Le Guyader, and C. Petitjean. A Geometrically-Constrained Deep Network for CT Image Segmentation. In IEEE International Symposium on Biomedical Imaging (ISBI), 2021. 15

[75] Z. Lambert, C. Le Guyader, and C. Petitjean. Enforcing Geometrical Priors in Deep Networks for Semantic Segmentation Applied to Radiotherapy Planning. J. Math. Imaging Vision, pages 1–24, 2022. 15

[76] Z. Lambert, C. Petitjean, B. Dubray, and S. Ruan. SegTHOR : Segmentation of Thoracic Organs at Risk in CT images. In 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–6, 2020. 15, 19, 94, 96

[77] C. Le Guyader and L. Vese. A combined segmentation and registration framework with a nonlinear elasticity smoother. Comput. Vis. Image Underst., 115(12) :1689–1709, 2011. 106

[78] C. Le Guyader and L. A. Vese. Self-Repelling Snakes for Topology-Preserving Segmentation Models. IEEE Trans. Image Process., 17(5) :767–779, 2008. 32

[79] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proc. IEEE, 86(11) :2278–2324, 1998. 25, 26, 27, 28

[80] M. C. H. Lee, K. Petersen, N. Pawlowski, B. Glocker, and M. Schaap. TETRIS : Template transformer networks for image segmentation with shape priors. IEEE Trans. Med. Imaging, 38(11) :2596–2606, 2019. 155

[81] B. Li, W. J. Niessen, S. Klein, M. Groot, M. A. Ikram, M. W. Vernooij, and E. E. Bron. A hybrid deep learning framework for integrated segmentation and registration : evaluation on longitudinal white matter tract changes. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pages 645–653. Springer, 2019. 106

[82] H. Li, D. Hu, H. Liu, J. Wang, and I. Oguz. Cats : Complementary CNN and Transformer Encoders for Segmentation. In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pages 1–5. IEEE, 2022. 155

[83] X. Li, H. Chen, X.Qi, Q. Dou, C.-W. Fu, and P.-. Heng. H-denseunet : Hybrid densely connected unet for liver and liver tumor segmentation from CT volumes. CoRR, abs/1709.07330, 2017. 29

[84] M. Lin, Q. Chen, and S. Yan. Network in network. arXiv preprint arXiv :1312.4400, 2013. 26

[85] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017. 30

[86] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. Med. Image Anal., 42 :60–88, 2017. 23, 25

[87] J. Liu, X. Wang, and X.-C. Tai. Deep Convolutional Neural Networks with Spatial Regularization, Volume and Star-shape Priori for Image Segmentation. CoRR, abs/2002.03989, 2020. 34, 67, 68, 98

[88] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015. 28, 29

[89] S. Luo, X.-C. Tai, L. Huo, Y. Wang, and R. Glowinski. Convex shape prior for multi-object segmentation using a single level set function. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 613–621, 2019. 154

[90] R. Malladi, R. Kimmel, D. Adalsteinsson, G. Sapiro, V. Caselles, and J. A. Sethian. A geometric approach to segmentation and analysis of 3D medical images. In Proceedings of the workshop on mathematical methods in biomedical image analysis, pages 244–252. IEEE, 1996. 21

[91] M Marcus and V. J. Mizel. Transformations by functions in Sobolev spaces and lower semicontinuity for parametric variational problems. Bull. Am. Math. Soc., 79(4) :790–795, 1973. 113

[92] T. McInerney and D. Terzopoulos. Deformable models in medical image analysis : a survey. Med. Image Anal., 1(2) :91–108, 1996. 21

[93] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net : Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In 2016 fourth international conference on 3D vision (3DV), pages 565–571. IEEE, 2016. 29, 30, 41, 94

[94] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning : A survey. IEEE Trans. Pattern Anal. Mach. Intell., 2021. 23

[95] L. Modica. The gradient theory of phase transitions and the minimal interface criterion. Arch. Ration. Mech. An., 98(2) :123–142, 1987. 169

[96] L. Moisan. How to discretize the Total Variation of an image ? PAMM, 7(1) :1041907–1041908, 2007. 72

[97] J.-J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. Comptes rendus hebdomadaires des séances de l'Académie des sciences, 255 :2897–2899, 1962. 58

[98] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. Comm. Pure Appl. Math., 42(5) :577–685, 1989. 22, 71

[99] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In Icml, 2010. 27

[100] P. V. Negrón Marrero. A numerical method for detecting singular minimizers of multidimensional problems in nonlinear elasticity. Numer. Math., 58 :135–144, 1990. 117

[101] B. E. Nelms, W. A. Tomé, G. Robinson, and J. Wheeler. Variations in the contouring of organs at risk : test case from a patient with oropharyngeal cancer. Int. J. Radiat. Oncol. Biol. Phys., 82(1) :368–378, 2012. 44

[102] S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-Paredes, A. Karthikesalingam, C. Chu, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. CoRR, abs/1809.04430, 2018. 29

[103] R. Nock and F. Nielsen. Statistical region merging. IEEE Trans. Pattern Anal. Mach. Intell., 26(11) :1452–1458, 2004. 19

[104] M. S. Nosrati and G. Hamarneh. Incorporating Prior Knowledge in Medical Image Segmentation : a Survey. CoRR, abs/1607.01092, 2016. 32

[105] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed : Algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys., 79(1) :12–49, 1988. 21

[106] N. Otsu. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern., 9(1) :62–66, 1979. 19

[107] J. Peng, H. Kervadec, J. Dolz, I. Ben Ayed, M. Pedersoli, and C. Desrosiers. Discretely-constrained deep network for weakly supervised segmentation. Neural Netw., 130 :297 – 308, 2020. 33, 63, 64, 66

[108] O. Petit, N. Thome, A. Charnoz, A. Hostettler, and L. Soler. Handling missing annotations for semantic segmentation with deep convnets. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pages 20–28. Springer, 2018. 45

[109] D. L. Pham, C. Xu, and J. L. Prince. Current methods in medical image segmentation. Annu. Rev. Biomed. Eng., 2(1) :315–337, 2000. 11, 19

[110] R. Prevost. Méthodes variationnelles pour la segmentation d'images à partir de modèles : applications en imagerie médicale. Theses, Université Paris Dauphine - Paris IX, October 2013. 20

[111] M. I. Razzak, S. Naz, and A. Zaib. Deep learning for medical image processing : Overview, challenges and the future. Classification in BioApps, pages 323–350, 2018. 35

[112] O. Ronneberger, P. Fischer, and T. Brox. U-net : Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015. 28, 29

[113] F. Rosenblatt. The perceptron : a probabilistic model for information storage and organization in the brain. Psychological review, 65(6) :386, 1958. 24

[114] H. R. Roth, H. Oda, Y. Hayashi, M. Oda, N. Shimizu, M. Fujiwara, K. Misawa, and K. Mori. Hierarchical 3D fully convolutional networks for multi-organ segmentation. arXiv preprint arXiv :1704.06382, 2017. 29

[115] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. Phys. D, 60(1-4) :259–268, 1992. 69

[116] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. Nature, 323(6088) :533–536, 1986. 25

[117] C. Rupprecht, E. Huaroc, M. Baust, and N. Navab. Deep active contours. CoRR, abs/1607.05074, 2016. 32, 66

[118] S. S. M. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In International workshop on machine learning in medical imaging, pages 379–387. Springer, 2017. 30

[119] E. Schreibmann, D. M. Marcus, and T. Fox. Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. J. Appl. Clin. Med. Phys., 15(4) :22–38, 2014. 32

[120] F. Ségonne. Active contours under topology control—genus preserving level sets. Int. J. Comput. Vision, 79(2) :107–117, 2008. 32, 104

[121] F. Ségonne, J. Pacheco, and B. Fischl. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. IEEE Trans. Med. Imaging, 26(4) :518–529, 2007. 104

[122] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. W. Pluim, U. Bauer, and B. H. Menze. clDice-a novel topology-preserving loss function for tubular structure segmentation. pages 16560–16569, 2021. 105

[123] C. Y. Siu, H. L. Chan, and L. M. Lui. Image Segmentation with Partial Convexity Shape Prior Using Discrete Conformality Structures. SIAM J. Imaging Sci., 13(4) :2105–2139, 2020. 32

[124] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration : A survey. IEEE Trans. Med. Imaging, 32(7) :1153–1190, 2013. 106

[125] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15 :1929–1958, 2014. 39

[126] M. Storath and A. Weinmann. Fast partitioning of vector-valued images. SIAM J. Imag. Sci., 7(3) :1826–1852, 2014. 22

[127] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. CoRR, abs/1707.03237, 2017. 30, 41, 124

[128] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. Chiang, Z. Wu, and X. Ding. Embracing imperfect datasets : A review of deep learning solutions for medical image segmentation. Med. Image Anal., 63 :101693, 2020. 45

[129] G. Taylor, R. Burmeister, Z. Xu, B. Singh, A. Patel, and T. Goldstein. Training Neural Networks Without Gradients : A Scalable ADMM Approach. In M. F. Balcan and K. Q. Weinberger, editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 2722–2731. PMLR, 2016. 34, 63

[130] R. Trullo. Deep learning based approaches for the segmentation of Organs at Risk in Thoracic Computed Tomography Scans. PhD thesis, University of Normandy, 2018. 37

[131] R. Trullo, C. Petitjean, B. Dubray, and S. Ruan. Multi-Organ Segmentation using Distance-Aware Adversarial Networks. J. Med. Imaging, 6(1) :014001, 2019. 33, 67

[132] A. Tsai, A. Yezzi, and A. S. Willsky. Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification. IEEE Trans. Image Process., 10(8) :1169–1186, 2001. 23

[133] N. Tustison, Br. Avants, M. Siqueira, and J. Gee. Topological well-composedness and glamorous glue : A digital gluing algorithm for topologically constrained front propagation. IEEE Transactions on Image Processing, 20 :1756–61, 11 2010. 104

[134] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017. 155

[135] B. Vemuri, J. Ye, Y. Chen, and C. Leonard. Image registration via level-set motion : applications to atlas-based segmentation. Med. Image Anal., 7(1) :1–20, 2003. 106

[136] L. A. Vese and T. F. Chan. A multiphase level set framework for image segmentation using the Mumford and Shah model. Int. J. Comput. Vision, 50(3) :271–293, 2002. 22

[137] L. A. Vese and C. Le Guyader. Variational methods in image processing. CRC Press Boca Raton, FL, 2016. 21

[138] S. K. Vinod, M. Min, M. G. Jameson, and L. C. Holloway. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J. Med. Imaging Radiat. Oncol., 60(3) :393–406, 2016. 44

[139] K. Vinogradova, A. Dibrov, and G. Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 13943–13944, 2020. 148

[140] B. Wirth. On the Gamma-limit of joint image segmentation and registration functionals based on phase fields. Interfaces Free Bound., 18(4) :441–477, 2016. 112, 115

[141] D. J. Withey and Z. J. Koles. Medical image segmentation : Methods and software. In 2007 Joint Meeting of the 6th International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging, pages 140–143. IEEE, 2007. 19, 31

[142] M. K. Wyburd, N. K. Dinsdale, A. IL. Namburete, and M. Jenkinson. TEDS-Net : Enforcing diffeomorphisms in spatial transformers to guarantee topology preservation in segmentations. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 250–260. Springer, 2021. 154

[143] Z. Xu and M. Niethammer. DeepAtlas : Joint Semi-Supervised Learning of Image Registration and Segmentation. CoRR, abs/1904.08465, 2019. 106

[144] A. Yezzi, S. Kichenassamy, A. Kumar, P. Olver, and A. Tannenbaum. A geometric snake model for segmentation of medical imagery. IEEE Trans. Med. Imaging, 16(2) :199–209, 1997. 21

[145] A. Yezzi, L. Zollei, and T. Kapur. A variational framework for joint segmentation and registration. In Mathematical Methods in Biomedical Image Analysis, pages 44–51. IEEE-MMBIA, 2001. 106

[146] R. A. Yotter, R. Dahnke, P. M. Thompson, and C. Gaser. Topological correction of brain surface meshes using spherical harmonics. Hum. Brain Mapp., 32(7) :1109–1124, 2011. 104

[147] D. Zhang and L. M. Lui. Topology-preserving 3D image segmentation based on hyperelastic regularization. J. Sci. Comput., 87(3) :1–33, 2021. 106

[148] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional Random Fields as Recurrent Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 1529–1537, 2015. 66

[149] H. Zhu, F. Meng, J. Cai, and S. Lu. Beyond pixels : A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. J. Visual Commun. Image Represent., 34 :12–27, 2016. 19

Annexe

Dans cette annexe, nous joignons un travail portant sur l'analyse du modèle pondéré de Van der Waals-Cahn-Hilliard pour la segmentation d'images qui a fait l'objet d'une publication ([73]) et dont le résumé en français suit. Ce travail préliminaire a servi de point de départ à notre réflexion quant au fait d'imposer l'alignement des contours entre prédiction et vérité terrain, et d'exploiter de façon sous-jacente les propriétés de la variation totale pondérée. Cela, en sus de l'appariement des intensités lumineuses des pixels dans l'entraînement des réseaux profonds pour la segmentation d'images. Même si l'approche revêtait un intérêt théorique et numérique, les premières simulations et comparaisons nous ont conduits à privilégier les modèles développés dans le Chapitre 3.

Dans l'article fondateur *The Gradient Theory of Phase Transitions and the Minimal Interface Criterion, 1987*, L. Modica ([95]) prouve certaines conjectures liées à la théorie des transitions de phase de Van der Waals-Cahn-Hilliard.
Cette théorie vise à surmonter le problème d'absence d'unicité du problème initialement considéré en imposant à l'interface de satisfaire un critère de surface minimale. Dans cette perspective, une famille de fonctionnelles paramétrées par $\varepsilon > 0$ et incluant une dépendance au gradient de l'inconnue est introduite, modélisant cette contrainte de surface minimale, et le comportement asymptotique lorsque $\varepsilon \to 0^+$ des solutions u_ε du problème de minimisation relatif est analysé à travers un résultat de Γ-convergence. Motivée par ce travail, cette contribution aborde la question de l'extension de ce résultat au cas pondéré. Il est démontré que ce nouveau modèle hérite des bonnes propriétés du modèle original non pondéré avec en particulier, la validité du résultat de Γ-convergence, utile pour la minimisation du périmètre pondéré, et pertinent pour la segmentation d'images.

Analysis of the weighted Van der Waals-Cahn-Hilliard model for image segmentation

Zoé Lambert
Normandie Univ
INSA Rouen, LMI
76000 Rouen, France
zoe.lambert@insa-rouen.fr

Carole Le Guyader
Normandie Univ
INSA Rouen, LMI
76000 Rouen, France
carole.le-guyader@insa-rouen.fr

Caroline Petitjean
Université de Rouen Normandie
LITIS
76801 Saint-Etienne-du-Rouvray Cedex, France
caroline.petitjean@univ-rouen.fr

***Abstract*—In the seminal paper** *The Gradient Theory of Phase Transitions and the Minimal Interface Criterion, 1987*, **L. Modica ([13]) proves some conjectures related to the Van der Waals-Cahn-Hilliard theory of phase transitions.**
This theory intends to overcome the issue of lack of uniqueness of the solution of the initially considered minimization problem —which aims to minimize the total energy of a fluid confined to a bounded container $\Omega \subset \mathbb{R}^n$ and with Gibbs free energy per unit volume, a prescribed function W of the density distribution u —by enforcing that the interface has minimal surface. In that purpose, a family of functionals parameterized by $\varepsilon > 0$ and including a dependency on the density gradient modelling this interfacial energy is introduced, and the asymptotic behavior as $\varepsilon \to 0^+$ of the solutions u_ε of the related minimization problem is analyzed through a Γ-convergence result. Motivated by this work, this paper addresses the question of extending this result to the weighted case. It is shown that this new model inherits the fine properties of the original unweighted one with in particular, the validity of the Γ-convergence result, useful for the minimization of weighted perimeter, relevant for image segmentation.

***Index Terms*—phase transition, weighted BV, Γ-convergence, classification**

I. Motivations

The Van der Waals-Cahn-Hilliard theory of phase transitions originates from the search of the stable configurations of a fluid made up of two unstable components confined to a bounded container $\Omega \subset \mathbb{R}^n$, and of the characterization of the interface between the two phases while the system reaches equilibrium (see [1, Chapter 5, Section 5.2], [13] and [19]). The problem amounts to minimizing the fluid total energy $E(u) = \int_\Omega W(u(x))\,dx$ over $u \in L^1(\Omega)$, W being a given function of the density distribution u denoting the fluid Gibbs free energy per unit volume, subject to the prescribed total mass constraint $\int_\Omega u(x)\,dx = m$. The function W is assumed to be $\mathcal{C}^2$, nonnegative, with exactly two zeros α and β such that $0 < \alpha < \beta$, and to satisfy $W'(\alpha) = W'(\beta) = 0$, and $W''(\alpha) > 0$, $W''(\beta) > 0$. Also, still following Aubert and Kornprobst ([1, page 225]), we suppose that there exist positive constants c_1 and c_2, l and an integer $p \geq 2$ such that $c_1 |u|^p \leq W(u) \leq c_2 |u|^p$ for $|u| \geq l$. This latter requirement ensures the well-posedness (in terms of functional spaces) of the minimization problem introduced later on. Provided $\alpha\,\mathrm{meas}(\Omega) < m < \beta\,\mathrm{meas}(\Omega)$ —'meas' standing for the Lebesgue measure—, the minimizers of this problem are exactly the set of L^1 functions taking only the values α and β and complying with the integral constraint. Equivalently, this minimization problem yields to partitions of Ω into two measurable sets A and B fulfilling $\alpha\,\mathrm{meas}(A) + \beta\,\mathrm{meas}(B) = m$. With no restriction on the shape of the interface between the sets $\{u = \alpha\}$ and $\{u = \beta\}$, the problem lacks uniqueness —from a physical viewpoint, the model fails to enhance small effects —, making it impossible to recover the physically reasonable criterion according to which the interface has minimal area. To overcome this issue and so to restore the small effects neglected by this primal model, the Van der Waals-Cahn-Hilliard theory introduces a family of functionals $(E_\varepsilon)_\epsilon$ parameterized by $\varepsilon > 0$ —the interfacial penalization is thus expressed through the dependency of the density gradient as will be seen just after—and focuses on the related perturbed problem:

$$\inf_{\substack{u \in H^1(\Omega) \\ \int_\Omega u\,dx = m}} \int_\Omega W(u) + \varepsilon^2 |\nabla u|^2\,dx. \qquad (P_\varepsilon)$$

Denoting by u_ε a minimizer of (P_ε) —existence is ensured by the direct method of the calculus of variations, nevertheless uniqueness still lacks in general —, a result relating u_{ε_j} and $u_0 = \lim_{j \to +\infty} u_{\varepsilon_j}$ for any L^1-convergent subsequence of $(u_\varepsilon)_\varepsilon$ is then established and stated as :

Theorem 1: Taken from [19, Theorem 1]

Suppose $u_{\varepsilon_j} \longrightarrow u_0$ in $L^1(\Omega)$ for some sequence of numbers $\varepsilon_j \longrightarrow 0$, where u_{ε_j} is a solution of (P_{ε_j}). Then u_0 is a solution of (P_0):

$$\inf_{\substack{u \in BV(\Omega) \\ W(u(x))=0 \text{ a.e.}, \int_\Omega u\,dx = m}} \mathrm{Per}_\Omega\{u = \alpha\}. \qquad (P_0)$$

The possible limit points u_0 are thus those taking only the values α or β, satisfying the integral constraint, while minimizing the perimeter of the interface between $\{x \in \Omega \,|\, u(x) = \alpha\}$ and $\{x \in \Omega \,|\, u(x) = \beta\}$, making then the interface not too

This project was co-financed by the European Union with the European Regional Development Fund (ERDF, 18P03390/18E01750/18P02733) and by the Haute-Normandie Regional Council via the M2SINUM project.

978-1-7281-8750-1/20/$31.00 ©2020 IEEE

irregular. As pointed out by Sternberg [19, Page 10], the proof relies on the derivation of an asymptotic expansion for the energy of (P_ε) and on the identification of the first non-trivial term which is here of order $\mathcal{O}(\varepsilon)$. Anticipating this order, a rescaling is often made, yielding the following family of functionals $(F_\varepsilon)_\varepsilon : L^1(\Omega) \to \mathbb{R}$ given by:

$$F_\varepsilon(u) = \begin{cases} \frac{1}{\varepsilon}\int_\Omega W(u(x))\,dx + \varepsilon\int_\Omega |\nabla u|^2\,dx, & \substack{u\in H^1(\Omega)\\ \int_\Omega u\,dx = m} \\ +\infty \text{ otherwise} \end{cases},$$

while $F_0 : L^1(\Omega) \to \mathbb{R}$ is defined by:

$$F_0 = \begin{cases} K\ \mathrm{Per}_\Omega\left\{u=\alpha\right\}, & \substack{u\in BV(\Omega),\ \int_\Omega u\,dx = m\\ W(u(x))=0 \text{ a.e.}} \\ +\infty \text{ otherwise} \end{cases},$$

with $K = 2\left\{\inf_{\substack{\gamma,\ \text{piecewise } c^1\\ \gamma(-1)=\alpha,\ \gamma(1)=\beta}} \int_{-1}^{1} \sqrt{W(\gamma(s))}\,|\gamma'(s)|\,ds\right\}$, while Per_Ω stands for the classical perimeter. The penalties of $+\infty$ in the previous definitions enable one to define F_ε and F_0 on $L^1(\Omega)$, functional space whose topology exhibits desirable compactness properties with respect to H^1 and BV as well as being larger than those latter ones.

Now, a straightforward connection/transposition of this approach to the binary image segmentation setting can be drawn (see again [1, Chapter 5, Section 5.2] or [17], [18]). Assuming, for the sake of simplicity, that the image to be recovered $u : \Omega \to \mathbb{R}$ from the observed data u_0 comprises two regions $R_1 = \{x \in \Omega \,|\, u(x) = \alpha\}$ and $R_2 = \{x \in \Omega \,|\, u(x) = \beta\}$, a suitable minimization criterion could be $\frac{1}{\varepsilon}\int_\Omega W(u(x))\,dx + \varepsilon\int_\Omega |\nabla u|^2\,dx$, combined with a classical fidelity term such as $\|u - u_0\|^2_{L^2(\Omega)}$. The potential W prescribes the number of classes (—two in the considered case —) and contributes to push the values of the recovered image towards the labels of the induced classes. With parameter ε decreasing, the diffusion process linked to the component $\varepsilon\int_\Omega |\nabla u|^2\,dx$ weakens, whilst sharpening the segmentation process. An interesting fact is that this approach reconciles the intrinsic discrete nature of segmentation —which consists in assigning a label to each image pixel —with the continuous dimension of variational methods. To put it in another way, the labels that are discrete in essence are viewed as the limit points of continuous variables. Motivated by these observations and the straightforward connections that can be made with image segmentation, we propose extending these theoretical results to the weighted case, while disregarding the total mass constraint —note however that the result would still hold with this additional integral constraint —. In line with the unweighted formulation, we suggest considering the family of functionals $(\tilde{F}_\varepsilon)_\varepsilon$ defined by:

$$\tilde{F}_\varepsilon(u) = \begin{cases} \frac{1}{\varepsilon}\int_\Omega w\ W(u(x))\,dx + \varepsilon\int_\Omega w\,|\nabla u|^2\,dx, & u\in H^1(\Omega), \\ +\infty \text{ otherwise,} \end{cases}$$

while $\tilde{F}_0 : L^1(\Omega) \to \mathbb{R}$ is defined by:

$$\tilde{F}_0 = \begin{cases} K\,\mathrm{Per}_{\Omega,w}\left\{u=\alpha\right\}, & u(x)\in\{\alpha,\beta\} \text{ a.e.} \\ +\infty & \text{otherwise} \end{cases},$$

with w a weight (whose smoothness will be made clearer afterwards) and $\mathrm{Per}_{\Omega,w}$ denoting the w-weighted perimeter defined later on, and investigate its asymptotic behaviour. Note nevertheless that as w is assumed to satisfy $0 < c \le w \le 1$ (see below for accurate justifications), it is equivalent to working with $H^1(\Omega) = W^{1,2}(\Omega)$ classical Sobolev space, or $H^1_w(\Omega) = W^{1,2}(\Omega, w)$ defined by:

$$W^{1,2}(\Omega,w) = \left\{ f \in L^2(\Omega,w) \,\middle|\, \begin{array}{c} \nabla f \text{ exists weakly} \\ |\nabla f| \in L^2(\Omega,w) \end{array} \right\},$$

where

$$L^2(\Omega,w) = \left\{ f : \Omega \to \mathbb{R} \text{ measurable } \middle|\ w|f|^2 \in L^1(\Omega) \right\}.$$

We would like to point out that prior related works such as [4], [9] or [11] (—even if this latter one is more theory-oriented —) consider weighted functional spaces in the context of segmentation. More precisely, in [4], Bresson *et al.* present a global minimization framework —unifying the geodesic active contours [6] and the piecewise-constant Mumford-Shah model ([14]) with the Rudin-Osher-Fatemi model ([16]) for image denoising) —designed to overcome the limitation of local minima and thus able to deal with global minimum. In that purpose, they let $g : \mathbb{R}^+ \to \mathbb{R}^+$ be an edge detector function satisfying $g(0) = 1$, g strictly decreasing and $\lim_{r\to+\infty} g(r) = 0$, and use the generalization of the notion of function of bounded variation to the setting of BV-spaces with a weight function, yielding as new regularizer, $TV_g(u)$, —u being the image to be recovered and segmented —, the weighted total variation (see Section II for the theoretical details). An intuitive interpretation of this weighted total variation is available in the case where the weight g is assumed sufficiently smooth: if $E \subset \mathbb{R}^2$ denotes a Lebesgue measurable set with regular boundary of class $\mathcal{C}^2$, then $TV_g(\chi_E) = \int_{\Omega\cap\partial E} g\,d\mathcal{H}^1$ ($\mathcal{H}^1$ denoting the 1-dimensional Hausdorff measure), which can be interpreted as a new definition of the curve length with a metric that depends on the image content. Thus including the weighted total variation in a segmentation model not only forces the recovered image u to be piecewise constant, but also enforces edge alignment.

Inspired by prior works by Dávila [7] and Ponce [15], the work [9] theoretically analyzes the nonlocal counterpart of this weighted semi-norm, quantity involving a sequence of integral operators. The work [11] is dedicated to the analysis of the Ambrosio-Tortorelli approximation scheme with weighted underlying metric, the starting point of the study being the bilevel learning scheme introduced in [8] to find the best tuning parameter in the Mumford-Shah model. Although related to our proposed work in terms of technical mathematical tools, [11] focuses on a different problem from ours. The underlying practical application we have in mind would fall within the domain of constrained deep networks for supervised/semi-supervised segmentation : in the training process, the weight could be computed from the ground truth and included in the pipeline in order to enforce edge alignment between the predicted segmentation and this ground truth.

Before turning to the statement of the main result (Section III), we introduce some preliminary mathematical tools (Section II) (—complemented by new theoretical results —), and particularly, the definition of the weighted BV space.

II. Mathematical background

Let Ω be a connected bounded open subset of $\mathbb{R}^2$ of class $\mathcal{C}^1$. The generalization of the notion of function of bounded variation to the setting of weighted BV-spaces associated with a Muckenhoupt's weight is tackled in [2]. We follow Baldi's notations to define the weighted BV-space related to weight w.

For a general weight w, some hypotheses are required. More precisely, Ω_0 being a neighborhood of $\bar{\Omega}$, the positive weight $w \in L^1_{loc}(\Omega_0)$ is assumed to belong to the global Muckenhoupt's $A_1 = A_1(\Omega)$ class of weight functions, *i.e.*, w satisfies the condition $(\mathcal{C}) : C\, w(x) \geq \frac{1}{|B(x,r)|}\int_{B(x,r)} w(y)\, dy$ a.e. in any ball $B(x,r) \subset \Omega_0$. Now, denoting by A_1^* the class of weights $w \in A_1$, w lower semicontinuous (lsc) and that satisfy condition $(\mathcal{C})$ pointwise, the definition of the weighted BV-space related to weight w is given by:

Definition 1 ([2, Definition 2]): Let w be a weight function in the class A_1^*. We denote by $BV(\Omega, w)$ the set of functions $u \in L^1(\Omega, w)$ (set of functions that are integrable with respect to the measure $w(x)\, dx$) such that:

$$\sup\left\{\int_\Omega u\, \mathrm{div}(\varphi)\, dx : |\varphi| \leq w \text{ everywhere, } \varphi \in \mathrm{Lip}_0(\Omega, \mathbb{R}^2)\right\} < \infty, \quad (1)$$

with $\mathrm{Lip}_0(\Omega, \mathbb{R}^2)$ the space of Lipschitz continuous functions with compact support. We denote by $\mathrm{TV}_w(u)$ the quantity (1). From now on, for theoretical purposes, we assume that $\exists c > 0$ such that $0 < c \leq w \leq 1$ —the latter inequality is consistent with the case where w is an edge-detector function —and that w is Lipschitz continuous. In particular, it suffices to take $C = \frac{1}{c}$ to satisfy $(\mathcal{C})$ pointwise. In the following remark, we point out a first observation relevant for the sake of consistency with the existing literature.

Remark 1: In [2], Baldi defines the BV-space taking as test functions elements of $\mathrm{Lip}_0(\Omega, \mathbb{R}^2)$ while classically in the literature, the test functions are chosen in $\mathcal{C}^1_c(\Omega, \mathbb{R}^2)$. In fact, these definitions coincide. This new result is justified hereafter.

Proof: Let $u \in L^1(\Omega, w)$ with the prescribed assumptions on w. We aim to prove that $A_u := \sup\left\{\int_\Omega u\, \mathrm{div}\, \varphi\, dx \,|\, |\varphi| \leq w \text{ everywhere, } \varphi \in \mathcal{C}^1_c(\Omega, \mathbb{R}^2)\right\}$ and $B_u := \sup\left\{\int_\Omega u\, \mathrm{div}\, \varphi\, dx \,|\, |\varphi| \leq w \text{ everywhere, } \varphi \in \mathrm{Lip}_0(\Omega, \mathbb{R}^2)\right\}$ coincide.

As $\mathcal{C}^1_c(\Omega, \mathbb{R}^2) \subset \mathrm{Lip}_0(\Omega, \mathbb{R}^2)$, one clearly has $A_u \leq B_u$. We now aim to prove the reverse inequality. First, we assume that $A_u < +\infty$, otherwise it is done. As $0 < c \leq w \leq 1$, with $\varphi \in \mathcal{C}^1_c(\Omega, \mathbb{R}^2)$ such that $|\varphi| \leq 1$, one has $\int_\Omega u\, \mathrm{div}\, \varphi\, dx = \frac{1}{c}\int_\Omega u\, \mathrm{div}(c\,\varphi)\, dx \leq \frac{1}{c} A_u$, meaning that $u \in BV(\Omega)$ in the classical sense (*i.e.*, with test functions in $\mathcal{C}^1_c(\Omega, \mathbb{R}^2)$).

Let $\{\varphi_k\} \in \mathrm{Lip}_0(\Omega, \mathbb{R}^2)$ with $\forall k \in \mathbb{N}$, $|\varphi_k| \leq w$ everywhere, be a maximizing sequence, *i.e.*, $\lim_{k\to+\infty} \int_\Omega u\, \mathrm{div}\, \varphi_k\, dx = B_u$. We know that for any $f \in \mathrm{Lip}_0(\Omega, \mathbb{R}^2)$, there exists a sequence $(f_j) \in \mathcal{C}^1_c(\Omega, \mathbb{R}^2)$ that uniformly converges to f in Ω. Then for each $k \in \mathbb{N}$, there exists a sequence $\{\varphi_{k,j}\}_{j\in\mathbb{N}} \in \mathcal{C}^1_c(\Omega, \mathbb{R}^2)$ such that $\forall j \in \mathbb{N}$, $|\varphi_{k,j}| \leq w$ everywhere (—this is ensured by the mollification process —) and such that $\{\varphi_{k,j}\}$ uniformly converges to φ_k in Ω when j tends to $+\infty$, that is,

$$\forall \epsilon > 0,\, \exists N_{\epsilon,k},\, \forall j \in \mathbb{N},\, \forall x \in \Omega,$$
$$\left(j \geq N_{\epsilon,k} \Longrightarrow |\varphi_{k,j}(x) - \varphi_k(x)| \leq \epsilon\right).$$

Let us take in particular $\epsilon = \frac{1}{k}$. Then

$$\exists N_k \in \mathbb{N},\, \forall j \in \mathbb{N},\, \forall x \in \Omega,$$
$$\left(j \geq N_k \Longrightarrow |\varphi_{k,j}(x) - \varphi_k(x)| \leq \frac{1}{k}\right).$$

From now on, one sets $j = N_k$ so that $\forall x \in \Omega$, $|\varphi_{k,N_k}(x) - \varphi_k(x)| \leq \frac{1}{k}$. Function u being an element of $BV(\Omega)$ in the classical sense, $\int_\Omega u\, \mathrm{div}\, \varphi_k\, dx = -\int_\Omega \varphi_k \cdot dDu$, from density results and Lebesgue dominated convergence theorem, and similarly $\int_\Omega u\, \mathrm{div}\, \varphi_{k,N_k}\, dx = -\int_\Omega \varphi_{k,N_k} \cdot dDu$. Indeed, the classical technique of approximation by smooth functions called mollification (see [10, Chapter 4, Section 4.2]) states that with $u \in BV(\Omega)$ and $\Psi \in \mathrm{Lip}_0(\Omega, \mathbb{R})$, for $\epsilon > 0$ small enough, one has $\Psi \star \rho_\epsilon \in \mathcal{C}^\infty_c(\Omega, \mathbb{R})$ —ρ_ϵ denoting the standard mollifier —, and

$$\int_\Omega u\, \frac{\partial(\Psi * \rho_\epsilon)}{\partial x_i}\, dx = -\int_\Omega \Psi * \rho_\epsilon\, dD_i u,\ i = 1, 2.$$

As $\epsilon \to 0$, one has $\Psi * \rho_\epsilon \to \Psi$ uniformly and

$$\frac{\partial(\Psi * \rho_\epsilon)}{\partial x_i} = \frac{\partial \Psi}{\partial x_i} * \rho_\epsilon \longrightarrow \frac{\partial \Psi}{\partial x_i}$$

almost everywhere, so by the Lebesgue dominated convergence theorem, one gets :

$$\int_\Omega u\, \frac{\partial \Psi}{\partial x_i}\, dx = -\int_\Omega \Psi\, dD_i u,\ i = 1, 2.$$

We thus get:

$$\left|\int_\Omega u\, \mathrm{div}\, \varphi_k\, dx - \int_\Omega u\, \mathrm{div}\, \varphi_{k,N_k}\, dx\right|$$
$$\leq \|\varphi_k - \varphi_{k,N_k}\|_{L^\infty(\Omega)} \int_\Omega d\|Du\| \leq \frac{1}{k}\int_\Omega d\|Du\|.$$

We then derive the following inequality:

$$\left|\int_\Omega u\, \mathrm{div}\, \varphi_{k,N_k}\, dx - B_u\right| \leq \left|\int_\Omega u\, \mathrm{div}\, \varphi_{k,N_k}\, dx - \int_\Omega u\, \mathrm{div}\, \varphi_k\, dx\right|$$
$$+ \left|\int_\Omega u\, \mathrm{div}\, \varphi_k\, dx - B_u\right|,$$

yielding, thanks to the previous inequalities, $\lim_{k\to+\infty} \int_\Omega u\, \mathrm{div}\, \varphi_{k,N_k}\, dx = B_u$. But by definition, $\int_\Omega u\, \mathrm{div}\, \varphi_{k,N_k}\, dx \leq A_u$, so by passing to the limit when k tends to $+\infty$, it leads to $B_u \leq A_u$. ∎

Now it is ensured that both definitions —whether it be with test functions in $\mathrm{Lip}_0(\Omega, \mathbb{R}^2)$ or in $\mathcal{C}^1_c(\Omega, \mathbb{R}^2)$ —coincide, we prove some extensions of classical results in the weighted BV case, most of the time differently from the proposed approaches developed in [5] (w is assumed smoother here compared to

[5], yielding more straightforward proofs). Recall that if $u \in BV(\Omega)$, u has a distributional derivative Du which is a vector-valued Radon measure and the total variation of this measure is denoted by $\|Du\|$ ([10, Chapter 5, Section 5.1]).

Theorem 2: With the prescribed assumptions on w, one has $TV_w(u) < +\infty$ if and only if $u \in BV(\Omega)$. When this condition holds, $TV_w(u) = \int_\Omega w\, d\|Du\|$.

Proof:

- Let us assume first that $u \in BV(\Omega)$. $\forall \varphi \in \text{Lip}_0(\Omega, \mathbb{R}^2)$ with $|\varphi| \leq w$, $\int_\Omega u \,\text{div}\,\varphi\, dx = -\int_\Omega \varphi \cdot dDu \leq \int_\Omega w\, d\|Du\|$, resulting in $TV_w(u) \leq \int_\Omega w\, d\|Du\| < +\infty$ due to the boundedness of w and by passing to the supremum in the previous inequality.
- Let us assume now that $TV_w(u) < \infty$. As $0 < c \leq w$ everywhere, $\forall \varphi \in \text{Lip}_0(\Omega, \mathbb{R}^2)$ with $|\varphi| \leq 1$,

  $$\int_\Omega u \,\text{div}\,\varphi\, dx = \frac{1}{c}\int_\Omega u \,\text{div}(c\varphi)\, dx \leq \frac{1}{c} TV_w(u) < +\infty,$$

  yielding $TV(u) < +\infty$ and $u \in BV(\Omega)$ since $L^1(\Omega, w) \subset L^1(\Omega)$. Now, let $\Psi \in \text{Lip}_0(\Omega, \mathbb{R}^2)$ with $|\Psi| \leq 1$. Then

  $$\int_\Omega \Psi \cdot w\, dDu = -\int_\Omega \text{div}(w\Psi)\, u\, dx \leq TV_w(u).$$

  By passing to the supremum over Ψ, considering the measure $w dDu$ on Ω and computing its total variation defined by $\int_\Omega w d\|Du\| = \sup\left\{\int_\Omega \Psi \cdot w dDu \,|\, \Psi \in \text{Lip}_0(\Omega, \mathbb{R}^2),\, |\Psi| \leq 1\right\}$ (see [5, page 24] and [20]), it follows that:

  $$\int_\Omega w\, d\|Du\| \leq TV_w(u).$$

  Gathering the two previous inequalities yields the desired result.

■

The next new result gives some insight into the meaning of the weighted total variation.

Theorem 3: Let $E \subsetneq \Omega$ be a regular bounded open set in $\mathbb{R}^2$ with boundary of class $\mathcal{C}^2$. Then $TV_w(\chi_E) = \int_{\Omega\cap\partial E} w\, d\mathcal{H}^1$, χ_E being the characteristic function of E defined by $\chi_E(x) = \begin{cases} 1 & \text{if} \quad x \in E \\ 0 & \text{otherwise} \end{cases}$. The weighted total variation can thus be viewed as a new definition of the curve length with a metric that depends on the weight.

Proof: The set E being bounded, $\int_\Omega \chi_E\, dx = \text{meas}(E)$, Lebesgue measure of E and $\chi_E \in L^1(\Omega)$. Similarly, $\int_\Omega w\, \chi_E\, dx < +\infty$ due to the assumptions on w and $\chi_E \in L^1(\Omega, w)$. Suppose $g \in \mathcal{C}^1_c(\Omega, \mathbb{R}^2)$. Then by the Gauss-Green theorem,

$$\int_\Omega \chi_E \,\text{div}\, g\, dx = \int_E \text{div}\, g\, dx = \int_{\partial E} g \cdot \nu\, d\mathcal{H}^1,$$

where $\nu(x)$ is the unit outward normal to ∂E at x. Now $|\nu(x)| = 1$ so that if $|g(x)| \leq w(x)$, then

$$\int_{\partial E} g \cdot \nu\, d\mathcal{H}^1 \leq \int_{\partial E} w\, d\mathcal{H}^1$$

By passing to the supremum over g, it yields $TV_w(\chi_E) \leq \int_{\partial E} w\, d\mathcal{H}^1$.
It remains to prove that $TV_w(\chi_E) \geq \int_{\partial E} w\, d\mathcal{H}^1$.
Since E has a $\mathcal{C}^2$-boundary, $\nu(x)$ is a $\mathcal{C}^1$-vector-valued function of x with $|\nu(x)| = 1$ and it can thus be extended to a function N defined on the whole of $\mathbb{R}^2$ such that $N \in \mathcal{C}^2(\mathbb{R}^2, \mathbb{R}^2)$ and $|N(x)| \leq 1$ for all x (see [12, p. 5]). Let $\eta \in \mathcal{C}^1_c(\Omega)$ be such that $|\eta| \leq w$ everywhere. Then we have, setting $g = \eta\, N$,

$$\int_\Omega \chi_E \,\text{div}\, g\, dx = \int_{\partial E} \eta\, d\mathcal{H}^1.$$

By applying first the definition of the weighted total variation and by passing then to the supremum over η, leads to:

$$TV_w(\chi_E) \geq \sup\left\{\int_{\partial E} \eta\, d\mathcal{H}^1,\ \eta \in \mathcal{C}^1_c(\Omega),\, |\eta| \leq w(x)\right\}.$$

The weight w being Lipschitz, there exists a sequence $(\eta_k)_{k\in\mathbb{N}} \in \mathcal{C}^1_c(\Omega)$ that uniformly converges to w in Ω. Using density arguments similar to those used in Remark 1, one proves that $\sup\left\{\int_{\partial E} \eta\, d\mathcal{H}^1,\ \eta \in \mathcal{C}^1_c(\Omega),\, |\eta| \leq w(x)\right\} = \int_{\partial E} w\, d\mathcal{H}^1$, which concludes the proof. ■

We end up this section with an adaptation of the co-area formula relating the weighted total variation of u with the perimeters of its level sets, which constitutes again a new result.

Theorem 4: Co-area formula

Let $u \in BV(\Omega, w)$. Then $E_t = \{x \in \Omega \,|\, u(x) > t\}$ has finite perimeter for $\mathcal{L}^1$ a.e. $t \in \mathbb{R}$ and $TV_w(u) = \int_{-\infty}^{+\infty} \|D_w \chi_{E_t}\|\, dt$.

Proof: The proof is rather long and relies on the adaptation to the weighted case of the technical elements developed in [10, Chapter 5, Section 5.5, Theorem 1]. We only give the main steps. Let $\varphi \in \text{Lip}_0(\Omega, \mathbb{R}^2)$ with $|\varphi| \leq w$. We first prove that $\int_\Omega u \,\text{div}\,\varphi\, dx = \int_{-\infty}^{+\infty} \left(\int_{E_t} \text{div}\,\varphi\, dx\right) dt$. The case $u \geq 0$ is first investigated and using the fact that then $u(x) = \int_0^{+\infty} \chi_{E_t}(x)\, dt$ and Fubini-Lebesgue theorem, it follows that:

$$\int_\Omega u \,\text{div}\,\varphi\, dx = \int_0^{+\infty} \left(\int_{E_t} \text{div}\,\varphi\, dx\right) dt.$$

The case $u \leq 0$ is straightforwardly derived as well as the general case by writing $u = u^+ + (-u^-)$. By passing to the supremum over φ, it yields:

$$TV_w(u) \leq \int_{-\infty}^{+\infty} \|D_w \chi_{E_t}\|\, dt.$$

The second part of the proof is more involved and is based on the density result [2, Theorem 3.4]. ■

III. Main result

Provided with these elements, we are now ready to state the main result which is expressed as follows.

Theorem 5: Main result. With the prescribed conditions on W and with K defined above, $\tilde{F}_\varepsilon$ Γ-converges to $\tilde{F}_0$ for the (L^1-strong topology) with

$$\tilde{F}_0 = \begin{cases} K\,\mathrm{Per}_{\Omega,w}(E) & \text{if } u(x) \in \{\alpha,\beta\} \text{ a.e.} \\ +\infty & \text{otherwise} \end{cases},$$

with $E = \{u = \alpha\}$ and where $\mathrm{Per}_{\Omega,w}$, the weighted perimeter, is defined by :

$$\mathrm{Per}_{\Omega,w}(E) = TV_w(\chi_E) = \int_{\Omega\cap\partial E} w\, d\mathcal{H}^1,$$

(if E is assumed sufficiently smooth for the furthest right equality).
For the notion of Γ-convergence, we refer the reader to [3] or [1, Chapter 2, Section 2.1, Subsection 2.1.4].

Remark 2: The consequence of this result is that if v_ε is a sequence of minimizers of $\tilde{F}_\varepsilon$ such that v_ε strongly converges to $\bar{v}$ in $L^1(\Omega)$, then $\bar{v}$ is a solution of the problem

$$\inf_{u\in BV(\Omega)} K\,\mathrm{Per}_{\Omega,w}(\{u=\alpha\}),\ u(x)\in\{\alpha,\beta\}\ \text{ a.e..}$$

Before dealing with the proof, we need the following lemma taken from [1, Chapter 5, Section 5.2, Lemma 5.2.2] that we recall for the sake of completeness.

Lemma 1: Taken from [1, Chapter 5, Section 5.2, Lemma 5.2.2] Let us denote by g the auxiliary function defined by:

$$g(u) = \inf_{\substack{\gamma(-1)=\alpha \\ \gamma(1)=u}} \int_{-1}^{1} \sqrt{W(\gamma(s))}\,|\gamma'(s)|\,ds,$$

the infimum being taken over functions $\gamma(\cdot)$ that are Lipschitz continuous. Note that this quantity is related to the constant K introduced above.
For every $u \in \mathbb{R}$, there exists a function $\gamma_u : [-1,1] \to \mathbb{R}$ such that $\gamma_u(-1) = \alpha$, $\gamma_u(1) = u$, and

$$g(u) = \int_{-1}^{1} \sqrt{W(\gamma_u(s))}\,|\gamma_u'(s)|\,ds.$$

The function g is Lipschitz continuous and satisfies $|g'(u)| = \sqrt{W(u)}$ a.e.. There exists a smooth increasing function $\zeta :]-\infty,+\infty[\to]-1,1[$ such that the function $\xi(\tau) = \gamma_\beta(\zeta(\tau))$ satisfies :

$$2g(\beta) = \int_{-\infty}^{+\infty} \left[W(\xi(\tau)) + |\xi'(\tau)|^2\right]\,d\tau,$$
$$\lim_{\tau\to-\infty} \xi(\tau) = \alpha \quad \lim_{\tau\to+\infty} \xi(\tau) = \beta,$$

with these limits being attained at an exponential rate.
We now come back to the proof which appears to be an extension of the one of [1, Theorem 5.2.1] but involving nevertheless substantial changes. We follow the steps of Aubert and Kornprobst.

Proof:

- The first step of the proof consists in proving that whenever v_ε converges strongly in L^1 to v_0, one has

$$\tilde{F}_0(v_0) \le \liminf_{\varepsilon\to 0} \tilde{F}_\varepsilon(v_\varepsilon).$$

We can restrict ourselves to the case $v_0(x) = \begin{cases} \alpha & \text{if } x\in A \\ \beta & \text{if } x\in B \end{cases}$, with A and B being two disjoint sets satisfying $A\cup B = \Omega$. Indeed, all the other cases can be straightforwardly ruled out since if $W(v_0) \neq 0$ on a set of positive measure $\tilde{F}_0(v_0) = +\infty$ by definition of $\tilde{F}_0$. Additionally,

$$\liminf_{\varepsilon\to 0} \frac{1}{\varepsilon}\int_\Omega w\,W(v_\varepsilon)\,dx = +\infty \le \liminf_{\varepsilon\to 0} \tilde{F}_\varepsilon(v_\varepsilon)$$

resulting in $\liminf_{\varepsilon\to 0} \tilde{F}_\varepsilon(v_\varepsilon) \to +\infty$, which means that the inequality to be proved is obviously true.
Let us set $h_\varepsilon(x) = g(v_\varepsilon(x))$ with g defined in Lemma 1. Applying Cauchy's inequality to $\tilde{F}_\varepsilon(v_\varepsilon)$ yields:

$$\tilde{F}_\varepsilon(v_\varepsilon) \ge 2\int_\Omega w\,\sqrt{W(v_\varepsilon)}|\nabla v_\varepsilon|\,dx.$$

But from Lemma 1, due to the Lipschitz continuous feature of g and the definition of $|g'|$, $|\nabla h_\varepsilon(x)| = \sqrt{W(v_\varepsilon)}\,|\nabla v_\varepsilon|$. In addition,

$$\begin{aligned}\|h_\varepsilon - g(v_0)\|_{L^1(\Omega)} &= \|g(v_\varepsilon) - g(v_0)\|_{L^1(\Omega)}, \\ &\le \kappa_g\,\|v_\varepsilon - v_0\|_{L^1(\Omega)},\end{aligned}$$

κ_g denoting the Lipschitz constant of g, implying thus that h_ε strongly converges to $g(v_0)$ in $L^1(\Omega)$ and so in $L^1(\Omega,w)$. The lower semicontinuity of the weighted total variation for the $L^1(\Omega,w)$-strong convergence ([2, Theorem 3.2]) enables one to conclude that:

$$\begin{aligned}\liminf_{\varepsilon\to 0} \tilde{F}_\varepsilon(v_\varepsilon) &\ge \liminf_{\varepsilon\to 0}\ 2\int_\Omega w\,\sqrt{W(v_\varepsilon)}|\nabla v_\varepsilon|\,dx, \\ &= \liminf_{\varepsilon\to 0}\ 2\int_\Omega w\,|\nabla h_\varepsilon(x)|\,dx, \\ &\ge 2\,TV_w(g(v_0)).\end{aligned}$$

To conclude, $g(v_0)$ is such that $g(v_0) = \begin{cases} 0 & \text{if } x\in A \\ g(\beta) & \text{if } x\in B \end{cases}$, implying that $TV_w(g(v_0)) = g(\beta)\,\mathrm{Per}_{\Omega,w}\{v_0=\alpha\}$ and yielding the desired inequality (again, thanks to the definition of the auxiliary function g).

- It remains to prove that for each $v_0 \in L^1(\Omega)$, there exists a sequence ρ_ε such that ρ_ε strongly converges to v_0 in $L^1(\Omega)$ and

$$\tilde{F}_0(v_0) \ge \limsup_{\varepsilon\to 0} \tilde{F}_\varepsilon(\rho_\varepsilon).$$

Again, we may assume that $v_0 \in BV(\Omega)$ with $v_0(x) = \begin{cases} \alpha & \text{if } x\in A \\ \beta & \text{if } x\in B \end{cases}$, otherwise $\tilde{F}_0(v_0) = +\infty$ and $\rho_\varepsilon = v_0$ is suitable. Without loss of generality (see [19, page 222]), we may assume that $\Gamma = \partial A \cup \partial B \in \mathcal{C}^2$. We follow the arguments of [1, Chapter 5, Section 5.2] and [19] and define the signed distance function $D : \Omega \to \mathbb{R}$ by :

$$d(x) = \begin{cases} -\mathrm{dist}(x,\Gamma) & \text{if } x\in A \\ +\mathrm{dist}(x,\Gamma) & \text{if } x\in B \end{cases}$$

In a neighborhood of Γ, d is smooth (in fact, $\mathcal{C}^2$) and $|\nabla d| = 1$. We also define the sequence :

$$\rho_\varepsilon(x) = \begin{cases} \xi\left(\frac{-1}{\sqrt{\varepsilon}}\right) & \text{if} \quad d(x) < -\sqrt{\varepsilon} \\ \xi\left(\frac{d(x)}{\varepsilon}\right) & \text{if} \quad |d(x)| \leq \sqrt{\varepsilon} \\ \xi\left(\frac{1}{\sqrt{\varepsilon}}\right) & \text{if} \quad d(x) > \sqrt{\varepsilon} \end{cases}$$

involving the function ξ introduced in Lemma 1. It can be proved that ρ_ε converges strongly to v_0 in $L^1(\Omega)$ thanks to Lemma 1. Now, due to the properties $\lim_{\tau\to-\infty} \xi(\tau) = \alpha$ and $\lim_{\tau\to+\infty} \xi(\tau) = \beta$, one has :

$$\limsup_{\varepsilon\to 0} \tilde{F}_\varepsilon(v_\varepsilon) = \limsup_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_{\{|d(x)|\leq\sqrt{\varepsilon}\}} w \left[W\left(\xi(\frac{d(x)}{\varepsilon})\right) + |\xi'\left(\frac{d(x)}{\varepsilon}\right)|^2 \right] dx,$$

and the co-area formula enables one to conclude that :

$$\limsup_{\varepsilon\to 0} \tilde{F}_\varepsilon(v_\varepsilon) = \limsup_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_{-\sqrt{\varepsilon}}^{\sqrt{\varepsilon}} \left(\int_{\{d=s\}} w(r)[W(\xi(\frac{s}{\varepsilon})) + |\xi'(\frac{s}{\varepsilon})|^2] d\mathcal{H}^1(r)) \, ds.$$

Making the change of variable $\tau = \frac{s}{\varepsilon}$ and as $2g(\beta) = \int_{-\infty}^{+\infty} \left[W(\xi(\tau)) + |\xi'(\tau)|^2\right] d\tau$, one can bound above the previous quantity by :

$$\limsup_{\varepsilon\to 0} \tilde{F}_\varepsilon(v_\varepsilon) \leq 2\, g(\beta) \limsup_{\varepsilon\to 0} \max_{|t|\leq\sqrt{\varepsilon}} \left(\int_{\{d=t\}} w(r)\, d\mathcal{H}^1(r) \right).$$

Let us consider $t < 0$ and let us denote by $S_t = \{x \in A \,|\, d(x) = t\}$ and by $V_t = \{x \in A \,|\, t < d(x) < 0\}$. According to [13, Lemma 3.], one can show that $\chi_{A\setminus V_t}$ strongly converges to χ_A in $L^1(\Omega)$ and so in $L^1(\Omega, w)$ due to the assumptions on w. The lower semicontinuity of the weighted total variation implies that:

$$\int_{\{d=0\}} w(r)\, d\mathcal{H}^1(r) = \text{Per}_{\Omega,w}(A) = TV_w(\chi_A)$$
$$\leq \liminf_{t\to 0^-} TV_w(\chi_{A\setminus V_t}) = \liminf_{t\to 0} \int_{\{d=t\}} w(r)\, d\mathcal{H}^1(r).$$

In fact, one can prove that $\lim_{t\to 0^-} TV(\chi_{A\setminus V_t}) = TV(\chi_A)$ (see again [13, Lemma 3.]), which, combined with the fact that $\chi_{A\setminus V_t}$ strongly converges to χ_A in $L^1(\Omega)$, gives the strict convergence of $\chi_{A\setminus V_t}$ to χ_A in BV. This convergence implies weak-$*$ convergence, meaning that $\forall \Psi \in \mathcal{C}^0(\Omega, \mathbb{R}^2)$, $\lim_{t\to 0^-} \int_\Omega \Psi\, dD\chi_{A\setminus V_t} = \int_\Omega \Psi\, dD\chi_A$. We then prove that $\limsup_{t\to 0^-} TV_w(\chi_{A\setminus V_t}) \leq TV_w(\chi_A)$, adapting the first part of the proof of [13, Lemma 3.]. The same reasoning (with a slight adaptation) holds for $t > 0$. Consequently,

$$\limsup_{\varepsilon\to 0} \tilde{F}_\varepsilon(v_\varepsilon) \leq \tilde{F}_0(v_0),$$

which completes the proof. ■

IV. Conclusion

This short note exemplifies the strength of continuous variational methods to model problems involving discrete variables. In the context of image segmentation whose nature is intrinsically discrete, we have proposed an extension of the classical Van der Waals-Cahn-Hilliard phase transition model to the weighted case. We have shown that the theoretical results still hold, substantiating the fact that (for ε small enough) continuous variables are relevant to model discrete-set-valued-quantities. The underlying practical application we have in mind by introducing this weight is related to constrained deep networks for supervised/semi-supervised segmentation in order to enforce edge alignment in addition to intensity matching.

References

[1] G. Aubert and P. Kornprobst. Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations. Applied Mathematical Sciences. Springer-Verlag, 2001.

[2] A. Baldi. Weighted BV functions. Houston J. Math., 27(3):683–705, 2001.

[3] A. Braides. Approximation of free-discontinuity problems, volume 1694 of Lecture Notes in Mathematics. Springer-Verlag Berlin Heidelberg, 1998.

[4] X. Bresson, S. Esedoğlu, P. Vandergheynst, J.-P. Thiran, and S. Osher. Fast global minimization of the active contour/snake model. J. Math. Imaging Vis., 28(2):151–167, 2007.

[5] C. S. Camfield. Comparison of BV Norms in Weighted Euclidean Spaces and Metric Measure Spaces. PhD thesis, University of Cincinnati, U.S., 2008.

[6] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic Active Contours. Int. J. Comput. Vis., 22(1):61–87, 1993.

[7] J. Dávila. On an open question about functions of bounded variation. Calc. Var. Partial Dif., 15(4):519–527, 2002.

[8] J.C. De Los Reyes, C.-B. Schönlieb, and T. Valkonen. The structure of optimal parameters for image restoration problems. J. Math. Anal. Appl., 434(1):464 –500, 2016.

[9] N. Debroux and C. Le Guyader. A joint segmentation/registration model based on a nonlocal characterization of weighted total variation and nonlocal shape descriptors. SIAM J. on Imaging Sci., 11(2):957–990, 2018.

[10] L.C. Evans and R.F. Gariepy. Measure Theory and Fine Properties of Functions. CRC Press, 1992.

[11] I. Fonseca and P. Liu. The Weighted Ambrosio–Tortorelli Approximation Scheme. SIAM J. Math. Anal., 49(6):4491–4520, 2017.

[12] E. Giusti. Minimal Surfaces and Functions of Bounded Variation, volume 80 of Monographs in Mathematics. Birkhäuser Basel, 1984.

[13] L. Modica. The gradient theory of phase transitions and the minimal interface criterion. Arch. Ration. Mech. An., 98(2):123–142, 1987.

[14] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Anal., pages 577–685, 1989.

[15] A. C. Ponce. A new approach to Sobolev spaces and connections to Γ-convergence. Calc. Var. Partial Dif., 19(3):229–255, 2004.

[16] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. Phys. D, 60(1-4):259–268, 1992.

[17] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia. A variational model for image classification and restoration. IEEE T. Pattern Anal., 22(5):460–472, 2000.

[18] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia. Two Variational Models for Multispectral Image Classification. In M. Figueiredo, J. Zerubia, and A. K. Jain, editors, Energy Minimization Methods in Computer Vision and Pattern Recognition, pages 344–356. Springer Berlin Heidelberg, 2001.

[19] P. Sternberg. The effect of a singular perturbation on nonconvex variational problems. Arch. Ration. Mech. An., 101(3):209–260, 1988.

[20] K. Yosida. Functional Analysis. Classics in Mathematics. Springer-Verlag Berlin Heidelberg, 1995.