Heart Segmentation and Evaluation of Fibrosis

Zhou Zhao

SORBONNE UNIVERSITÉ

DOCTORAL THESIS

Heart Segmentation and Evaluation of Fibrosis

Author:
Zhou ZHAO

Supervisor:
Thierry Géraud
Élodie Puybareau

Reviewer:
Frédérique Frouin, INSERM-Institut Curie
Antoine Vacavant, Université Clermont Auvergne

Examiner:
Isabelle Bloch, Sorbonne Université
Alasdair Newson, Télécom ParisTech
Florence Rossant, Institut Supérieur d'Electronique de Paris
Caroline Petitjean, Université de Rouen Normandie

Invitee:
Jérôme Lacotte, Institut Cardiovasculaire Paris Sud

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

Laboratoire de Recherche et Développement de l'EPITA
École Doctorale Informatique, Télécommunications et Électronique

Acknowledgements

I would like to sincerely thank Frédérique Frouin and Antoine Vacavant, who accepted to review this thesis. I am also grateful to Isabelle Bloch, Alasdair Newson, Florence Rossant, Caroline Petitjean, and Jérôme Lacotte, who agreed to be members of my jury.

Recalling stages of the PhD, there are too many people who need to be thanked.

When I arrived in France the first day, Daniela helped me to check in the dormitory and accompanied me to the supermarket to buy necessities, which made me feel less unfamiliar in the unfamiliar environment, and then she helped me to register at school, all of which were done step by step patiently, thank her sincerely.

I am very grateful to Thierry Géraud and Élodie Puybareau who supervised me during these three years. Whenever I encountered problems that were difficult to solve, they patiently discussed solutions with me and never put any pressure on me. Let me really enjoy the fun of scientific research. At the stages of PhD, I thought I was a lucky person to meet two good supervisors.

I would like to sincerely thank the lab engineer, Clément, who managed the server machine and ensured that the lab was kept in good working condition. I warmly thank my colleagues, Nicolas, Duy, Minh, Yizi, Michaël, Anissa, Florian, Baptiste, Caroline.... I am really lucky to work with them during my PhD.

I am deeply grateful to the China Scholarship Council (CSC) for funding my living expenses, and thanks EPITA (l'école des ingénieurs en intelligence informatique) for funding my accommodation expenses, which made my life less stressful throughout my PhD.

I would like to thank my family and my girlfriend AN Dongjie for supporting me unconditionally.

Abstract

Atrial fibrillation is the most common heart rhythm disease. Due to a lack of understanding in matter of underlying atrial structures, current treatments are still not satisfying. Recently, with the popularity of deep learning, many segmentation methods based on deep learning have been proposed to analyze atrial structures, especially from late gadolinium-enhanced magnetic resonance imaging. However, two problems still occur: 1) segmentation results include the atrial-like background; 2) boundaries are very hard to segment. Most segmentation approaches design a specific network that mainly focuses on the regions, to the detriment of the boundaries.

Therefore, in this dissertation, we propose two different methods to segment the heart, one two-stage and one end-to-end trainable method. For the two-stage method, it can be decomposed in three main steps: a localization step, a Gaussian-based contrast enhancement step, and a segmentation step. This architecture is supplied with a hybrid loss function that guides the network to study the transformation relationship between the input image and the corresponding label in a three-level hierarchy (pixel-, patch- and map-level), which is helpful to improve segmentation and recovery of the boundaries. We demonstrate the efficiency of our approach on three public datasets in terms of regional and boundary segmentations. For the end-to-end trainable method. we propose an attention full convolutional network framework based on the ResNet-101 architecture, which focuses on boundaries as much as on regions. The additional attention module is added to have the network pay more attention on regions and then to reduce the impact of the misleading similarity of neighboring tissues. We also use a hybrid loss composed of a region loss and a boundary loss to treat boundaries and regions at the same time. The efficiency of proposed approach is verified on three public datasets.

Finally, for evaluating the fibrosis degree, we proposed two methods, one is to combine deep learning with morphology, and the other is to use deep learning directly. For the first method, we calculate the left atrial wall based on the segmentation results in the previous chapter by morphologically dilating, and then thresholds to evaluate the fibrosis degree. For the second method, we provide one cascaded UNet architecture and uses multi-modalities information to complete the segmentation of the myocardium, scar and edema. We demonstrate the efficiency of our approach on one public dataset.

Keywords: Deep Learning, Cardiac, Segmentation, Attention, Fully Convolutional Network, Hybrid Loss, Fibrosis Assessment, Morphological Image Processing.

Résumé

La fibrillation auriculaire est la maladie du rythme cardiaque la plus courante. En raison d'un manque de compréhension des structures auriculaires sous-jacentes, les traitements actuels ne sont toujours pas satisfaisants. Récemment, avec la popularité de l'apprentissage profond, de nombreuses méthodes de segmentation basées sur l'apprentissage profond ont été proposées pour analyser les structures auriculaires, en particulier à partir de l'imagerie par résonance magnétique renforcée au gadolinium tardif. Cependant, deux problèmes subsistent : 1) les résultats de la segmentation incluent le fond de type atrial ; 2) les limites sont très difficiles à segmenter. La plupart des approches de segmentation conçoivent un réseau spécifique qui se concentre principalement sur les régions, au détriment des frontières.

Par conséquent, dans cette thèse, nous proposons deux méthodes différentes pour segmenter le cœur, une méthode en deux étapes et une méthode entraînable de bout en bout. La méthode en deux étapes peut être décomposée en trois étapes principales : une étape de localisation, une étape d'amélioration du contraste à base de gaussienne et une étape de segmentation. Cette architecture est dotée d'une fonction de perte hybride qui guide le réseau pour étudier la relation de transformation entre l'image d'entrée et l'étiquette correspondante dans une hiérarchie à trois niveaux (pixel-, patch- et carte), ce qui permet d'améliorer la segmentation et la récupération des frontières. Nous démontrons l'efficacité de notre approche sur trois ensembles de données publiques en termes de segmentations régionales et de frontières. Pour la méthode entraînable de bout en bout, nous proposons un cadre de réseau convolutif complet d'attention basé sur l'architecture ResNet-101, qui se concentre sur les frontières autant que sur les régions. Le module d'attention supplémentaire est ajouté pour que le réseau accorde plus d'attention aux régions et pour réduire l'impact de la similarité trompeuse des tissus voisins. Nous utilisons également une perte hybride composée d'une perte de région et d'une perte de frontière pour traiter les frontières et les régions en même temps. L'efficacité de l'approche proposée est vérifiée sur trois jeux de données publics.

Enfin, pour évaluer le degré de fibrose, nous avons proposé deux méthodes, l'une consistant à combiner l'apprentissage profond avec la morphologie, et l'autre à utiliser directement l'apprentissage profond. Pour la première méthode, nous calculons la paroi auriculaire gauche sur la base des résultats de segmentation du chapitre précédent en dilatant morphologiquement, puis des seuils pour évaluer le degré de fibrose. Pour la seconde méthode, nous fournissons une architecture UNet en cascade et utilisons des informations multi-modalités pour compléter la segmentation du myocarde, de la cicatrice et de l'œdème. Nous démontrons l'efficacité de notre approche sur un jeu de données public.

Mots-clés: Apprentissage profond, cardiaque, segmentation, attention, réseau entièrement convolutif, perte hybride, évaluation de la fibrose, traitement morphologique des images.

Résumé long

Résumé La fibrillation auriculaire est la maladie du rythme cardiaque la plus courante. En raison d'un manque de compréhension des structures atriales sous-jacentes, les traitements actuels ne sont pas encore satisfaisants. Afin d'aider les médecins dans leur diagnostic, avec la popularité de l'apprentissage profond, nous proposons deux méthodes différentes pour segmenter le cœur, une méthode en deux étapes et une méthode entraînable de bout en bout. Ensuite, sur la base des résultats de la segmentation du cœur, nous continuons à évaluer le degré de fibrose en combinant l'apprentissage profond avec la morphologie. Enfin, nous démontrons l'efficacité de notre approche sur un jeu de données public.

1 Introduction

La fibrillation auriculaire (FA) est la maladie du rythme cardiaque la plus courante, correspondant à l'activation d'un substrat électrique au sein du myocarde auriculaire. La FA est déjà une maladie endémique, et sa prévalence est en pleine expansion, en raison de l'augmentation de l'incidence de l'arythmie et de l'augmentation de sa prévalence liée à l'âge. En effet, 1 à 2 % de la population souffre actuellement de FA, et le nombre de personnes touchées devrait doubler ou tripler au cours des deux ou trois prochaines décennies, tant en Europe qu'aux États-Unis [1].

Au cours des dernières années, plusieurs groupes ont testé la capacité du LGE-CMR à détecter une fibrose préexistante. Bien que ces rapports suggèrent que l'étendue de la fibrose peut prédire les récidives après les procédures d'ablation, l'absence de reconstruction automatisée en 3D du LA, le manque de valeurs de référence pour la normalité ont conduit à la publication de plusieurs protocoles d'acquisition et de post-traitement d'images et de seuils pour identifier la fibrose, limitant finalement la validation externe et la reproductibilité de cette technique.En raison de ces limites techniques [2–4], l'évaluation de la fibrose du LA n'a pas encore été largement adoptée dans la pratique clinique [5].

Actuellement, avec la popularité de l'apprentissage profond, certaines méthodes basées sur l'apprentissage profond ont été proposées pour segmenter l'oreillette et évaluer la fibrose LA. Par exemple, Bai et al. [6] et Vigneault et al. [7] ont conçu un cadre de réseau basé sur les FCN 2D pour segmenter directement l'oreillette gauche et droite. En outre, les réseaux proposés peuvent également être appliqués pour segmenter les ventricules après la formation sans changer de cadre de réseau. De

même, Xiong et al. [8], Preetha et al. [9], Bian et al. [10] et Chen et al. [11] ont également conçu un cadre de réseau de segmentation basé sur les FCN 2D pour segmenter l'oreillette. Yang et al. [12, 13] ont utilisé une méthode basée sur un atlas pour identifier l'oreillette gauche, puis ont utilisé un réseau d'apprentissage profond pour détecter les tissus fibrotiques dans la zone de l'oreillette gauche. En relation avec la méthode de segmentation de bout en bout, Chen et al. [14] ont proposé un réseau neuronal profond pour segmenter à la fois l'oreillette gauche et les cicatrices auriculaires.

Par conséquent, pour aider les médecins à établir un diagnostic et réduire leur charge de travail, nous proposons de nouveaux cadres de réseaux neuronaux pour segmenter l'oreillette et évaluer la fibrose de l'oreillette. Tout d'abord, nous proposons deux méthodes différentes pour segmenter le cœur, une méthode en deux étapes et une méthode entraînable de bout en bout. La méthode en deux étapes peut être décomposée en trois étapes principales : une étape de localisation, une étape de renforcement du contraste à base de gaussienne et une étape de segmentation. Pour la méthode entraînable de bout en bout, nous proposons un cadre de réseau convolutif complet d'attention basé sur l'architecture ResNet-101, qui se concentre sur les frontières autant que sur les régions. Le module d'attention supplémentaire est ajouté pour que le réseau accorde plus d'attention aux régions et pour réduire l'impact de la similarité trompeuse des tissus voisins. Ensuite, sur la base des résultats de la segmentation cardiaque, nous combinons l'apprentissage profond avec la morphologie pour évaluer la fibrose de l'oreillette gauche. Nous calculons la paroi de l'oreillette gauche à partir des résultats de la segmentation de l'oreillette gauche par dilatation morphologique, puis nous fixons des seuils pour évaluer le degré de fibrose. Enfin, nous démontrons l'efficacité de notre approche sur certains jeux de données publics.

2 Méthodes de Segmentation du Cœur

2.1 Méthode en Deux Étapes

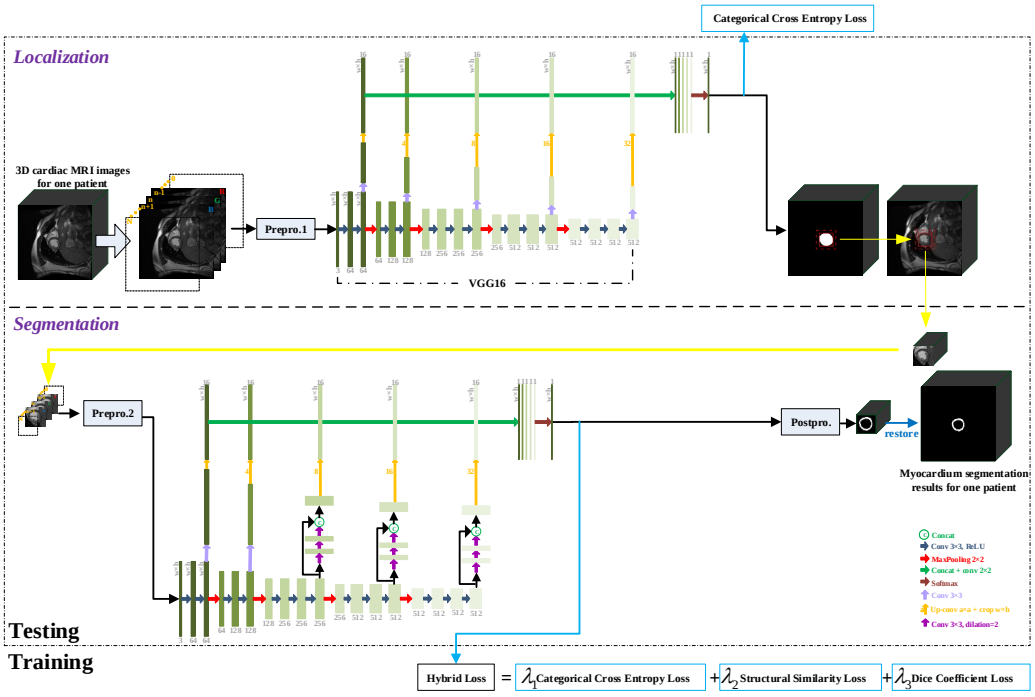


FIGURE 1: Architecture des réseaux à deux étages.

La vue d'ensemble de nos réseaux se compose de deux parties (localisation et segmentation) comme le montre la Fig. 1. La première partie (le "réseau de localisation") est utilisée pour localiser approximativement la position de l'objet. La seconde partie est consacrée à la segmentation de l'objet (le "réseau de segmentation"). **Réseau de localisation:** Tout d'abord, nous nous appuyons sur l'architecture originale du réseau VGG16 [15], pré-entraîné sur des millions d'images naturelles d'ImageNet pour la classification d'images [16]. Nous éliminons ensuite ses couches entièrement connectées pour ne conserver que le sous-réseau composé de cinq "étages" basés sur la convolution (le réseau de base). Chaque étage est composé de deux couches convolutionnelles, d'une fonction d'activation ReLU et d'une couche de max-pooling. Comme les couches de max-pooling diminuent la résolution de l'image d'entrée, nous obtenons un ensemble de cartes de caractéristiques fines à grossières (avec 5 niveaux de caractéristiques). Inspirés par les travaux de [17–20], nous avons ajouté des couches convolutionnelles *spécialisées* (avec un noyau de taille 3×3) avec K (par exemple $K = 16$) cartes de caractéristiques après les couches convolutionnelles ascendantes placées à la fin de chaque étape. Les sorties des couches spécialisées présentent la même résolution que l'image d'entrée, et sont concaténées ensemble. Nous ajoutons une couche convolutionnelle 1×1 à la sortie de la couche de concaténation pour combiner linéairement les cartes de caractéristiques fines à grossières.

Réseau de segmentation: Par rapport aux travaux sur la localisation nous ajoutons trois couches convolutionnelles avec 256 ou 512 dilatés (dilatation = 2) [21] 3×3 filtres, et une couche de concaténation dans le réseau de segmentation basé sur le réseau de localisation précédent. réseau de localisation précédent.

2.1.1 Perte hybride

Pour obtenir une segmentation régionale de haute qualité, nous définissons ℓ_R comme une perte de région : $\ell_R = \ell_{CCE} + \ell_{SSIM} + \ell_{DC}$, où ℓ_{CCE}, ℓ_{SSIM} et ℓ_{DC} désignent respectivement la perte d'entropie croisée catégorielle (CCE) [22], la perte de similarité structurelle (SSIM) [23] et la perte de coefficient de dés (DC) [24].

La perte CCE [22] est couramment utilisée pour la classification et la segmentation multi-classes. Elle est définie comme suit:

$$\ell_{CCE} = -\sum_{i=1}^{C}\sum_{a=1}^{H}\sum_{b=1}^{W} y^{i}_{(a,b)} \ln {y_*}^{i}_{(a,b)}, \tag{1}$$

où C est le nombre de classes de chaque image, H et W sont la hauteur et la largeur de l'image, $y^{i}_{(a,b)} \in \{0,1\}$ est l'étiquette de vérité du sol à un coup de la classe i à la position (a,b). et ${y_*}^{i}_{(a,b)}$ est la probabilité prédite que (a,b) appartient à la classe i.

La perte SSIM peut évaluer la qualité de l'image [23], et peut être utilisée pour capturer l'information structurelle, ce qui diminuera le taux de mauvaise segmentation des tissus similaires environnants. Par conséquent, nous l'avons intégré dans notre perte d'apprentissage pour apprendre les différences entre le domaine segmenté et les tissus similaires autour du domaine segmenté. Si **S** et **G** sont respectivement la carte de probabilité prédite et le masque de vérité terrain, la fonction de perte SSIM de **S** et **G** est définie comme suit

$$\ell_{SSIM} = 1 - \frac{(2\mu_S\mu_G + \varepsilon_1)(2\sigma_{SG} + \varepsilon_2)}{(\mu_S^2 + \mu_G^2 + \varepsilon_1)(\sigma_S^2 + \sigma_G^2 + \varepsilon_2)} \tag{2}$$

où μ_S, μ_G et σ_S, σ_G sont les moyennes et les écarts types de **S** et **G** respectivement, σ_{SG} est leur covariance, $\varepsilon_1 = 0.01^2$ et $\varepsilon_2 = 0.03^2$ sont utilisés pour éviter une division par zéro.

La perte DC [24] est utilisée pour mesurer la similarité entre deux ensembles comme défini dans Eq. 2.36. Mais pour la tâche de segmentation multi-classes, Eq. 2.36 ne convient pas en raison du problème de déséquilibre des classes dans de tels cas. Par conséquent, nous étendons la définition de la perte DC à la segmentation multi-classes comme suit:

$$dice_i = \left(\epsilon + 2\sum_{n=1}^{N_i} y^{i}_{n}\, {y_*}^{i}_{n}\right) / \left(\epsilon + \sum_{n=1}^{N_i}\left(y^{i}_{n} + {y_*}^{i}_{n}\right)\right) \tag{3}$$

$$\ell_{DC} = 1 - \sum_{i=1}^{C} dice_i / \left(N_i + \epsilon\right), \tag{4}$$

où N_i désigne les numéros de la classe i et $\epsilon > 0$ est un facteur lisse.

2.2 Méthode de Bout en Bout

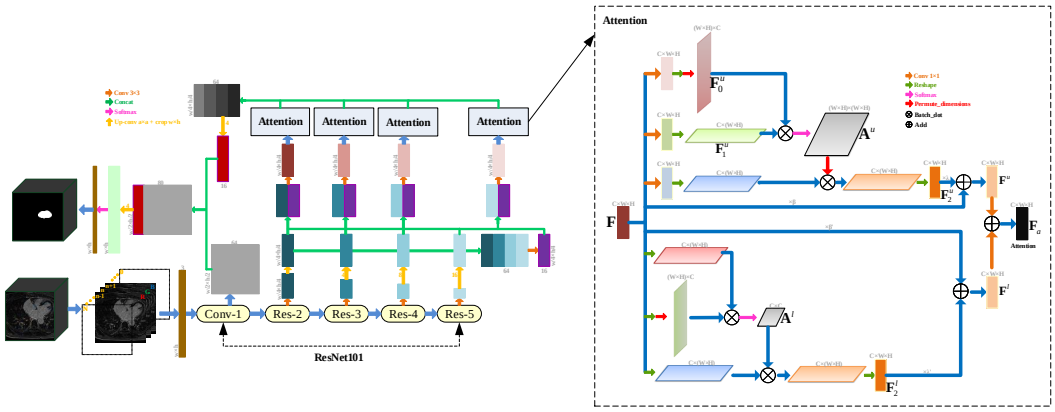


FIGURE 2: Architecture des réseaux de bout en bout.

Nous proposons un nouveau réseau d'attention (voir Fig. 2) utilisant ResNet-101 pré-entraîné sur ImageNet [25] pour calculer les cartes de caractéristiques. Nous éliminons ses couches de mise en commun des moyennes et ses couches entièrement connectées, et ne conservons que le sous-réseau composé d'une étape basée sur la convolution et de quatre "étapes" basées sur les résidus. Comme la résolution diminue à chaque étape, nous obtenons un ensemble de cartes de caractéristiques fines à grossières (avec cinq niveaux de caractéristiques). Nous ajoutons des couches convolutionnelles *spécialisées* (avec un noyau de taille 3×3) avec K (par exemple $K = 16$) cartes de caractéristiques placées à la fin de quatre "étages" basés sur les résidus. Elles sont concaténées ensemble après les couches convolutionnelles ascendantes. Ces dernières cartes de caractéristiques sont combinées avec chacune des sorties des couches spécialisées, puis introduites dans le module d'attention pour générer les caractéristiques d'attention. Enfin, nous concaténons les caractéristiques d'attention avec les sorties de *Conv1* et nous les introduisons dans la couche softmax.

Le module d'attention est inspiré de [26]. $\mathbf{F} \in \mathbb{R}^{C\times W\times H}$ agit comme une carte de caractéristiques d'entrée pour le module d'attention, où C, W, H sont respectivement le canal, la largeur et la hauteur de la carte de caractéristiques. La branche supérieure **F** est alimentée dans une couche convolutive, une couche de Reshape et ensuite une couche de Transpose, résultant en une carte de caractéristiques $\mathbf{F}_0^u \in \mathbb{R}^{(W\times H)\times C}$. Dans la deuxième branche (considérons l'ordre de haut en bas), la carte de caractéristiques d'entrée **F** suit les mêmes opérations moins la couche Transpose, ce qui donne $\mathbf{F}_1^u \in \mathbb{R}^{C\times(W\times H)}$. Ensuite, les couches Multiply et Softmax suivent ; elles sont appliquées sur $\mathbf{F}_0^u$ et $\mathbf{F}_1^u$ pour obtenir la carte d'attention spatiale $\mathbf{A}^u \in \mathbb{R}^{(W\times H)\times(W\times H)}$. L'entrée F est introduite dans une couche convolutive différente dans la troisième branche, puis elle est multipliée par A u introduite dans la couche Transpose, ce qui donne $\mathbf{F}_2^u$. Par conséquent, la sortie $\mathbf{F}^u$ de la branche supérieure peut être formulée comme suit: $\mathbf{F}^u = \lambda \times \mathbf{F}_2^u + \beta \times \mathbf{F},$, où $\lambda \in \mathbb{R}^C$ est initialisé à [0,..,0], et $\beta \in \mathbb{R}^C$ est

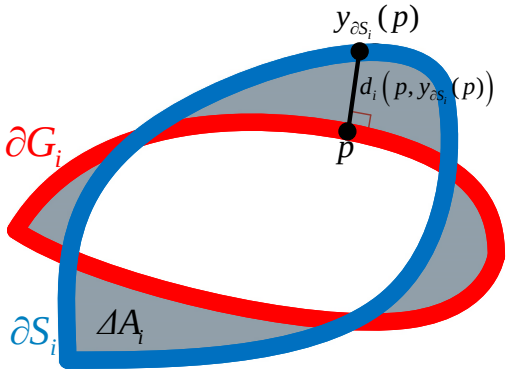


FIGURE 3: Illustration du calcul de la perte aux frontières

initialisé à [1,..,1]. Les valeurs λ et β sont utilisées pour apprendre progressivement l'importance de la carte d'attention spatiale.

Dans la branche inférieure, le module d'attention se concentre principalement sur les canaux les plus importants. La carte d'attention des canaux $\mathbf{A}^l$ peut être obtenue par différentes combinaisons de convolution. Enfin, la sortie $\mathbf{F}^l$ de la branche inférieure peut être définie comme suit: $\mathbf{F}^l = \lambda' \times \mathbf{F}_2^l + \beta' \times \mathbf{F},$, où $\lambda' \in \mathbb{R}^C$ est initialisé à [0,..,0], et $\beta' \in \mathbb{R}^C$ est initialisé à [1,..,1]. La carte de caractéristiques $\mathbf{F}_2^l$ dénote les résultats du produit de l'entrée $\mathbf{F}$ avec $\mathbf{A}^l$ alimenté dans une convolution passant par le bloc de transposition. Par conséquent, la carte de caractéristiques d'attention $\mathbf{F}_a$ est définie comme :

$$\mathbf{F}_a = Conv\left(\mathbf{F}^u\right) + Conv\left(\mathbf{F}^l\right). \tag{5}$$

2.2.1 Perte hybride

La perte hybride se compose de deux parties : la perte de région et la perte de frontière. Elle est définie comme suit : $\ell_{\mathrm{H}} = \ell_{\mathrm{R}} + \ell_{\mathrm{B}}$, où ℓ_{B} la perte de frontière. Elles sont expliquées ci-après.

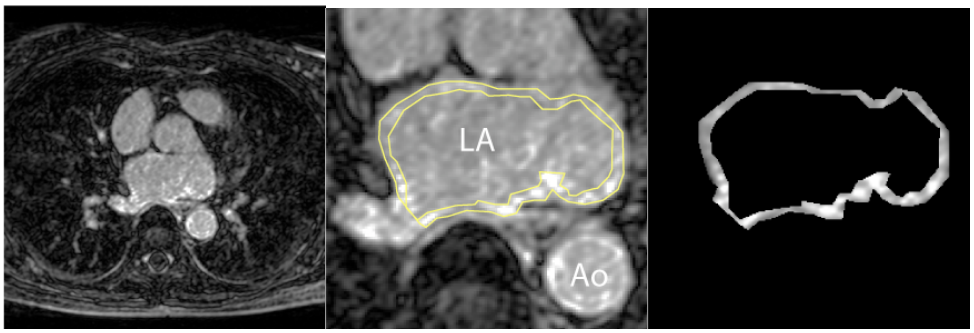
Les fonctions de perte mentionnées précédemment sont principalement destinées à la segmentation de régions, nous proposons donc une fonction de perte de frontière multi-classe basée sur la distance de Kervadec [27] pour pouvoir affiner les segmentations. Comme le montre la Fig. 3, ΔA désigne la différence entre la frontière $\mathbf{G}_{\mathbf{B}}^i$ de la vérité terrain de la classe i et la frontière $\mathbf{S}_{\mathbf{B}}^i$ de la prédiction de la classe i. Lorsque ΔA tend vers zéro, cela signifie que les résultats de la segmentation deviennent meilleurs autour des frontières. Ainsi, pour une classe i donnée, lorsque la prédiction et la vérité terrain sont suffisamment proches, ce qui est facilement obtenu grâce à notre perte régionale, la minimisation de la différence entre leurs frontières peut être obtenue en minimisant la distance de Kervadec [27]:

$$\ell_{\mathrm{B}}^i = \int_{\partial G_i} \left\| y_{\partial S_i}(p) - p \right\|^2 \mathrm{d}p \tag{6}$$

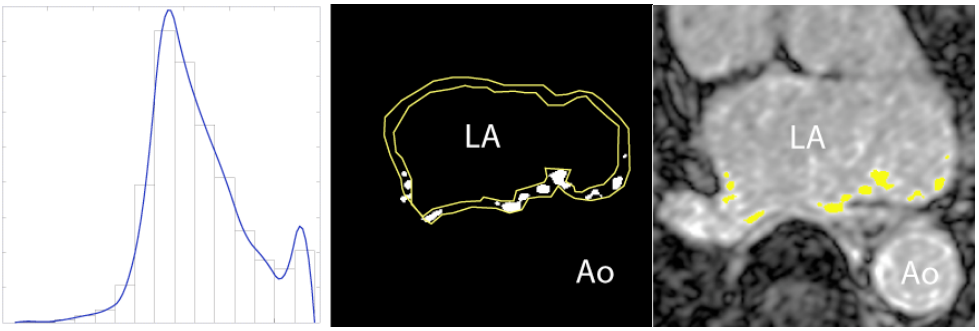
où ∂G_i et ∂S_i désignent les frontières de $\mathbf{G}_\mathbf{B}^i$ et (binarisé) $\mathbf{S}_\mathbf{B}^i$ et $\|\cdot\|$ désigne la norme L2. Lorsque p est un point dans ∂G_i, $y_{\partial S_i}(p)$ désigne le point correspondant sur la frontière ∂S_i i le long de la direction normale à∂G_i (voir Fig. 3). On peut montrer [27] que minimiser ℓ_B^i est équivalent à minimiser l'aire de la surface $\Delta A_i = (\mathbf{G}_\mathbf{B}^i \backslash \mathbf{S}_\mathbf{B}^i) \bigcup (\mathbf{S}_\mathbf{B}^i \backslash \mathbf{G}_\mathbf{B}^i)$ (voir Fig. 3). Ainsi, notre perte de frontière multi-classe s'ensuit naturellement:

$$\ell_\mathrm{B} = \sum_{i=1}^{C} \int_{\partial G_i} \left\| y_{\partial S_i}(p) - p \right\|^2 \mathrm{d}p \tag{7}$$

3 Méthodes d'évaluation de la Fibrose



(a) MRI d'entrée (b) Contours myocardiques (c) Paroi auriculaire gauche



(d) Histogramme (e) 3SD seuil (f) Détection

FIGURE 4: Schéma du processus proposé.

Sur la base des méthodes de segmentation précédentes, nous continuons à évaluer la fibrose en utilisant une méthode de morphologie. La Fig. 4 présente le workflow attendu : segmentation du volume cardiaque conduisant à l'identification de la paroi auriculaire gauche, analyse de la radiométrie au sein de la paroi, seuillage pour quantifier le degré de fibrose. La segmentation cardiaque peut être complétée par les méthodes de segmentation précédentes. La partie analyse peut s'appuyer sur une approche de morphologie mathématique.

4 Résultats Expérimentaux

Nous évaluons notre méthode sur le MICCAI 2018 Atrial Segmentation Challenge [1](AtriaSeg18). Son objectif est de segmenter l'oreillette gauche. Il contient 100 3D MRIs annotées provenant de patients souffrant de fibrillation auriculaire. L'espacement des pixels des MR images est de $0.625 \times 0.625 \times 0.625$ mm/pixel. L'ensemble de données comprend deux tailles d'images différentes: $88 \times 576 \times 576$ et $88 \times 640 \times 640$.

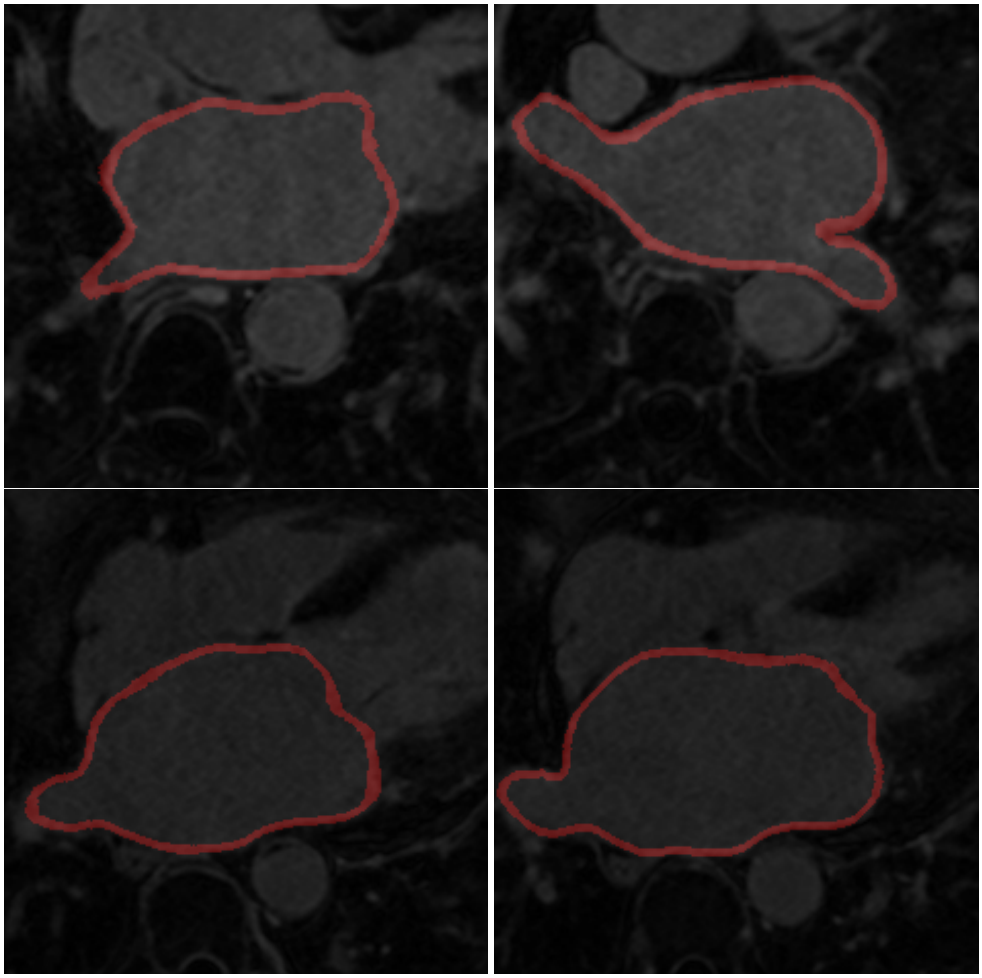


FIGURE 5: Segmentation de la paroi de l'oreillette gauche

Le jeu de données AtriaSeg18 fournit l'étiquette de la cavité auriculaire gauche (LA), de sorte que les résultats de la segmentation du cœur du jeu de données AtriaSeg18 sont la cavité auriculaire gauche (LA) et le bord endocardique est obtenu. Ensuite, la limite endocardique est dilatée morphologiquement (par 4 couches de pixels, 2,5 mm), puis ajustée manuellement pour créer la coquille de la surface épicardique de l'oreillette gauche [28]. Dans une dernière étape, la segmentation de l'endocarde est soustraite de la couche épicardique pour définir la segmentation de la paroi, comme indiqué sur la Fig 5.

[1]http://atriaseg2018.cardiacatlas.org/

Après avoir obtenu la segmentation de la paroi, nous supposons que l'image ne comprend que l'oreillette gauche A est défini comme $A = ES \times I$, où ES désigne le résultat de la segmentation endocardique (image binaire) et I désigne l'image grise du cœur. Ensuite, nous calculons la valeur moyenne M et écart-type SD de $A > 0$, et le seuil est fixé à $M + 3SD$. Enfin, la fibrose est ob Enfin, la fibrose est détectée par $W > (M + 3SD)$ (W indique que l'image ne comprend que la paroi de l'oreillette gauche), comme le montre la Fig 6.

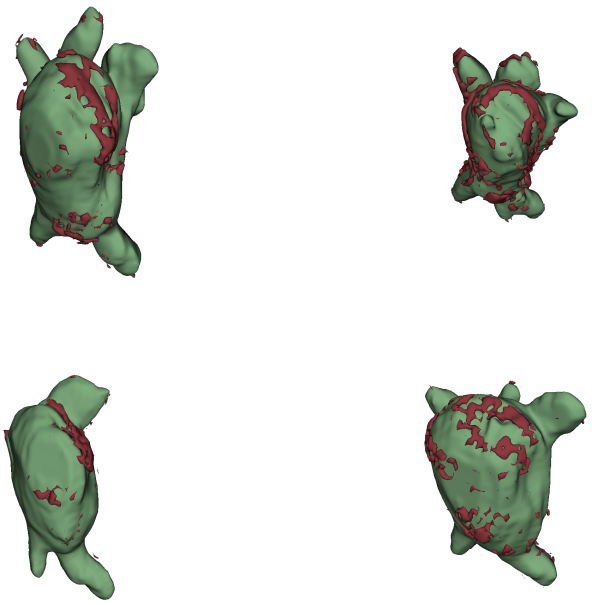


FIGURE 6: Vue 3D de la fibrose et de la paroi de l'oreillette gauche; la couleur rouge indique la fibrose et la couleur verte la paroi de l'oreillette gauche.

Comme le jeu de données AtriaSeg18 ne fournit pas le label de la cicatrice, nous continuons à tester notre méthode sur le Left Atrial and Scar Quantification & Segmentation Challenge (LAScarQS2022) [29–31], et le LAScarQS2022 vise à segmenter l'oreillette gauche et à évaluer la cicatrice. Il comprend deux tâches (Tâche 1 et Tâche 2) et la Tâche 1 contient les données de cicatrices, nous utilisons donc uniquement le jeu de données de la Tâche 1. La tâche 1 contient 60 IRM 3D annotées de patients souffrant de fibrillation auriculaire pour l'entraînement et la validation. La taille des voxels des images IRM est différente: $1.25 \times 1.25 \times 2.5$ mm, $1.4 \times 1.4 \times 1.4$ mm et $1.3 \times 1.3 \times 4.0$ mm. L'ensemble de données comprend deux tailles d'images différentes: $44 \times 576 \times 576$ pixels et $44 \times 640 \times 640$ pixels.

TABLE 1: Étude d'ablation de SD sur le jeu de données LAScarQS2022 en utilisant une validation croisée 5 fois.

SD différente	DC de la cicatrice
SD	0.328±0.035
2SD	0.305±0.067
3SD	0.062±0.038

5 Conclusion

Nous combinons des méthodes d'apprentissage profond largement utilisées avec une approche de morphologie mathématique pour segmenter le cœur et évaluer la fibrose. Le temps de calcul de l'ensemble du pipeline est inférieur à 4 secondes pour un volume 3D entier, ce qui le rend utilisable dans la pratique clinique.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

1.1 Medical Context

The heart is an organ that supplies blood and oxygen to all parts of the body. As shown in Fig.1.1, it is a hollow organ composed mainly of cardiac muscle and comprising the left atrium, left ventricle, right atrium, and right ventricle. The left ventricle is connected to the aorta, the right ventricle to the pulmonary artery, the left atrium to the pulmonary vein, and the right atrium to the superior and inferior vena cava. Both the left and right atria and the left and right ventricles are separated by the septum, so they are not connected. There are valves between the atria and the ventricles. These valves ensure that blood flows only from the atria to the ventricles, but not back. The heart and the circulatory system together form the cardiovascular system.

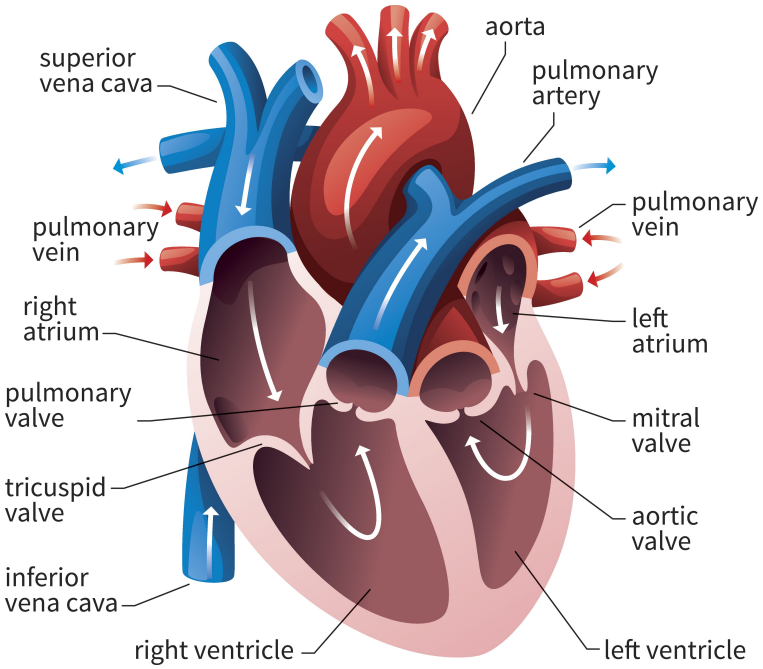


FIGURE 1.1: The anatomy of the heart[1]

[1]https://www.thoughtco.com/thmb/Z5FC4sAv87cdE4L6tyTyybL9ww8=/4402x3856/filters:fill(auto,1)/human-heart-circulatory-system-598167278-5c48d4d2c9e77c0001a577d4.jpg

The diagnosis and treatment of heart rhythm disorders depend increasingly on medical imaging from various scanners. For example, 1) Calcium deposits in plaque are discovered via a computerized tomography (CT) scan, commonly known as a calcium-score screening heart scan, in patients with heart disease. They are the most effective way to detect atherosclerosis before symptoms appear. Atherosclerosis of the coronaries increases with coronary calcium levels. 2) A magnetic resonance imaging (MRI) scan is a non-invasive procedure that employs radio and magnetic waves produced by an MRI scanner to provide precise images of the interior of your heart. It is used to detect congenital heart disease, cardiomyopathy, heart valve disease, and other conditions. 3) A heart positron emission tomography (PET) scan is a noninvasive nuclear imaging test. It creates images of your heart by using radioactive tracers (called radionuclides). Cardiovascular PET scans are used by doctors to diagnose coronary artery disease and heart attack damage. PET scans can distinguish between healthy and damaged heart muscle. PET scans can also help determine whether you will benefit from a percutaneous coronary intervention, such as angioplasty and stenting, coronary artery bypass surgery, or another procedure. 4) A noninvasive nuclear imaging test is a single-photon emission computerized tomography (SPECT) scan of the heart. It creates images of your heart by injecting radioactive tracers into your blood. SPECT is used by doctors to diagnose coronary artery disease and determine whether or not a heart attack has occurred. SPECT imaging can reveal how well blood flows to the heart and how well the heart functions. 5) Echocardiogram, cardiac echo, and transthoracic echo are all terms for ultrasound (TTE). It creates a moving image of the heart by using ultrasonic waves that bounce off it. It allows doctors to see the heart in motion, including the heartbeat. It is most effective for detecting heart structure and function abnormalities such as dilated cardiomyopathy or restrictive cardiomyopathy. It also aids in the detection of cardiac chamber enlargement, irregular heart rhythms, and heart valve disease.

Atrial fibrillation (AF), as shown in Fig.1.2, is the most common heart rhythm disease, corresponding with the activation of an electrical substrate within the atrial myocardium. AF is already an endemic disease, and its prevalence is soaring, due to both an increasing incidence of the arrhythmia and an age-related increase in its prevalence. Indeed, 1–2% of the population suffer from AF at present, and the number of affected individuals is expected to double or triple within the next two to three decades both in Europe and in the USA [1].

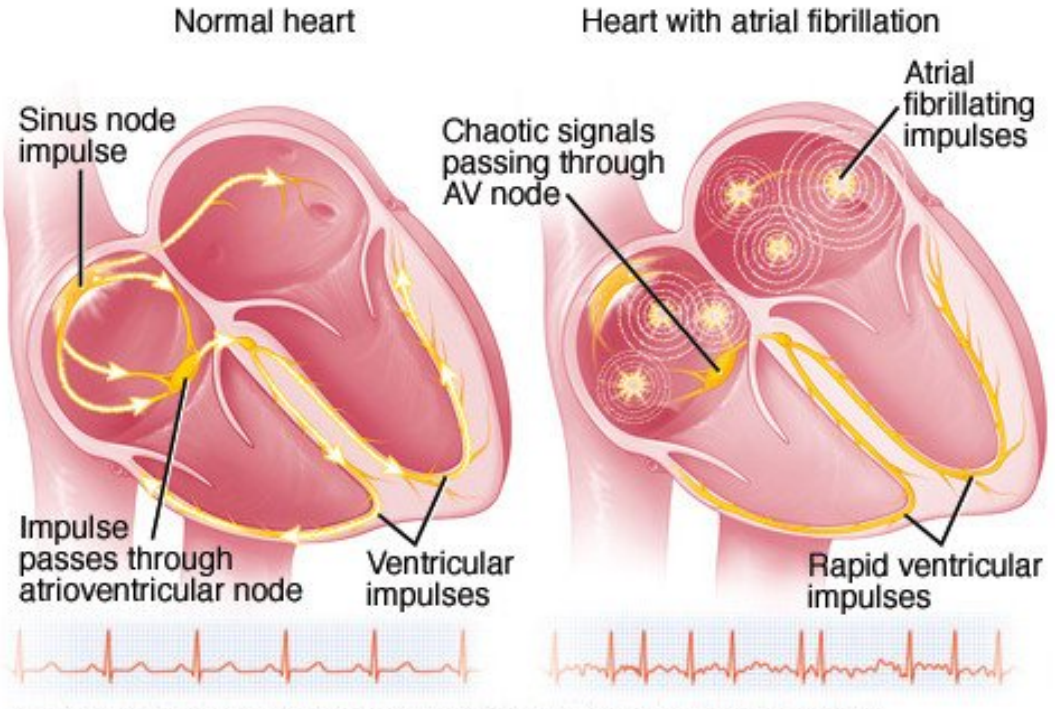


FIGURE 1.2: Atrial fibrillation[2]

Due to the limited effects of anti-arrhythmic drugs, AF can only be cured by percutaneous radiofrequency catheter ablation (CA) targeting triggers and critical areas responsible for AF perpetuation in left atrium (LA). Identification and quantification of AF electrical substrate prior to AF ablation remains an unsolved issue as the number of targets remains unpredictable using clinical criterias. AF CA is still a challenging intervention requiring a perioperative 3D mapping to identify AF substrate to select the best ablation strategy [1].

Exploration of LA substrate has suggested that AF may be a self-perpetuating disease with a voltage or electrogram (EGM) amplitude reduction which is an indicator of the severity of tissue corresponding with collagen deposition in the myocardial interstitial space. Non-invasive assessment of myocardial fibrosis has proved useful as a diagnostic, prognostic, and therapeutic tool. Visualization and quantification of gadolinium in late gadolinium-enhanced cardiac magnetic resonance (LGE-CMR) sequences estimate the extracellular matrix volume and have been used as a LA fibrosis surrogate [36].

[2]https://twitter.com/MayoClinic/status/1007688695590342657/photo/1

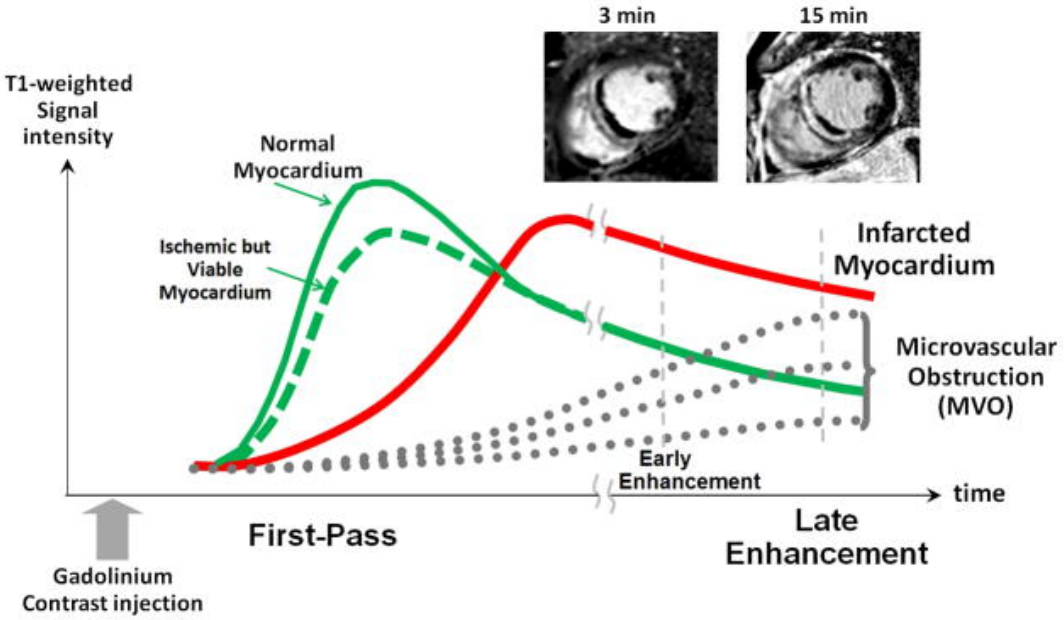


FIGURE 1.3: The principle of late gadolinium-enhanced cardiac magnetic resonance (LGE-CMR) [32]

As shown in Fig. 1.3, following administration of a bolus of gadolinium contrast agent, the contrast will reach the various tissue compartments within the myocardium at different rates until a dynamic steady state is reached. Signal intensity time course from administration of bolus through late enhancement illustrating slower wash-in and wash-out of gadolinium contrast into infarcted tissue compared with normal ischemic tissue. Late enhancement imaging is typically performed 10–30 minutes following administration of gadolinium when there is sufficient contrast between normal and infarcted tissue. T1-weighted, T2-weighted and Fluid Attenuated Inversion Recovery (FLAIR) scans are the most common LGE-CMR sequences [37]. T1-weighted images are produced by using short Time to Echo (TE)[3] and Repetition Time (TR)[4] times. The contrast and brightness of the image are predominately determined by T1 properties of tissue. Conversely, T2-weighted images are produced by using longer TE and TR times. In these images, the contrast and brightness are predominately determined by the T2 properties of tissue. The FLAIR sequence is similar to a T2-weighted image except that the TE and TR times are very long, and it is very sensitive to pathology.

Over the last years, several groups tested the ability of LGE-CMR to detect preexisting fibrosis. Although these reports suggested that the extent of fibrosis may predict recurrences after ablation procedures, the lack of 3D automated LA reconstruction, the lack of reference values for normality has prompted the publication

[3]Time to Echo (TE) is the time between the delivery of the Radio Frequency (RF) pulse and the receipt of the echo signal.

[4]Repetition Time (TR) is the amount of time between successive pulse sequences applied to the same slice.

of several image acquisition and post-processing protocols and thresholds to identify fibrosis, eventually limiting the external validation and reproducibility of this technique [2–4].

Because of these technical limits, the assessment of LA fibrosis has not yet been widely adopted in the clinical practice [5]. The aims of this project involving EPITA and the Institut Cardiovasculaire Paris Sud (ICPS) are to provide a normalized, systematic, consistent, reproducible and automatically 3D LA LGE-CMR reconstruction to identify LA fibrotic tissue prior to AF ablation.

1.2 Traditional Methods for Left Atrial Fibrosis/Scar Segmentation

Most traditional methods present the expected workflow for left atrial fibrosis or scar segmentation as follows:

(1) Segmentation of the heart volume leading to the identification of the left atrial wall

(2) Analysis of the radiometry within the wall, thresholding to quantify the fibrosis degree.

For the first step, many methods are applied such as level-set [38], region growing [39] and watershed [40] and so on (as shown in Fig.1.4). In term of the **level-set** method [38], the whole process was divided into two steps. First, the median filter was used to obtain the velocity image, and then the gradient magnitude filter was used to process the velocity image. Taking segmentation of the atrial wall as an example, first used the median filter to process the atrial images. After obtaining the velocity images, the gradient magnitude filter detected the edge zones with sharp gradients around the epicardial boundary, and then stops at these edge zones. Finally, the atrial wall is obtained by subtracting the result of the level set method from the endocardium segmentation mask. For the **Region growing** method [39], it was also often used in medical image segmentation tasks. It achieved segmentation by placing seed points in the segmentation area and choosing different thresholds according to different situations. As shown in the Fig.1.4, because the threshold depended on different situations, the final segmentation result was not stable, and the method had no independent adjustment ability. For the **watershed** segmentation [40], in medical image segmentation, mainly used the image as a topographic surface and markers for controlling. The seed points were placed in the target segmentation area and the adjacent area similar to the target structure.

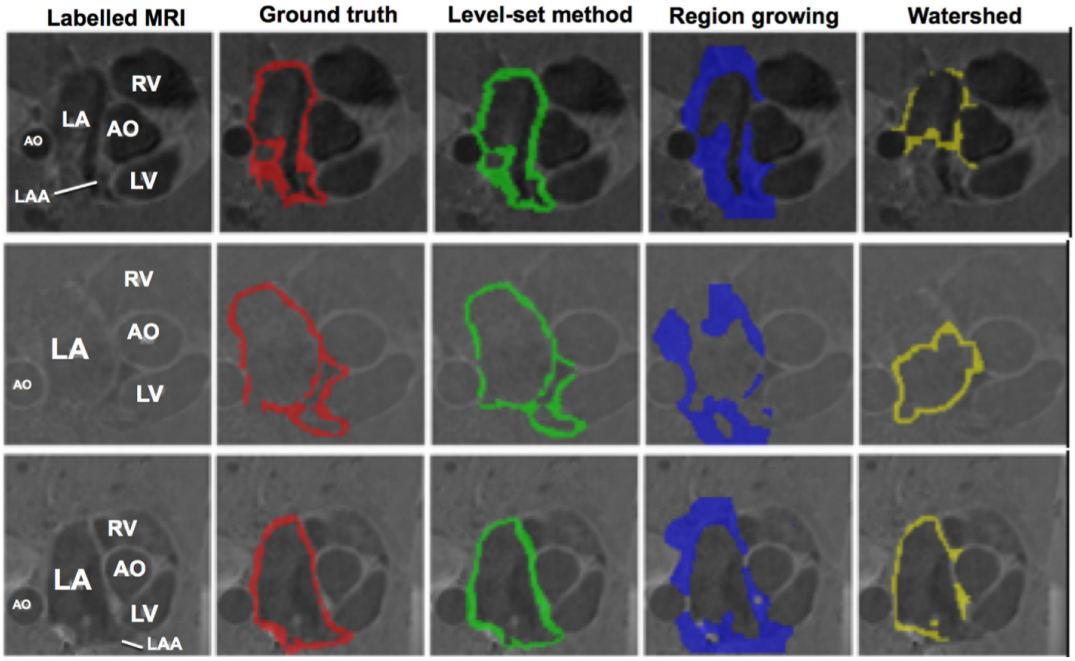


FIGURE 1.4: From the MRI image database. Each row represents a separate case. Each column represents (from left to right): original MRI, segmentation for ground truth, level-set method, region-growing and watershed segmentation. Abbreviations: LA – left atrium, AO – aorta, LV – left ventricle, RV – right ventricle, LAA – left atrial appendage [33]

For the second step, based on the left atrial wall, many fibrosis or scar detection and segmentation algorithms are proposed. Firstly, we give an overview of the previously published fibrosis or scar detection and segmentation algorithms (as shown in Table. 1.1). It could be concluded from Table. 1.1 that for detecting fibrosis or scars, most researchers chose the SD algorithm [41–43]. But other methods also had certain advantages. For example, FWHM [44] was further used to classify scars as cores or peri-core areas [45], and other methods had been proposed to automatically calculate thresholds [42] such as clustering [46, 47], and Graph-cuts [48]. MIP algorithm was used to visualize the infarcted area [49], which was very useful for visualizing the number of scars on the surface of the atrial. For the detection of fibrosis before ablation, the global threshold of the image could be calculated and adjusted according to the data of each slice to achieve the best detection effect [3].

All the methods in Table. 1.1 except [3] and [49] can be used to detect scars in the myocardium. But facing the task of scar segmentation, many difficulties need to be solved urgently, especially the nearby enhanced structures such as the aortic wall and valves. There are also differences between the atrial myocardium and the ventricular myocardium. For example, the thickness of the atrial myocardium is thinner than that of the ventricle, and it is more difficult to segment. Therefore, only using some fixed models to detect the scar of the atrial myocardium cannot achieve good results. Some researchers have used it, but we still think that it is not suitable for scar segmentation of the ventricular myocardium. The reason is simple:

TABLE 1.1: Overview of previously published scar detection and segmentation methods [35]

Reference	Model	Modality	LV/LA	Algorithm
Oakes et al.[3]	Human	CMR	LA	SD
Kim et al.[41]	Canine	CMR	LV	SD
Kolipaka et al.[42]	Human	CMR	LV	SD
Schmidt et al..[43]	Human	CMR	LV	SD
Amado et al.[44]	Animal	CMR	LV	FWHM
Yan et al.[45]	Human	CMR	LV	SD
Positano et al.[46]	Human	CMR	LV	Clustering
Detsky et al.[47]	Human	CMR	LV	Clustering
Lu et al.[48]	Human	CMR	LV	Graph-cuts
Knowles et al.[49]	Human	CMR	LA	MIP
Hennemuth et al. [50]	Human	CMR	LV	EM
Tao et al.[51]	Human	CMR	LV	Otsu

Note: LV denotes Left ventricle and LA denotes left atrium. Most methods employed simple standard deviation (SD) thresholding from a base healthy tissue intensity value. Others such as full-width-at-half-maximum (FWHM), maximum intensity projection (MIP) and expectation-maximisation (EM) fitting have also been proposed.

using a single fixed model cannot deal with all the different variables encountered randomly. These variables may come from outside (image resolution, noise, image acquisition time, etc.), and it may also come from the inside (the size and shape of the scar, etc.). This fact has been supported in [3] that in order to obtain a suitable segmentation, the threshold must be constantly re-adjusted according to the data on each slice.

1.3 State-of-the-art for Left Atrial Fibrosis/Scar Segmentation

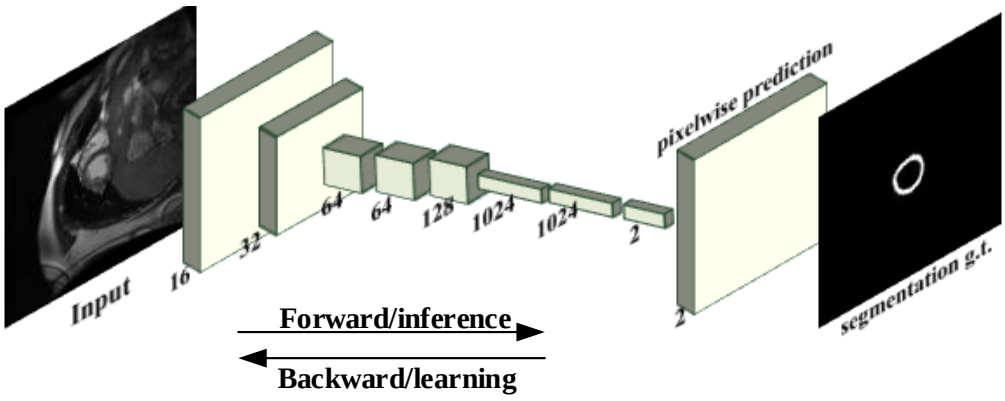


FIGURE 1.5: Fully convolutional networks

Among these traditional methods mentioned above, the most common method offers good accuracy for fibrosis or scar detection and segmentation, but often looses efficiency due to heavy calculations with the registration algorithm. Recently, deep learning becomes more and more famous and is used in many field. Many researchers have combined deep learning approaches with traditional segmentation methods for the purpose of scar segmentation. For the deep learning methods in the field of medical image segmentation, most of the proposed network frameworks are mainly based on fully convolutional networks (FCNs) [52] or on U-Net [53], as shown in Fig.1.5 and Fig.1.6. They use upsampling layers and combine the feature maps from lower to higher resolutions. Many extensions to these networks have been proposed.

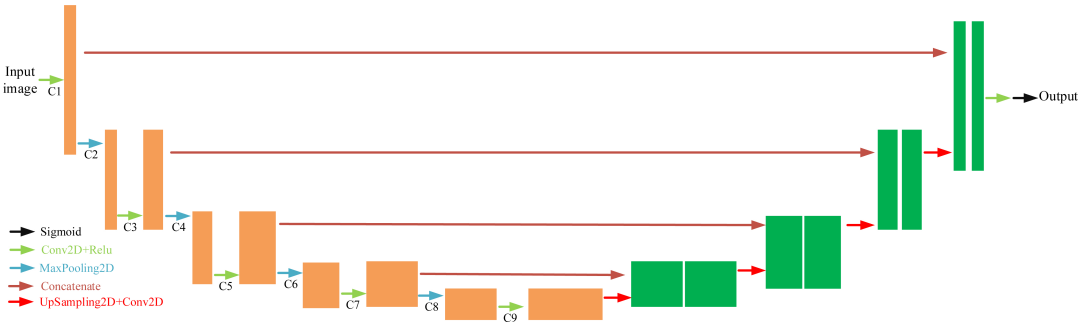


FIGURE 1.6: U-net architecture

Firstly, some researchers mainly focus on atrial segmentation, because it can be used as a basis for scar segmentation and atrial fibrosis quantification from LGE images. For example, Bai et al. [6] and Vigneault et al. [7] designed one network framework based on 2D FCNs to directly segment the left and right atrium. In addition, the proposed networks can also be applied to segment ventricles after training without changing in network framework. Similarly, Xiong et al. [8] , Preetha et al. [9], Bian et al. [10], and Chen et al. [11] also designed a segmentation network framework based on 2D FCNs to segment the atrial. Compared with the previously proposed 2D network frameworks, their network structures were optimized, which made the network pay more attention to feature learning. However, in medical image segmentation, most of the data is 3D volume. Therefore, in order to capture 3D global information, some 3D networks [54–58] and multi-view networks [59, 60] were constantly proposed. Especially the fully automatic two-stage segmentation framework proposed by Xia et al. [54], which mainly includes two steps of localization and segmentation. First, the first 3D UNet was used to roughly locate the center of the target, and then the target region was cropped out, and input the second 3D UNet performed precise segmentation, and the final segmentation result won the first place in the left atrium segmentation challenge 2018.

As attention mechanisms become more popular, it is increasingly used for the cardiac segmentation. For example, Zhou et al. [61] designed a cross-modal attention

module between the encoders and decoder, which leveraged the correlated information between modalities to benefit the cross-modal cardiac segmentation. Based on 3D U-Net et al. [62], Li et al. [58] designed one attention based hierarchical aggregation module, and through the ablation study, the module was proved that it was an effective way to force the network to focus the left atrium of cardiac. Zhang et al. [63] designed three types of attention modules including the spatial attention module that selectively aggregates the features at each position, the channel attention module that focuses on integrating associated features among all channels, and the region attention module that highlights useful feature regions from the whole feature maps, through directly inserting into a FCN, which achieved one good segmentation performance on left and right ventricle of cardiac. Tong et al. [64] presented an interleaved attention mechanism, which can effectively combined low-level and high-level features, and made more discriminative information pass forward to the refinement stage, through applying to recurrent fully convolutional architecture, the performance of cardiac MRI segmentation was improved. Wei et al. [65] proposed a spatial constrained channel attention module to pay more attention to left ventricle of cardiac, decrease the impact of surrounding similar tissues, which can effectively deal with segmentation of multiply connected domains.

Then, in order to segment fibrosis or scars, we mainly base on LGE MR images, because it can show scars and fibrosis [41]. Before deep learning was widely used in the field of medical images, traditional segmentation methods, such as intensity threshold-based or clustering methods, were used for scar segmentation. These methods are very sensitive to the local intensity changes of the image [66], and different parameters need to be designed according to different data each time, and they are not suitable for being widely used. At the moment, they need to manually segment the region of interest to reduce the workload [67]. Therefore, these semi-automatic methods cannot be widely used in hospitals to reduce the workload on doctors.

Therefore, only using traditional segmentation methods to segment scars is not a development trend. Combining with widely used deep learning methods is the current development trend. For example, Yang et al. [12, 13] used one atlas-based method to identify the left atrium, and then used a deep learning network to detect fibrotic tissue in the left atrium area. Related to the end-to-end segmentation method, Chen et al. [14] proposed a deep neural network to segment both the left atrium and atrial scars. In particular, to achieve better segmentation accuracy, they also proposed a multi-view framework with one attention module to integrate different visual information.

Currently, there are still many challenges in fully automatic end-to-end scar segmentation, because the proportion of scars in the entire image is very low, it is easy to cause serious overfitting of the network, and because of the differences of patients, LGE images will also be generated abnormal. To achieve one fast segmentation speed, Fahmy et al. [68] designed one network based on UNet to segment

both the myocardium and the scars, but the segmentation results on the scar regions were very low. Subsequently, Zabihollahy et al. [66] and Moccia et al. [69] proposed one semi-automatic method that kept the higher segmentation accuracy on the test sets for the scar segmentation, first by manually segmenting the myocardium, and then applying a 2D network to distinguish between scars and normal myocardium. Excitingly, an RNN method proposed by Xu et al. [70] could automatically delineate the myocardial infarction area from the MR image sequence without contrast agent. Compared with the manual annotation on the LGE MR images, their method obtained a higher dice score and provided a new method for infarction assessment.

1.4 Public datasets

Among these deep learning methods mentioned above, the most common method is mainly data-driven, and study the transformation relationship between the input image and the corresponding label. So, obtaining the labeled patient data is pivotal for deep learning methods. We make a summary of public datasets on heart segmentation in recent years as shown in Table 1.2.

TABLE 1.2: Summary of public datasets on heart segmentation

Source	Data	Image size/pixels	Voxel size/mm
HVSMR16 [71]	10 3D CMR	390×390×165	0.9×0.9×0.85
MM-WHS2017 [72]	60 CT, 60 bSSFP MRI	324×325×171	0.94×0.94×1.20
AtriaSeg18 [73]	150 LGE MRI	608×608×88	0.625×0.625×0.625
LVQuan19 [74, 75]	56 CMR	347×347×20	1.18×1.18×1.18
LAScarQS2022) [29–31]	298 LGE MRI	608×608×44	1.32×1.32×2.3

HVSMR16 [71] (MICCAI Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease[5]) is to segment myocardium and blood pool, it contains 10 training cardiovascular magnetic resonance (CMR) scans. For each patient, three kinds of images were provided: the full-volume axial images, the cropped axial images around the heart and thoracic aorta, and the cropped short axis reconstruction. The average voxel size is 0.9×0.9×0.85 mm. The average image sizes: 390×390×165 pixels.

MM-WHS2017 [72] (Multi-Modality Whole Heart Segmentation [6]) aims to segment 7 substructures of the whole heart. It contains 60 cardiac MRI and 60 CT images. The average voxel size is 0.94×0.94×1.20 mm. The average sizes: 324×325×171 pixels.

AtriaSeg18 [73] (MICCAI 2018 Atrial Segmentation Challenge [7]) aims to segment the left atrium and contains 150 annotated 3D MRIs from patients with atrial fibrillation. The voxel size of the MR images is 0.625 × 0.625 × 0.625 mm. The dataset includes two different image sizes: 88×576×576 pixel and 88×640×640 pixel.

[5]http://segchd.csail.mit.edu/index.html
[6]http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs17/index.html
[7]http://atriaseg2018.cardiacatlas.org/

LVQuan19 [74, 75] (MICCAI 2019 left ventricle (LV) Full Quantification Challenge dataset[8]) is to segment the myocardium of the left ventricle and estimate a set of clinical significant LV indices such as regional wall thicknesses, cavity dimensions, and cardiac phase and so on. It contains the processed SAX MR sequences of 56 patients. For each patient, 20 temporal frames are given and cover a whole cardiac cycle. All ground truth (GT) values of the LV indices are provided for every single frame. The pixel spacings of the MR images range from 0.6836 mm/pixel to 1.5625 mm/pixel, with mean values of 1.1809 mm/pixel. The LV dataset includes two different image sizes: 256×256 or 512×512 pixels.

LAScarQS2022 [29–31] (Left Atrial and Scar Quantification & Segmentation Challenge [9]) aims to segment the left atrium and evaluates the scar. It includes two tasks (Task 1 and Task 2) and Task 1 contains the scar data. Task 1 contains 60 annotated 3D MRIs from patients with atrial fibrillation for training and validating. The voxel size of the MR images is different: $1.25\times\ 1.25\times\ 2.5$ mm, $1.4\times\ 1.4\times\ 1.4$ mm, and $1.3\times\ 1.3\times\ 4.0$ mm. The dataset includes two different image sizes: $44\times576\times576$ pixels and $44\times640\times640$ pixels.

However, due to the long time taken to form MRI images (10–30 minutes) [32], there are some difficulties in implementing heart segmentation tasks and fibrosis assessment tasks using MRI images from the aforementioned public datasets. As shown in Fig. 1.7, there are 1) poor contrast between myocardium and surrounding structures, 2) brightness due to blood flow, 3) non-homogeneous partial volume due to limited MRI resolution, 4) noise due to motion artifacts and heart dynamics, 5) shape and intensity variability due to different patients and pathologies. So, we should take these problems into account when we design segmentation and evaluation methods.

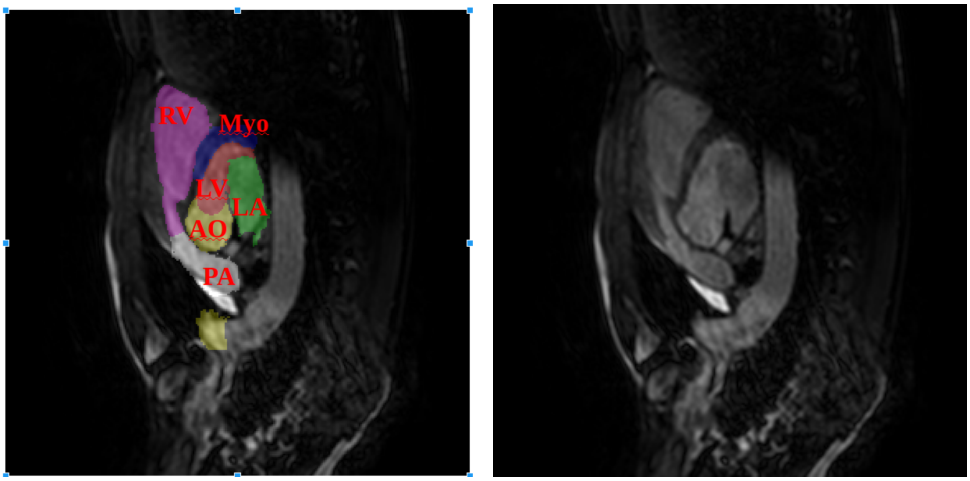


Image with lable Original image

FIGURE 1.7: MRI images. RV: right ventricle blood cavity; Myo: myocardium of the left ventricle; LV: left ventricle blood cavity; LA: left atrium blood cavity; AO: ascending aorta; PA: pulmonary artery.

[8] https://lvquan19.github.io
[9] https://zmiclab.github.io/projects/lascarqs22/

1.5 Main contributions

The main contribution of thesis is to assist doctors in diagnosis by designing neural network frameworks to reduce the workload of doctors. Throughout the design process, we found that there were many difficulties to segment atrial and evaluate fibrosis from cardiac MR images, for example, the presence of poor contrast between the segmented tissue and surrounding structures, the brightness heterogeneities due to blood flow, the shape and intensity variabilities of the structures across patients and pathologies, and so on.

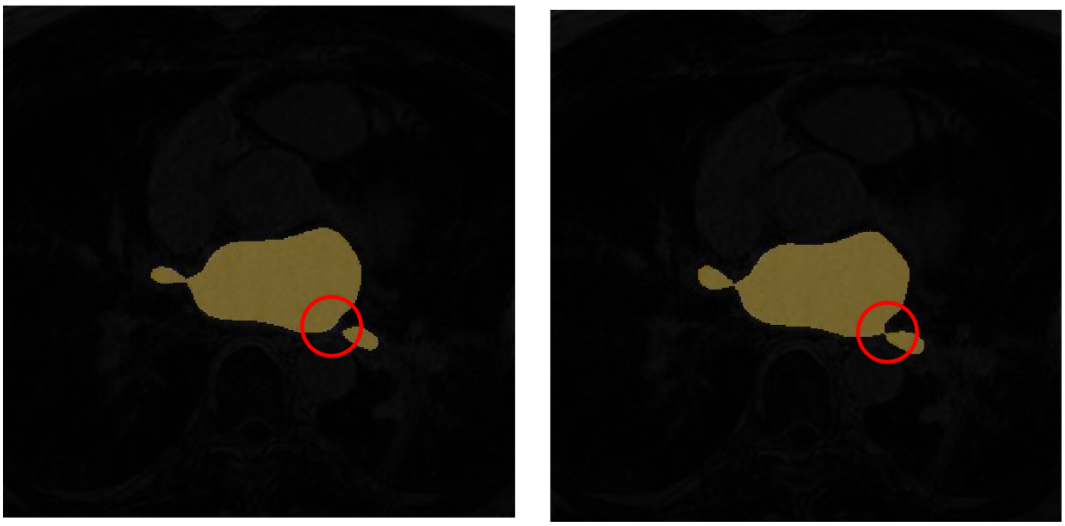


FIGURE 1.8: Detail segmentation at the left atrial

Although the fact that most methods continue to improve segmentation accuracy, the problem of low accuracy of boundaries and small objects segmentation still exists, which is due to the fact that they mainly pay attention to region accuracy, more than to the quality of the boundaries and small objects. As shown in Fig. 1.8, boundaries and details segmentation are especially important when the accuracy of region segmentation is about the same. Therefore, we consider two challenging problems applied on cardiac imaging : 1) how to improve the segmentation accuracy on small parts of objects; 2) how to balance the importance of the regions and the boundaries of objects.

1.5.1 Design of Neural Network Framework

Two-stage framework: For segmenting the left atrial, we propose a two-stage architecture, which is consist of a localization network and a segmentation network. These two networks mainly rely on the original VGG16 network architecture, pre-trained on millions of natural images of ImageNet for image classification. We then discard its fully connected layers to keep only the sub-network made of five

convolution-based "stages" (the base network). Each stage is made of two convolutional layers, a ReLU activation function, and a max-pooling layer. Since the max-pooling layers decrease the resolution of the input image, we obtain a set of fine to coarse feature maps (with 5 levels of features). We added specialized convolutional layers (with a 3×3 kernel size) with K (e.g. K = 16) feature maps after the up-convolutional layers placed at the end of each stage. The outputs of the specialized layers show the same resolution than the input image, and are concatenated together. We add a 1×1 convolutional layer at the output of the concatenation layer to linearly combine the fine to coarse feature maps.

End-to-end framework: For segmenting the left atrial, we propose a end-to-end architecture, using ResNet-101 pre-trained on ImageNet to compute feature maps. We discard its average pooling and fully connected layers, and keep only the sub-network made of one convolution-based and four residual-based "stages". Since the resolution decreases at each stage, we obtain a set of fine to coarse feature maps (with five levels of features). We add specialized convolutional layers (with a 3×3 kernel size) with K (e.g. K =16) feature maps placed at the end of four residual-based "stages". They are concatenated together after up-convolutional layers. These last feature maps are combined with each of the outputs of the specialized layers, and then fed into the attention module to generate the attention features. Finally, we concatenate the attention features and fed them into the softmax layer.

For segmenting the fibrosis/scar, we propose a hybrid network using five U-Net frameworks, which is composed of three U-Net to segment myocardium, left and right ventricle, and whole heart, and the remaining two U-Net to segment edema and scar.

1.5.2 Design of Attention

To decrease the impact of similar tissues on segmentation results, we built on the biological visual system, which concentrates on certain image regions requiring detailed analysis.

Gaussian attention: In the two-stage framework, between the localization network and the segmentation network, we propose one Gaussian attention method, which is to multiply the positioning target area by the Gaussian weight.

Attention module: In the end-to-end framework, we design one attention module embedded in the neural framework, which consists of one position attention branch and one channel attention branch. The attention module can make full use of spatial information and information between channels.

1.5.3 Design of Hybrid Loss

We propose the hybrid loss function that guides the network to study the transformation relationship between the input image and the corresponding label. To let the network to balance boundaries, small objects and regions during the process of training, we not only design region loss, but also boundary loss. For the region loss, we combines Categorical Cross Entropy (CCE), Structural Similarity (SSIM) and Dice Coefficient (DC) to guide the training process at three levels: pixel-level, patch-level, and map-level. For the boundary loss, it is used in calculating the difference between the boundary of the ground truth and the boundary of the prediction.

1.6 Manuscript organization

The thesis is divided into three parts.

The **first part** explains the main concept of work proposed in this thesis. It consists of two chapters.

- **Chapter 1: Introduction.** This chapter mainly describes the research background of cardiac segmentation and evaluation of fibrosis and some related research methods, and explains the significance and contribution of our research in this field.

- **Chapter 2: Theoretical Background.** This chapter is a briefly introduction to the relevant background knowledge required. We explain the fundamentals of deep learning, mainly explaining the convolutional layer and pooling layer, as well as how to train the network and evaluate the prediction results. Finally, we focus on explaining the principle of attention.

The **second part** of this thesis proposes different methods to segment heart and evaluate fibrosis. It consists of three chapters.

- **Chapter 3: Heart Data Preparation.** Deep learning is mainly based on big data, so it is very important to choose a suitable method to preprocess heart data. This chapter mainly explores how different preprocessing methods affect prediction results of network. We compare centralized and standardized, and find that the standardized makes the network more robust to noise through a large number of experiments.

- **Chapter 4: Two-stage Segmentation Method.** For cardiac magnetic resonance images, ambiguities often appear near the boundaries of the target domains due to tissue similarities. This chapter, to address this issue, we propose a new architecture, which can be decomposed in three main steps: a localization step, a Gaussian-based contrast enhancement step, and a segmentation step. This architecture is supplied with a hybrid loss function that guides the network to study the transformation relationship between the input image and the

corresponding label in a three-level hierarchy (pixel-, patch- and map-level), which is helpful to improve segmentation and recovery of the boundaries. We demonstrate the efficiency of our approach on three public datasets in terms of regional and boundary segmentations.

- **Chapter 5: End-to-end Segmentation Method.** This chapter proposes an attention full convolutional network framework based on the ResNet-101 architecture, which focuses on boundaries as much as on regions. The additional attention module is added to have the network pay more attention on regions and then to reduce the impact of the misleading similarity of neighboring tissues. We also use a hybrid loss composed of a region loss and a boundary loss to treat boundaries and regions at the same time. We demonstrate the efficiency of the proposed approach on the MICCAI 2018 Atrial Segmentation Challenge public dataset.

The **third part** of this thesis proposed different methods to evaluate fibrosis. It consists of one chapters.

- **Chapter 6: Evaluation of Fibrosis.** For left atrial fibrosis/scar, this chapter mainly designs two different segmentation methods: (1) Based on the segmentation results of left atrial of Chapter 4 or Chapter 5, combining mathematical morphology approaches, the segmentation results of fibrosis is obtained by setting the fixed threshold. (2) We directly segment the pathology tissue such as the scar by deep learning methods without mathematical morphology approaches. Firstly, we begin with a segmentation of the anatomical tissue (left ventricle (LV), right ventricle (RV), whole heart (WH), myocardium (myo)) around myocardial pathology, and then let the network learn a relationship between these segmentation results to obtain the myocardial pathology. The effect of class imbalance can be reduced by the segmentation of surrounding anatomical tissues, because it helps the network to focus on the small lesions regarding to the surrounding tissues.

The **fourth part** of this thesis makes one conclusion and perspectives. It consists of one chapter.

- **Chapter 7: Conclusion and Perspectives.** This chapter mainly summarizes the dissertation and introduces some research directions that can be explored.

Part II

Concept

Chapter 2

Theoretical Background

2.1 Fundamentals of Deep Learning

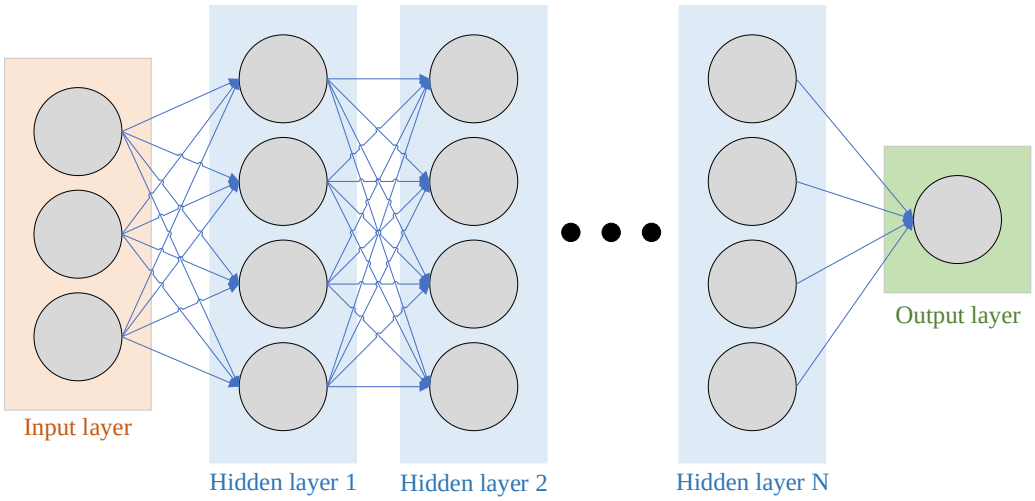


FIGURE 2.1: Neural networks; gray circle denotes neuron

Deep learning models are deep artificial neural networks. Each neural network (as shown in Fig.2.1) consists of an input layer, an output layer, and multiple hidden layers. Convolutional neural network (CNN), which is the most common type of deep neural networks for image analysis. CNN have been successfully applied to advance the state-of-the-art on many image classification, object detection and segmentation tasks [76].

2.1.1 Convolutional Neural Networks (CNNs)

As shown in Fig.2.2, a MR image is input into a CNN, and then hierarchical features are learned by convolutions and pooling layers. Some CNN frameworks are now well known such as LeNet [77], AlexNet [16], VGG [78], Inception [79], and ResNet [80] and so on. These frameworks are mainly used in extracting features at different levels for input images, and then use these features to perform different tasks, for example, these feature maps are flattened and reduced into a vector by fully connected layers, and then the vector can be varied for different tasks. It can be

probabilities for a set of classes (image classification) or coordinates of a bounding box (object localization) or a predicted label for the center pixel of the input (patch-based segmentation) or a real value for regression tasks. Therefore, it is very important to fully grasp the role of each layer of CNN, which often contains convolutional layers, pooling layers and/or fully-connected layers.

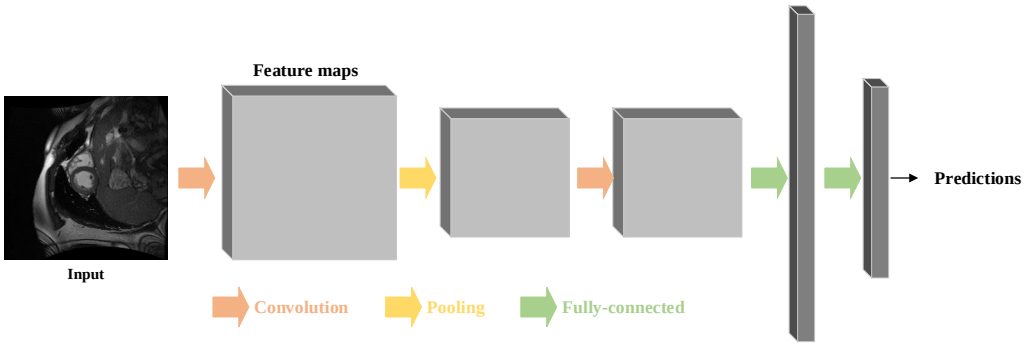


FIGURE 2.2: Convolutional Neural Networks (CNNs)

2.1.1.1 Convolutional Layers

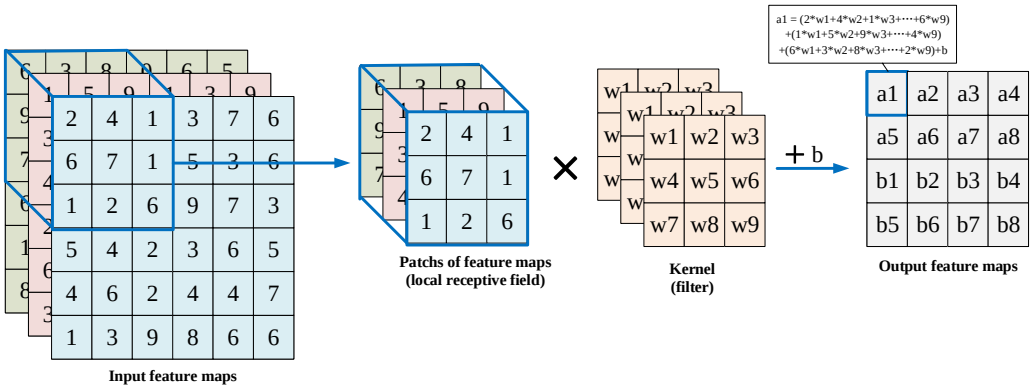


FIGURE 2.3: The calculation process of convolutional layer with the convolution kernel 3×3

The convolutional layer $CONV_l$ is the key part of CNN, and its primary parameters are convolution kernels and convolution filters. Fig. 2.3 shows the calculation process of convolutional layer with the convolution kernel 3×3. For one convolutional layer, if the convolution filter and convolution kernel are set to n_l and $k \times k$, respectively, which means to extract n_l feature maps by the $k \times k$ convolution kernel. In general, the convolution kernel is set to small such as 3×3, which can reduce the number of training parameters of network. However, if using the small convolution kernel, the receptive field (the area of the input image that potentially impacts the activation of a particular convolutional kernel/neuron) is also small. To increase the region of receptive field, the network usually build very deep, which means to

increase the number of convolutional layers. In fact, increasing the depth of convolution neural networks (the number of hidden layers) to enlarge the receptive field can lead to improved model performance. If directly using the big convolution kernel such as 7×7, compared to three convolution layers with 3×3 convolution kernel, the receptive field remains same, but the number of weights is increased by about twice. An online resource[1] is applied to clearly illustrate and visualize the change of receptive field by changing the number of hidden layers and the size of kernels.

2.1.1.2 Pooling Layers

The pooling layer used in CNN framework is mainly to reduce the redundant information and image size, and retain more important features. The most common type of pooling layers used is Max Pooling; the less common Average Pooling is sometimes seen in very deep neural networks.

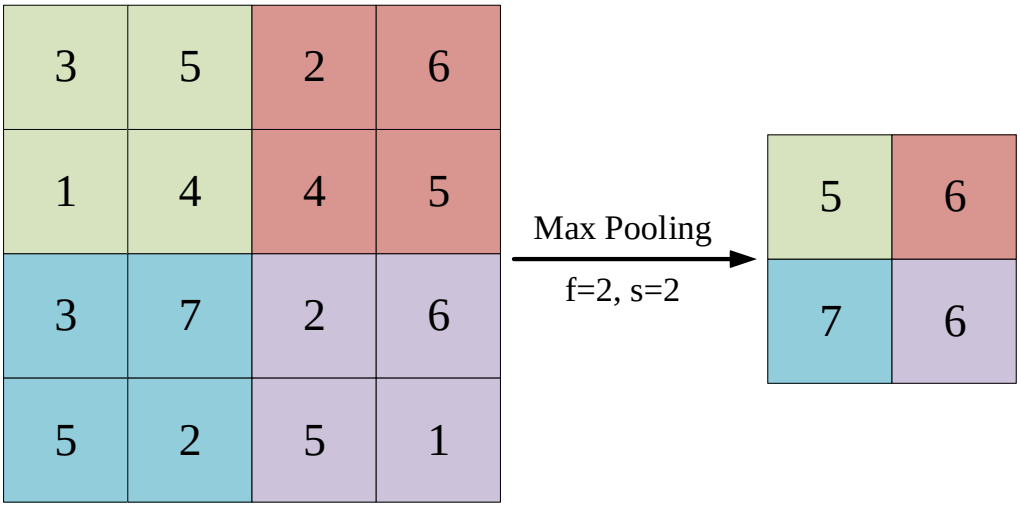


FIGURE 2.4: Max Pooling

The pooling layer does not contain any trainable parameters, and its operation mode is similar to the convolution operator by sliding a small matrix of size $f \times f$ across the input image with stride s, but unlike convolution, pooling is used in each channel individually. As shown in Fig.2.4, an example of Max Pooling, the operation involves taking the max value for $f = 2$ and $s = 2$ applied on a 4×4 matrix of feature map.

2.1.1.3 Activation Function

According to Fig.2.1, we only consider one neuron N:

$$N = \sum (w_i \times input_i) + b \tag{2.1}$$

[1] https://fomoro.com/research/article/receptive-field-calculator

where w_i is weight corresponding to $input_i$, b denotes one bias.

So the neuron is to calculate a weighted sum of its input, and then add a bias (see Fig.2.3), for the activation function, it is defined as how the weighted sum of the input is transformed into an output. As shown in Eq. 2.1, N ranges from negative infinity to positive infinity, which let neuron do not know the bounds of the value. Therefore, if we want neurons to make a purposeful choice of input values, there must be certain restrictions by the activation function.

However, for the neural networks, there are three types of layers as shown in Fig. 2.1. Each of the multiple hidden layers commonly uses the same activation function, but the activation function of the output layer will constantly change according to different tasks.

It is well known that neural networks are trained by backpropagation using error algorithms, requires the activation function to be differentiable. Until now, many activation functions have been proposed and widely used in neural networks, although perhaps only a small part of the activation functions is actually used in the hidden layer or the output layer.

Activation Function for Hidden Layer: There are three activation functions for most commonly using in the hidden layer as follows:

- Rectified Linear Activation (ReLU) [81];
- Logistic (Sigmoid) [82];
- Hyperbolic Tangent (Tanh) [83].

For the **ReLU** [81] activation function, it is defined as:

$$f(x) = max(0, x) \tag{2.2}$$

ReLU [81] activation function (as shown in Fig. 2.5) means from Eq. 2.2 that if the input value (x) is negative, then 0 is returned, otherwise, x is returned. However, it has some potential problems such as non-differentiable at zero, not zero-centered, and unbounded and so on. So, in order to solve its disadvantages, many activation functions were subsequently expanded based on ReLU such as Exponential linear units (ELU) [84], and Leaky ReLU [85],

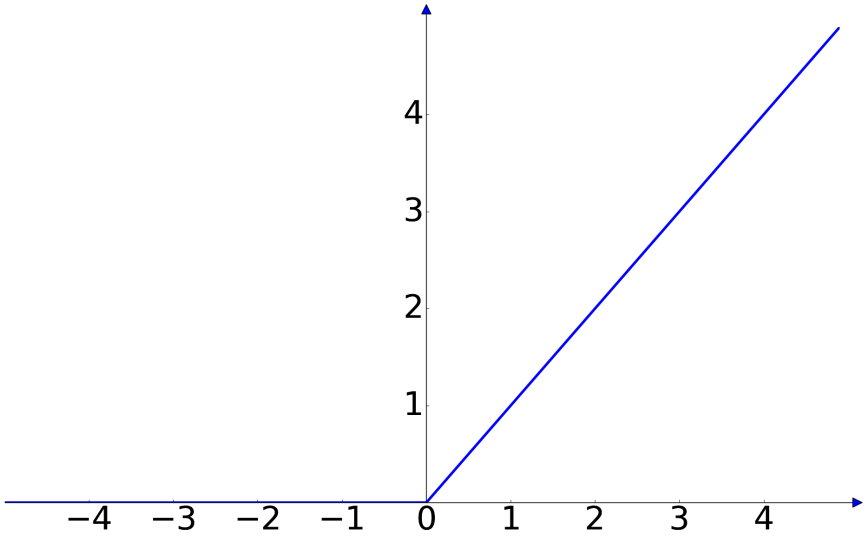


FIGURE 2.5: ReLU activation function

For the **Sigmoid** [82] activation function, it is the logistic function shown in the Fig. 2.6 and defined by the formula:

$$S(x) = \frac{1}{1+e^{-x}} \tag{2.3}$$

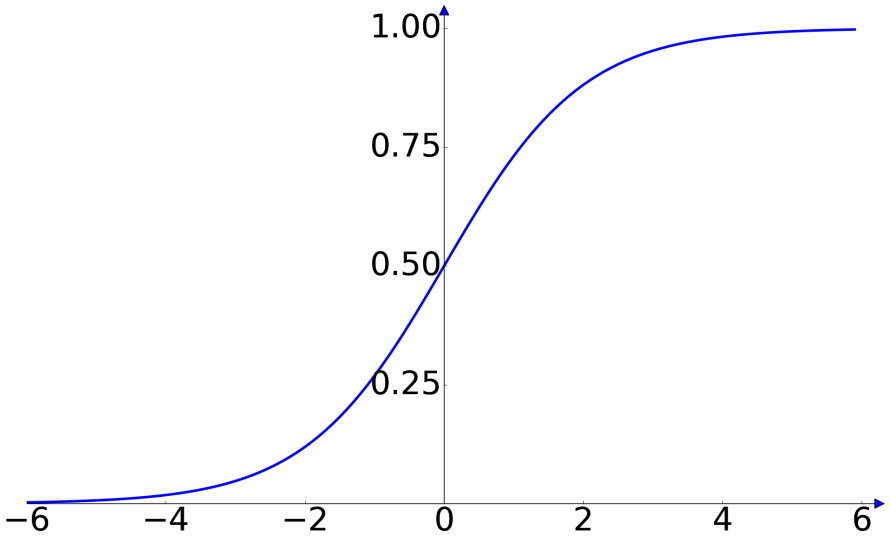


FIGURE 2.6: Sigmoid activation function

The sigmoid activation function is familiar S-shape, and its output value range is from 0 to 1. The larger the input value, the closer the output value is to 1, otherwise, the closer to 0. However, sigmoid has some similar problems with ReLU such as not zero-centered, but it improves the non-differentiable at zero and unbounded

problems. If we want to use the sigmoid or ReLU in the hidden layer, the input data should preferably be scaled to the range of 0 to 1.

For the **Tanh** [83] activation function (as shown in Fig. 2.7), it is a scaled sigmoid function and defined as:

$$T(x) = 2 * S(2x) - 1 \tag{2.4}$$

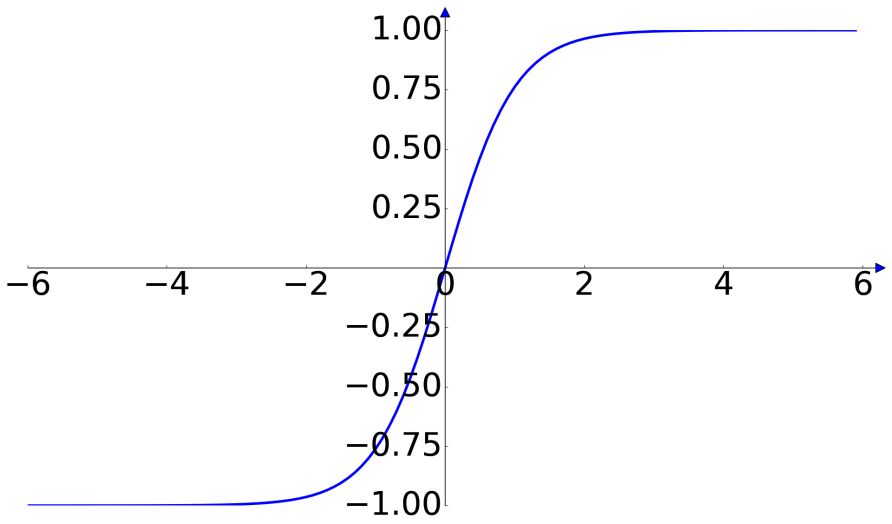


FIGURE 2.7: Tanh activation function

The Tanh activation function combines the advantages of ReLU and sigmoid, and is zero-centered. Its output value range is from -1 to 1. But but we need to pay attention to if we want to use in the hidden layer, the input data should be scaled to the range of -1 to 1, which is different with ReLU and sigmoid.

Activation Function for Output Layer: There are also three activation functions for most commonly using in the output layer as follows:

- Linear [86];
- Logistic (Sigmoid);
- Softmax [87].

For the linear [86] activation function, it directly returns the weighted sum of the input and not change the value. The linear activation function in output layer is mainly used for the regression task. Sigmoid is used for the binary classification task.

For the softmax [87] activation function, it outputs a vector of values that sum to 1 that can be interpreted as probabilities of class membership and is defined as:

$$\sigma(\boldsymbol{z}) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_i}} \quad for \;\; i = 1, \cdots, k \;\; and \;\; \boldsymbol{z} = (z_1, \cdots, z_k)\, \epsilon \mathbb{R}^k \tag{2.5}$$

It uses the standard exponential function to each element z_i of the input vector $\mathbf{z}$ and normalizes these values by dividing by the sum of all these exponentials, which ensures that the sum of the components of the output vector $\sigma(\mathbf{z})$ is 1. The softmax activation can be used not only for binary classification, but also for multi-classification tasks.

2.1.2 Training Neural Networks

For the neural network model to be successfully used, it must be trained for a long time based on big data. Therefore, there are certain requirements for the provided dataset. The dataset must contains paired images and labels for training and validating. Model parameters are updated through a loss function and an optimizer such as adam [88], RMSprop [89] and stochastic gradient descent [90] and so on. If you want to learn more about optimizers, please refer to the literature [91]. During the process of training, the loss function continuously calculates the error between the prediction and the label in each iteration, then minimizes the error value by providing signals for the optimizer to update the network parameters through backpropagation [92].

2.1.2.1 Loss Functions

Categorical Cross Entropy (CCE) [93] loss is commonly used for multi-class classification and segmentation. It is defined as:

$$\ell_{\mathrm{CCE}} \;=\; -\sum_{i=1}^{C}\sum_{a=1}^{H}\sum_{b=1}^{W} y^{i}_{(a,b)} \ln {y_*}^{i}_{(a,b)}, \tag{2.6}$$

where C is the number of classes of each image, H and W are the height and width of image, $y^{i}_{(a,b)} \in \{0, 1\}$ is the ground truth one-hot label of class i in the position (a, b) and ${y_*}^{i}_{(a,b)}$ is the predicted probability of class i.

Dice Coefficient (DC) [94] loss is used to measure the similarity between two sets as defined in Eq. 2.7.

$$\ell_{\mathrm{DC}} \;=\; 1 - \frac{2\,|A \bigcap B|}{|A| + |B|} \tag{2.7}$$

where A and B denote the 2D image matrix of prediction and target, respectively.

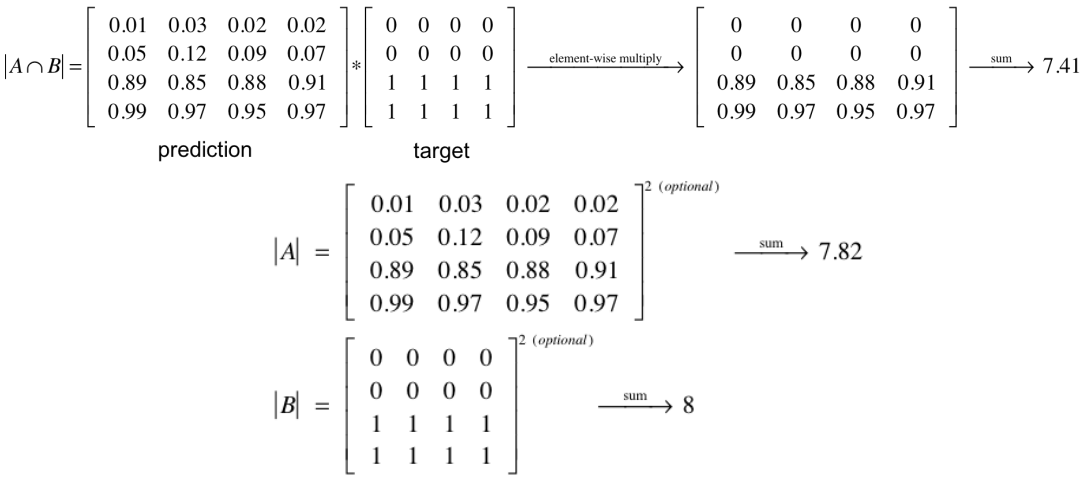


FIGURE 2.8: An example of dice coefficient calculation process

The loss function mentioned above is the simplest form, and some new loss functions are extended based on them, for example, to solve class imbalance problem, the weighted cross-entropy loss [95] and weighted dice loss [96] are presented, which is weighted to calculate rare classes or small objects.

2.1.2.2 Optimizers

During the training process of the network, optimizers are used for changing the parameters (weights) of the network to minimize the loss function. To successfully train the network, choosing the right optimizer is crucial. Therefore, we need to fully understand the pros and cons of various optimizers. Nowadays, the main optimizers are Stochastic Gradient Descent (SGD) [97], Adaptive gradient algorithm (Adagrad) [98], Root Mean Square Prop (RMSprop) [89], and Adaptive Moment Estimation (Adam) [88], etc. Next we will explain them one by one.

Stochastic Gradient Descent (SGD) [97]: It is defined as:

$$\theta = \theta - \eta \nabla_\theta J\left(\theta; x^i, y^i\right) \tag{2.8}$$

where θ is a network's parameters $\theta \epsilon \mathbb{R}$. $J(*)$ denotes an objective function (loss function). $\nabla_\theta J(*)$ denotes the opposite direction of the gradient of the objective function. η is the learning rate, which denotes the step size of updating gradient. x^i and y^i denotes the input and label of each training example, respectively.

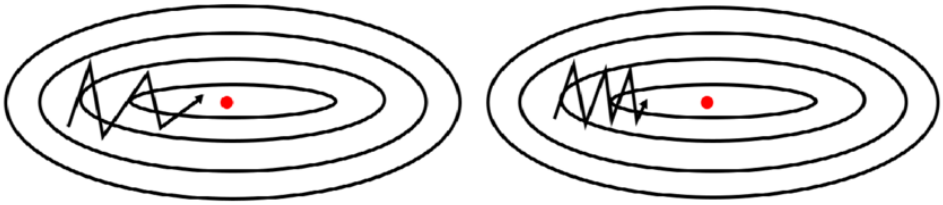
The network's parameters are updated based on each training example by SGD optimizer, which does not perform redundant computations for large datasets, and new training example can be added. However, every iteration is not toward the direction of global optimization, because SGD does not update the network's parameters based on the entire sample. Although the training speed is fast, the accuracy is reduced, which is not the global optimum. SGD updates the parameters frequently, which will cause serious fluctuations for the objective function.

SGD is easily trapped in the case of ravines. Ravines means that one direction of the surface is steeper than the other. At this time, SGD will oscillate and it will not be close to the minimum value. To solve this problem, momentum is added into Eq. 2.8 that is redefined as:

$$v_t = \gamma v_{t-1} + \eta \nabla_\theta J\left(\theta; x^i, y^i\right) \tag{2.9}$$

$$\theta = \theta - v_t \tag{2.10}$$

where γ is a fraction and usually set to 0.9. Essentially, we push a small ball down a mountain. There are no obstacles in the whole process of rolling down. The speed of the ball is getting faster and faster, so its momentum is also increasing. The same phenomenon also appears in updating the network parameters:the increase of momentum must be the same as the direction of the gradient, otherwise, it will decrease. Finally, we gain convergence quickly and reduce the oscillation as shown in Fig. 2.9.



(a) with momentum (b) without momentum

FIGURE 2.9: SGD with or without momentum

Adaptive gradient algorithm (Adagrad) [98]: It can make larger updates to parameters related to uncommon features, and smaller updates to parameters related to frequently occurring features, which let the Adagrad optimizer adapt the learning rate to the parameters. The update rules are as follows:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i} \tag{2.11}$$

$$G_{t,ii} = \sum_{\tau=1}^{t} (g_{\tau,i})^2 \tag{2.12}$$

where $g_{t,i}$ denotes the gradient of θ_i at time step t:

$$g_{t,i} = \nabla_\theta J\left(\theta_i\right) \tag{2.13}$$

Therefore, the advantage of Adagrad is that it eliminates the need to manually tune the learning rate and the learning rate is usually set to 0.01, but its disadvantage is also very obvious. Since the denominator in Eq. 2.11 is accumulating the square gradient during training, which causes the learning rate to shrink and eventually

become infinitely small. At this time, the optimizer can no longer acquire additional knowledge.

Root Mean Square Prop (RMSprop) [89]: It is to solve the problem of Adagrad's radically diminishing learning rates and is defined as

$$E\left[g^2\right]_t = \gamma E\left[g^2\right]_{t-1} + (1-\gamma)\, g_t^2 \tag{2.14}$$

where $E(\cdot)$ denotes the expectation, therefore, $E\left[g^2\right]_t$ is calculated by the previous average and the current gradient according to Eq. 2.14, and γ is usually set to 0.9. And then, the update rules are as follows:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E\left[g^2\right]_t + \epsilon}} \cdot g_t \tag{2.15}$$

Adaptive Moment Estimation (Adam) [88]: It is equivalent to RMSprop plus momentum. The decaying averages of past and past squared gradients m_t and v_t is calculated, respectively, as follows:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1)\, g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)\, g_t \end{aligned} \tag{2.16}$$

where m_t denotes the estimate of the first moment at time step t. v_t denotes the estimate of the second moment at time step t. If m_t and v_t are initialized as vectors of 0's, they will be biased towards 0, so the bias is corrected as follows:

$$\begin{aligned} \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \end{aligned} \tag{2.17}$$

Therefore, the Adam optimizer's update rules are as follows:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t \tag{2.18}$$

where β_1=0.9, β_2= 0.999, ϵ=10e-8.

In summary, choosing the right optimizer is very important for training the network. If the training data is sparse data, we should choose the self-applicable optimizer such as Adagrad, RMSprop, Adam, but, in general, Adam is the best choice.

2.1.2.3 Metrics

Before training the network, we need to set up some observation metrics to know whether the network is moving in our predetermined. We usually use classification accuracy and logarithmic loss as observation metrics during the training process.

For the classification accuracy, it is the ratio of number of correct predictions to the total number of input samples and is defined as follows:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made} \tag{2.19}$$

However, it does not work well for tasks with the class imbalance, for example, a binary classification task, one class A accounts for 96% of the entire sample, and the other class B only accounts for 4%. Then the classification accuracy can easily reach 96% at the beginning of training.

For the logarithmic loss, if there are N training samples corresponding to M classes, it can be calculated as follows:

$$loss = \frac{-1}{N} \sum_{i=1}^{N} \sum_{J=1}^{M} y_{ij} * log\left(p_{ij}\right) \tag{2.20}$$

where y_{ij} denotes whether the sample i belongs to the class j or not. p_{ij} denotes the probability that the sample i belongs to the class j. It can work well for classification tasks. Generally, the classifier can be provided with higher accuracy by minimizing the logarithmic loss. There are many other observation metrics such as Confusion Matrix [99], Area under Curve [100], Mean Absolute Error [101] and Mean Squared Error [102] and so on.

2.1.2.4 Backpropagation

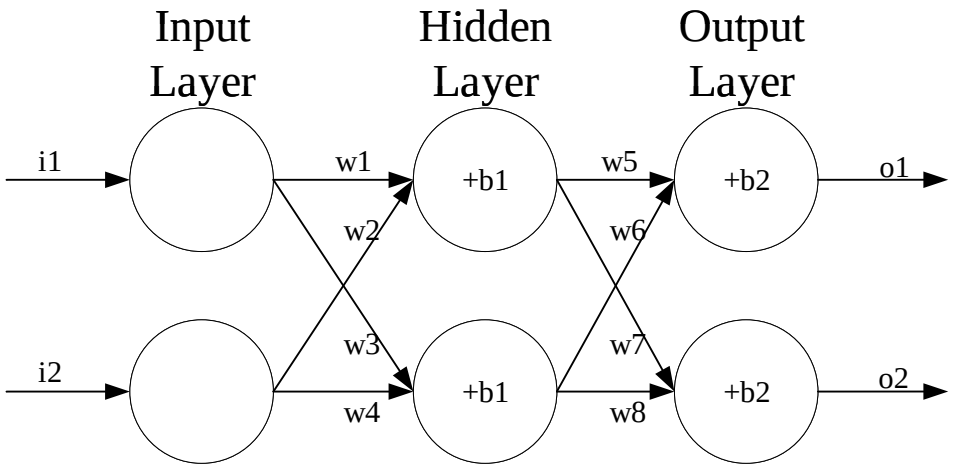


FIGURE 2.10: A simple example of backpropagation

In order to facilitate the understanding of the principle of backpropagation [103], we have created a simple structure as shown in Fig. 2.10 to explain the calculation process of backpropagation. As shown in Fig. 2.10, there are two inputs, two outputs

and one hidden layer. We assume that:

$$\begin{aligned} i1 &= 0.04 \\ i2 &= 0.2 \\ w1 &= 0.12 \\ w2 &= 0.24 \\ w3 &= 0.25 \\ w4 &= 0.3 \\ w5 &= 0.5 \\ w6 &= 0.35 \\ w7 &= 0.45 \\ w8 &= 0.55 \\ b1 &= 0.36 \\ b2 &= 0.8 \\ o1 &= 0.02 \\ o2 &= 0.98 \end{aligned} \tag{2.21}$$

The purpose of backpropagation is mainly to optimize the weights so that the network learn how to correctly transform inputs to outputs. We want the final output $o1$ and $o2$ to reach 0.02 and 0.98, respectively. Therefore, we first need to calculate the predicted value of the forward path of the network.

There are two neurons in the hidden layer, called $h1$ and $h2$. We use the logistic function as the activation function. So out_{h1} is calculated by

$$\begin{aligned} out_{h1} &= sigmoid(i1 \times w1 + i2 \times w2 + b1) \\ &= sigmoid(0.04 \times 0.12 + 0.2 \times 0.24 + 0.36) \\ &= \frac{1}{1 + e^{-0.4128}} \\ &= 0.6017590759 \end{aligned} \tag{2.22}$$

out_{h2} is calculated by the same process and $out_{h2} = 0.60587366843$. We continue to calculate the final outputs out_{pre1} and out_{pre2}: $out_{pre1} = 0.78799802619$; $out_{pre2} = 0.80281786099$. Then calculating the total error between the final outputs and prediction outputs by squared error function:

$$\begin{aligned} E_{total} &= \sum_{i}^{2} \frac{1}{2}(oi - out_{prei})^2 \\ &= E_{o1} + E_{o2} \\ &= \frac{1}{2}(0.02 - 0.78799802619)^2 + \frac{1}{2}(0.98 - 0.80281786099)^2 \\ &= 0.310607239 \end{aligned} \tag{2.23}$$

We have completed the calculation of the forward path of the network. Then we need to calculate the backward path of the network. During the process of calculating the backward path, its purpose is to minimize the total error by minimizing the error of each neuron, for example, we consider how $w5$ affects the total error E_{total}, and it is denotes $\frac{\partial E_{total}}{\partial w5}$. $\frac{\partial E_{total}}{\partial w5}$ denotes the partial derivative of E_{total} with respect to $w5$. According to the chain rule[2]:

$$\frac{\partial E_{total}}{\partial w5} = \frac{\partial E_{total}}{\partial out_{pre1}} \times \frac{\partial out_{pre1}}{\partial w5} \tag{2.24}$$

Due to $E_{total} = \frac{1}{2}(o1 - out_{pre1})^2 + \frac{1}{2}(o2 - out_{pre2})^2$, so

$$\begin{aligned} \frac{\partial E_{total}}{\partial out_{pre1}} &= 2 \times \frac{1}{2}(o1 - out_{pre1})^{2-1} \times 1 + 0 \\ &= -0.767998026 \end{aligned} \tag{2.25}$$

Due to $out_{pre1} = sigmoid(out_{h1} \times w5 + out_{h2} \times w6 + b2)$, so

$$\begin{aligned} \frac{\partial out_{pre1}}{\partial w5} &= out_{pre1} \times (1 - out_{pre1}) \times 1 \times out_{h1} \times w5^{1-1} + 0 + 0 \\ &= 0.78799802619 \times (1 - 0.78799802619) \times 1 \times 0.6017590759 \times 1 \\ &= 0.100528148 \end{aligned} \tag{2.26}$$

Therefore,

$$\begin{aligned} \frac{\partial E_{total}}{\partial w5} &= \frac{\partial E_{total}}{\partial out_{pre1}} \times \frac{\partial out_{pre1}}{\partial w5} \\ &= -0.767998026 \times 0.100528148 \\ &= -0.077205419 \end{aligned} \tag{2.27}$$

We can update $w5$ by

$$w5_{update} = w5 - \eta \times \frac{\partial E_{total}}{\partial w5} \tag{2.28}$$

where η denotes the learning rate, if η is equal to 0.01:

$$\begin{aligned} w5_{update} &= 0.5 - 0.01 \times (-0.077205419) \\ &= 0.500772054 \end{aligned} \tag{2.29}$$

$w6$, $w7$, and $w8$ can be updated by the same process. Next, we will continue to calculate the $w1$, $w2$, $w3$ and $w4$, for example, we consider how $w1$ affects the total error E_{total} by

$$\begin{aligned} \frac{\partial E_{total}}{\partial w1} &= \frac{\partial E_{total}}{\partial out_{h1}} \times \frac{\partial out_{h1}}{\partial w1} \\ &= (\frac{\partial E_{o1}}{\partial out_{h1}} + \frac{\partial E_{o2}}{\partial out_{h1}}) \times \frac{\partial out_{h1}}{\partial w1} \end{aligned} \tag{2.30}$$

[2]https://en.wikipedia.org/wiki/Chain_rule

According to the Eq. 2.30, $\frac{\partial E_{total}}{\partial out_{h1}}$ needs to take into consideration its effect on the both output out_{pre1} and out_{pre2}. Based on Eq. 2.26, $\frac{\partial E_{o1}}{\partial out_{h1}}$ can be simply calculated by

$$\begin{aligned}\frac{\partial E_{o1}}{\partial out_{h1}} &= -0.767998026 \times 0.78799802619 \times (1 - 0.78799802619) \times 0.5 \\ &= -0.064149776\end{aligned} \tag{2.31}$$

Following the same process for $\frac{\partial E_{o2}}{\partial out_{h1}}$, we can get $\frac{\partial E_{o2}}{\partial out_{h1}} = 0.043590891$. Therefore,

$$\begin{aligned}\frac{\partial E_{total}}{\partial out_{h1}} &= 0.043590891 - 0.064149776 \\ &= -0.020558885\end{aligned} \tag{2.32}$$

Due to $out_{h1} = sigmoid(i1 \times w1 + i2 \times w2 + b1)$, so

$$\begin{aligned}\frac{\partial out_{h1}}{\partial w1} &= out_{h1} \times (1 - out_{h1}) \times i1 \\ &= 0.6017590759 \times (1 - 0.6017590759) \times 0.04 \\ &= 0.009585804\end{aligned} \tag{2.33}$$

Finally,

$$\begin{aligned}\frac{\partial E_{total}}{\partial w1} &= \frac{\partial E_{total}}{\partial out_{h1}} \times \frac{\partial out_{h1}}{\partial w1} \\ &= -0.020558885 \times 0.009585804 \\ &= -0.000197073 \\ w1_{update} &= w1 - \eta \times \frac{\partial E_{total}}{\partial w1} \\ &= 0.12 - 0.01 \times (-0.000197073) \\ &= 0.120001971\end{aligned} \tag{2.34}$$

$w2$, $w3$, and $w4$ can be updated by the same process. The above calculation process is the update principle of backpropagation for all weights of the network.

2.1.2.5 Over-fitting

In the field of medical image analysis, due to the small dataset, such as just a few patient data, it often leads to over-fitting problems in training neural networks. In order to alleviate this problem, many new methods have been proposed as follows:

(1) **Data augmentation** [104] is a method to artificially create new training data from existing training data by using affine transformations such as rotation, scaling, and flipping and so on;

(2) **Regularization** is a method which makes slight modifications to the learning algorithm such that the model generalizes better. L1 and L2 [105] regularization are commonly used, which penalize the sum of the absolute weights and the sum of the squared weights, respectively;

(3) **Dropout** [106] is also a regularization technique that randomly drops some units (both hidden and visible) in the neural network during the process of training, which prevents complex co-adaptations on training data;

(4) **Transfer learning** [107] is a method where knowledge is transferred from one model to another, which is achieved by loading the weights of a pre-trained model into the current model, and keeps the framework of the pre-trained models unchanged in the different tasks.

2.1.3 Evaluation Metrics

In order to evaluate the performance of medical image segmentation methods, many evaluation metrics are proposed, which are mainly divided into three types: (a) **volume-based** metrics such as Dice metric [94] and Jaccard similarity index [108]; (b) **surface distance-based** metrics such as Hausdorff distance [109]; (c) **clinical performance** metrics such as ventricular volume and mass. In this dissertation, we mainly report the accuracy of methods in terms of the Dice metric [94] and Hausdorff distance [109], which already includes the evaluation of regions and boundaries and can fully evaluate methods.

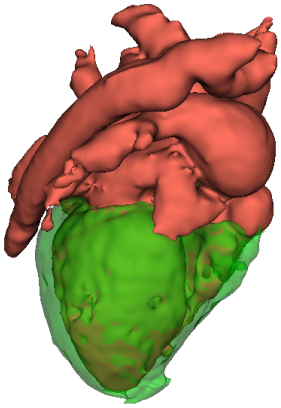


FIGURE 2.11: An example of 3D image matrix of prediction

Firstly, the Dice metric [94] is defined in Eq. 2.35. In the evaluation process, the evaluation metrics is based on one patient (3D volume) rather than one image, therefore, A_{3D} and B_{3D} of Eq. 2.35 are the 3D image matrix of prediction (as shown in Fig. 2.11) and target, respectively.

$$Dice = \frac{2\,|A_{3D} \bigcap B_{3D}|}{|A_{3D}| + |B_{3D}|} \tag{2.35}$$

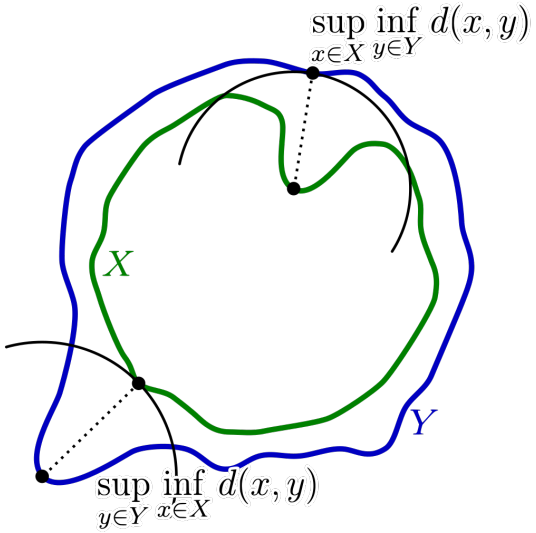


FIGURE 2.12: Hausdorff distance[3]

For the Hausdorff distance [109] as defined in Eq. 2.36, it is the longest distance you can be forced to travel by an adversary who chooses a point in one of the two sets, from where you then must travel to the other set. In other words, it is the greatest of all the distances from a point in one set to the closest point in the other set. It is also based on 3D space, which is the same as dice metric.

$$d_H(X, Y) = max\left\{ \sup_{x \epsilon X} \left(\inf_{y \epsilon Y} d(x, Y) \right), \sup_{y \epsilon Y} \left(\inf_{x \epsilon X} d(X, y) \right) \right\} \tag{2.36}$$

where X and Y denote two non-empty subsets of a metric space, sup represents the supremum.

2.2 Attention Method

In the field of medical imaging, due to the existence of a large amount of redundant information, the over-fitting of the network is aggravated. An important property of the human visual system is to not process a whole scene at once. Instead, humans exploit a sequence of partial glimpses, and selectively focus on salient parts in order to capture the visual structure in a better way [110, 111]. For this reason, attention methods have been developed: they focus on important regions, filter irrelevant information, and make up the limited receptive field of CNNs. They get good performance on segmentation tasks [112–115]. The block diagram of the attention module is shown in Fig. 2.13. The attention module teaches the network to pay attention to important features (e.g., features relevant to anatomy) and ignore redundant features.

[3]https://en.wikipedia.org/wiki/Hausdorff_distance

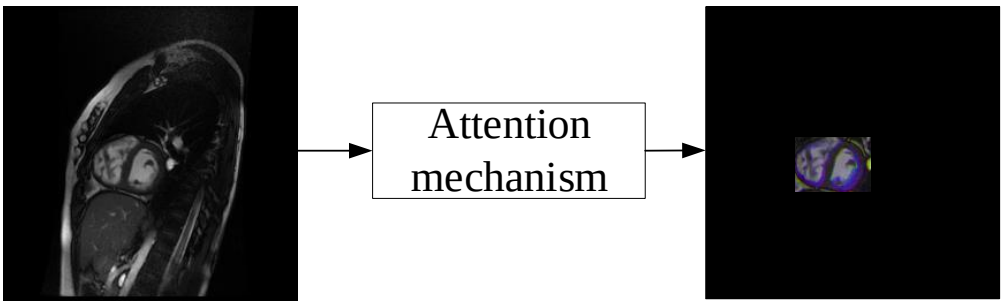


FIGURE 2.13: Block diagram of the attention module

For the attention unit in Fig. 2.13, it has been widely used for medical image segmentation. The common attention units mainly include Channel Attention Block (CAB), Region Attention Block (RAB), Position Attention Module (PAM), Channel Attention Module (CAM) [116]. CAM and PAM evolved based on CAB and RAB.

Channel Attention Block (CAB): Its purpose is to select the more important channel among all input channels, which means that each channel will be given a corresponding weight. The entire realization process is shown in Fig. 2.14.

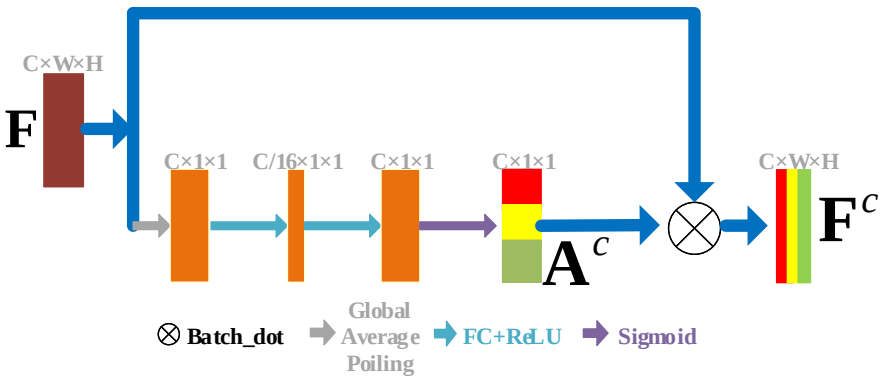


FIGURE 2.14: Channel Attention Block (CAB)
. C, W and H denote channel, width and height, respectively.

First, CAB will use the global average pooling layer to compress the spatial information, which transforms the shape of the input feature map from $C \times W \times H$ to $C \times 1 \times 1$. We assume that $\mathbf{F} = [M_1, M_2, M_3, ..., M_C]$ $(M_i \epsilon R^{H \times W}, i\epsilon[1, C])$ as the input feature maps. After the global average pooling layer, the output g is acquired by

$$g_i = \frac{1}{H \times W} \sum_{j=1}^{H} \sum_{k=1}^{W} M_i(j,k) \tag{2.37}$$

where (j,k) denotes the spatial position.

Then g goes through two fully connected layers, which squeezes and expands the dimensions of feature maps. Finally, CAB uses one sigmoid layer to score each channel and get attention maps $\mathbf{A}^C$. The final output $\mathbf{F}^C$ after passing CAB can be calculated by

$$\mathbf{F}^C = \mathbf{F} \times \mathbf{A}^C \tag{2.38}$$

Region Attention Block (RAB): It can reduce redundant information and make the network concentrated in the target region. The entire realization process is shown in Fig. 2.15.

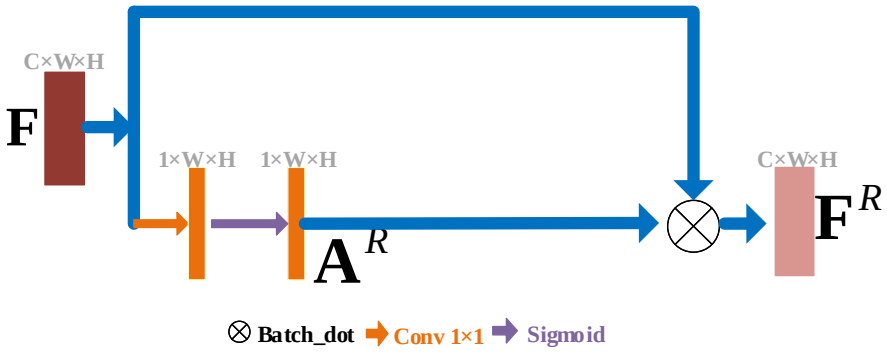


FIGURE 2.15: Region Attention Block (RAB)

RAB focuses on spatial information, first, the number of channel is compressed from C to 1 by one convolutional layer, then using the sigmoid layer to score each pixel of feature maps and get attention maps $\mathbf{A}^R$. The final output $\mathbf{F}^R$ after passing RAB can be calculated by

$$\mathbf{F}^R = \mathbf{F} \times \mathbf{A}^R \tag{2.39}$$

Position Attention Module (PAM): It is used in obtaining the similarity of pixels at different locations. Therefore, the corresponding weight of each pixel depends on the degree of similarity. The entire realization process is shown in Fig. 2.16.

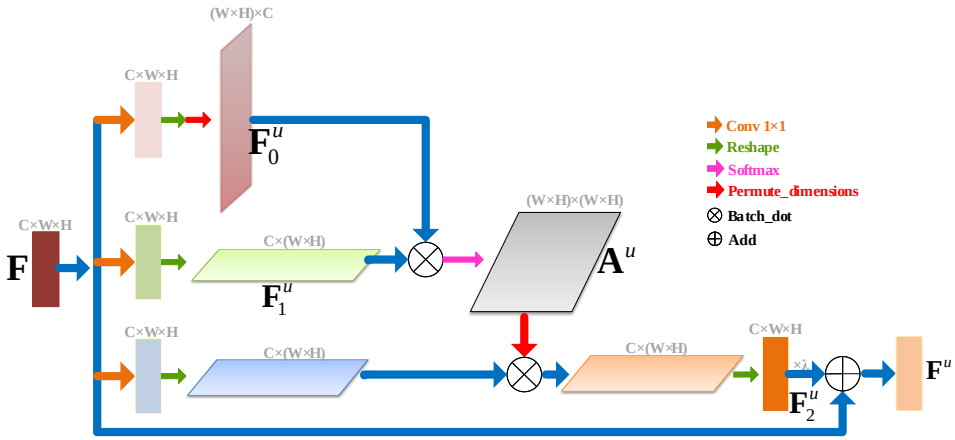


FIGURE 2.16: Position Attention Module (PAM)

F is fed into a convolutional, a Reshape and then a Transpose layers, resulting in a feature map $\mathbf{F}_0^u \in \mathbb{R}^{(W\times H)\times C}$. In the second branch (consider the order from top to bottom), the input feature map **F** follows the same operations minus the Transpose layer, resulting in $\mathbf{F}_1^u \in \mathbb{R}^{C\times(W\times H)}$. Then, the Multiply and the Softmax layers follow; they are applied on $\mathbf{F}_0^u$ and $\mathbf{F}_1^u$ to obtain the spatial attention map $\mathbf{A}^u \in \mathbb{R}^{(W\times H)\times(W\times H)}$. The input **F** is fed into a different convolutional layer in the third branch, and is then multiplied by $\mathbf{A}^u$ fed into the Transpose layer, resulting in $\mathbf{F}_2^u$. Therefore the output $\mathbf{F}^u$ can be formulated as follows:

$$\mathbf{F}^u = \lambda \times \mathbf{F}_2^u + \mathbf{F}, \tag{2.40}$$

where $\lambda \in \mathbb{R}^C$ is initialized to [0,..,0]. The values λ is used to gradually learn the importance of the spatial attention map.

Channel Attention Module (CAM): Its purpose is mainly to discover the relationship between the different channels. The entire realization process is shown in Fig. 2.17.

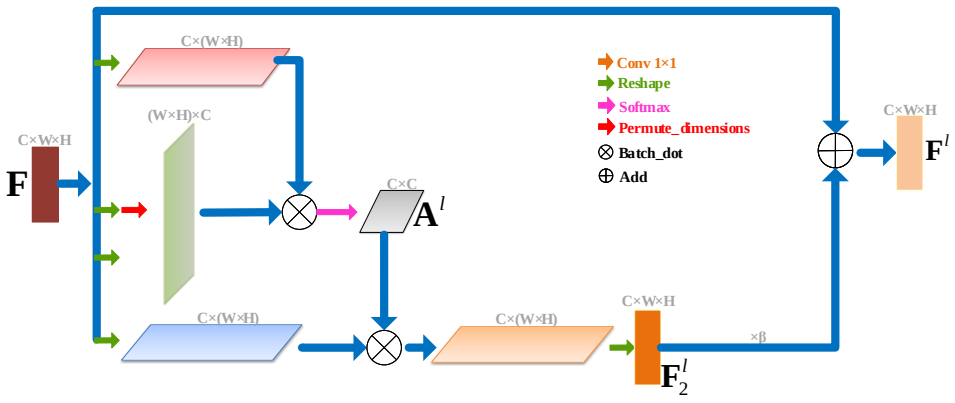


FIGURE 2.17: Channel Attention Module (CAM)

The channel attention map $\mathbf{A}^l$ can be obtained by different combinations of convolutional, Reshape and Transpose layers. Finally, the output $\mathbf{F}^l$ can be defined as follows:

$$\mathbf{F}^l = \beta \times \mathbf{F}_2^l + \mathbf{F}, \tag{2.41}$$

where $\beta \in \mathbb{R}^C$ is initialized to [0,..,0]. The feature map $\mathbf{F}_2^l$ denotes the results of the product of the input $\mathbf{F}$ with $\mathbf{A}^l$ fed into a convolutional passing through the transpose block.

The above attention units belong to soft attention. Soft attention is parameterization, so it is differentiable. Therefore, we can embed it into the network framework and train it together with other layers of the network. The gradient can be back propagated to other parts of the network through the attention unit.

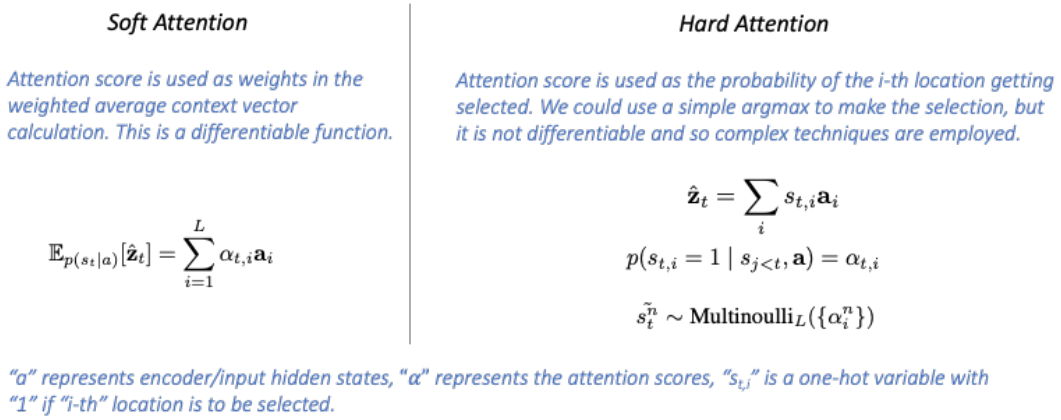


FIGURE 2.18: Soft vs Hard Attention [34]

Compared with soft attention, there must be hard soft. However, there is relatively little research on hard attention by researchers. For the hard attention, we need to select feature maps of the input by using attention scores, which means one

problem, because we can choose one function such as argmax to finish the selection, however, as we all know, it is not differentiable. Therefore, we can not embed it into the network for directly training, and need to more complex methods to solve this problem. Fig. 2.18 details the different between soft attention and hard attention. Soft attention can process everything but weights various regions differently. Hard attention can select only a fraction of the data for processing [117]. Hard-attention models address various use-cases, and can be motivated by interpretability [118], reduction of high-resolution data acquisition cost [119], or computational efficiency [120].

2.3 Conclusion

In this chapter, we have explained the fundamentals of deep learning and attention method. These basic knowledge will facilitate the understanding of the content of the subsequent chapters.

Part III

Heart Segmentation Methods

Chapter 3

Heart Data Preparation

There are many ways to preprocess medical image data before feeding into the network such as data augmentation, crop, resample, centralization (subtracting mean) and standardization (subtracting mean and then dividing standard deviation) and so on. Although we have used data augmentation cropping and resampling during the process of preprocessing, in this part we mainly explore the impact of centralization and standardization on network output.

3.1 Data Preprocessing Exploration

We design a series of experiments, which are mainly based on two network frameworks and two public datasets. These two frameworks are UNet as shown in Fig. 1.6 and an improved FCN framework [19] as shown in Fig. 3.1. These two public datasets are MRBrainS2018[1] and 2018 atrial segmentation challenge[2].

3.1.1 Architecture of Network

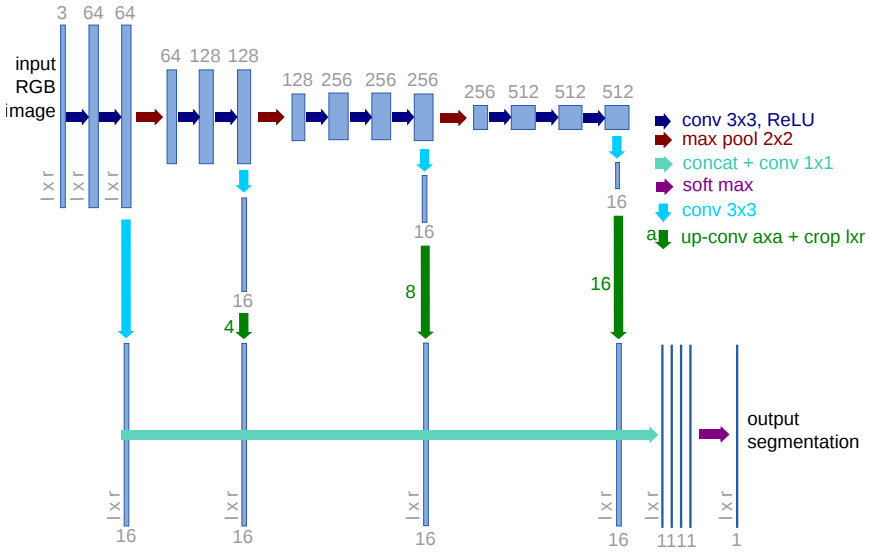


FIGURE 3.1: Architecture of the proposed network

[1] https://mrbrains18.isi.uu.nl
[2] https://atriaseg2018.cardiacatlas.org/

UNet [53]: The framework of network is shown in Fig. 1.6. The yellow part denotes the contracting path and the green part denotes the expansive path. For the contracting path, it mainly includes convolutional layers and max-pooling layers. C1, C3, C5, C7 are the repeated application of two 3×3 convolution operation with a rectified linear unit (ReLU), and its number of filters is [32 64 128 256 512 256 128 64 32]. C2, C4, C6, C8 are the max pooling operation with stride 2, after passing the max-pooling layer, the number of channels is doubled. For the expansive path, it mainly includes upsampling layers, convolutional layers and concatenate layer. After passing the upsampling layers and convolutional layers, the shape of feature maps doubles as before, but the number of channels is halved. Then concatenating the corresponding to feature maps from the contracting path. Finally, the output of network is obtained by the 1×1 convolutional layer.

Improved FCN framework [19]: The network architecture is illustrated in Fig. 3.1. The architecture is based on the 16-layer VGG network [15] pre-trained on millions of natural images of ImageNet for image classification [16], but there is a little difference with VGG16 network that the fully connected layers of VGG16 network is removed, and only keep the four blocks of convolutional parts called "*base framework*". The base framework consists of convolutional layers: $z_i = w_i \times x + b_i$, Rectified Linear Unit (ReLU) layers for non-linear activation function: $f(z_i) = \max(0, z_i)$, and max-pooling layers between two successive blocks, where x is the input of each convolutional layer, w_i is the convolution parameter, and b_i is the bias term. The three max-pooling layers divide the base network into four blocks of fine to coarse feature maps. Inspired by the work in [17, 18], specialized convolutional layers (with a 3×3 kernel size) with K (e.g. $K = 16$) feature maps are added after the convolutional layers at the end of each block. All the specialized layers are then rescaled to the original image size, and concatenated together. A last convolutional layer with kernel size 1×1 is added at the end of the network. This last layer combine linearly the fine to coarse feature maps in the concatenated specialized layers, and provide the final segmentation result.

3.1.2 Dataset Description

MRBrainS2018: It[3] provides 30 MRI scans, which contains three modalities such as T1-weighted, T1-weighted inversion recovery and T2-FLAIR. Seven of them are released as the training dataset. Another 23 scans are kept unreleased for test dataset. Its aim is to segment the 8 brain structure such as cortical gray matter, basal ganglia, white matter, white matter lesions, cerebrospinal fluid in the extracerebral space, ventricles, cerebellum and brain stem. The dataset includes same image size: 48× 240×240.

[3] https://mrbrains18.isi.uu.nl/

Atrial dataset [73]: 2018 atrial segmentation challenge released 100 annotated 3D MRIs from patients with atrial fibrillation. Its aim is to segment the left atrium. The pixel spacing of the MR images is 0.625 x 0.625 x 0.625 mm/pixel. The dataset includes two different image sizes: $88\times576\times576$ and $88\times640\times640$.

3.2 Experimental Results

For the 2018 atrial segmentation challenge dataset, based on the UNet framework, we obtain two segmentation results by using different preprocessing method as shown in Table. 3.1, but the difference between the two segmentation results is very small, only 0.59% in term of dice coefficient. Based on the improved FCN framework [19], for using the centralization method to preprocess the training data, the atrial segmentation results can reach 90.96%. If changing preprocessing method to standardization, the segmentation results do not show significant fluctuations.

TABLE 3.1: Segmentation results on the 2018 atrial segmentation challenge.

Method	Preprocessing	Dice/%
UNet [53]	centralization	89.86
	standardization	90.45
Improved FCN framework [19]	centralization	90.96
	standardization	90.03

We continue to experiment on the MRBrainS2018 dataset, which is different from the previous experiment because it is a multi-classification task rather than a binary classification task. The experimental results are shown in Table. 3.2. Based on the improved FCN framework [19], the segmentation results of 8 brain structures are obtained, although the segmentation results of white matter lesions and brain stem exist certain fluctuations for different preprocessing method, the segmentation results of another 6 structures remain stable. Therefore, these experiments does not indicate which preprocessing method is better. So we conducted another supplementary experiment.

TABLE 3.2: Segmentation results using the improved FCN framework [19] on the MRBrainS2018.

Segment labels	Dice/%	
	centralization	standardization
Cortical gray matter	85.30	84.96
Basal ganglia	79.02	80.92
White matter	84.68	84.43
White matter lesions	61.3	58.66
Cerebrospinal fluid in the extracerebral space	84.1	84.81
Ventricles	93.89	94.55
Cerebellum	91.84	91.63
Brain stem	86.11	83.97

As we all know, noise is everywhere. To evaluate the quality of a network, anti-noise is also one of the indicators, so we continue to explore the sensitivity of the network to noise for different preprocessing methods. In order to add different noises to the original image, we use a python library called **imgaug**[4], which can help user with augmenting images for machine learning projects. We mainly use functions *AdditiveGaussianNoise()* and *SaltAndPepper()* of imgaug library. For the *AdditiveGaussianNoise()*, if user wants to add gaussian noise to an image, it will sample once per pixel from a normal distribution $N(0, s)$, where s is sampled per image and varies between 0 and $s * 255$ as follows:

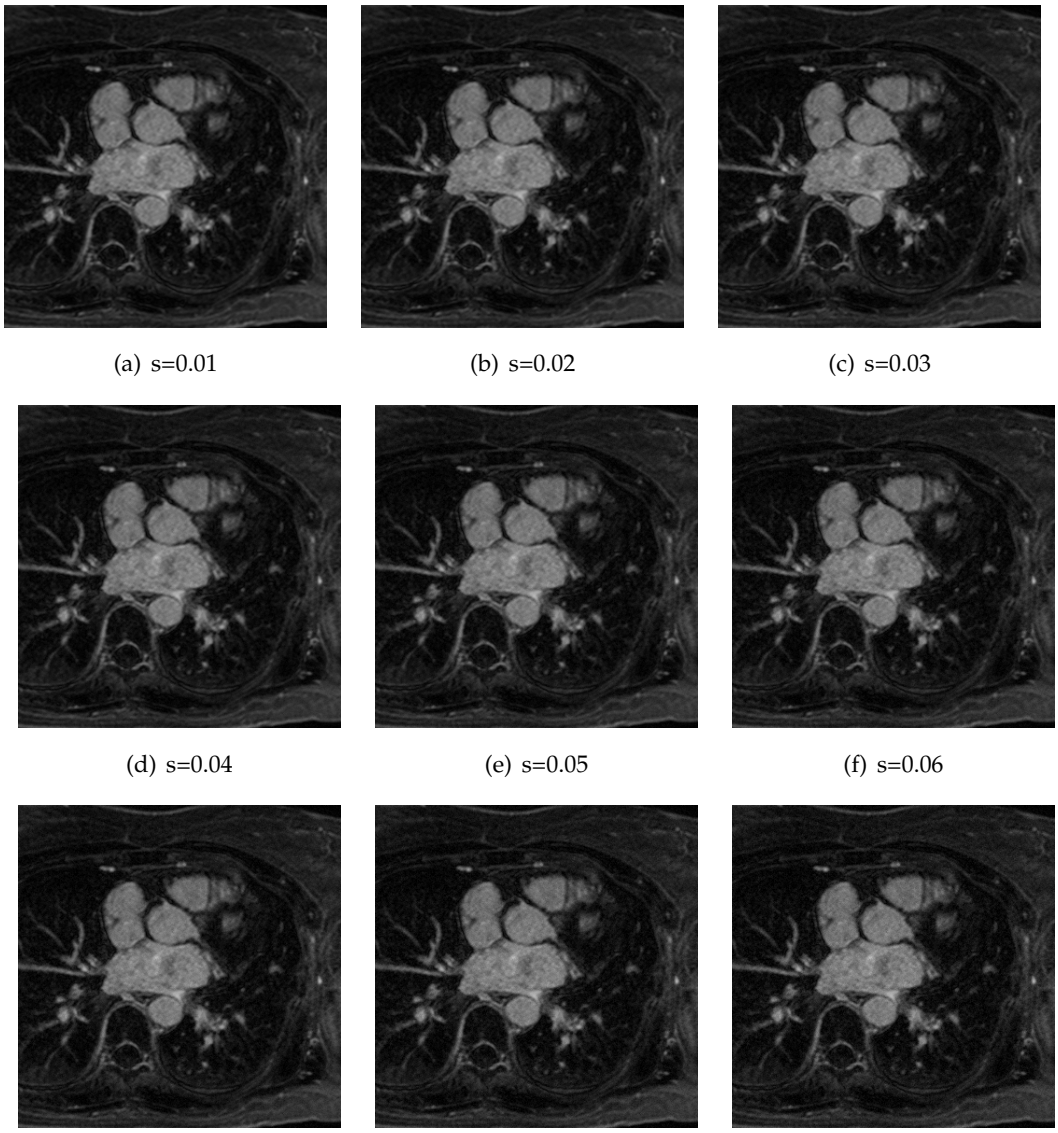
```
import imgaug.augmenters as iaa
aug = iaa.AdditiveGaussianNoise(scale=(0, s*255))
```

For the *SaltAndPepper()*, it means that replaces p such as 10% of all pixels with salt and pepper noise as follows:

```
import imgaug.augmenters as iaa
aug = iaa.SaltAndPepper(p)
```

For adding the gaussian noise to images, as shown in Fig. 3.2, we change the parameter s of function *AdditiveGaussianNoise()* from 0.01 to 0.09. As s increases, there is more and more noise in the image. For adding the salt and pepper noise, we test the network based on the parameter p of function *SaltAndPepper()* in two different orders of magnitude, which is from 0.01 to 0.09 and from 0.001 to 0.009, respectively. As can be seen in Fig. 3.3, the same phenomenon occurs as when Gaussian noise is added, i.e., as the parameter p increases, the noise becomes stronger.

[4]https://imgaug.readthedocs.io/en/latest/

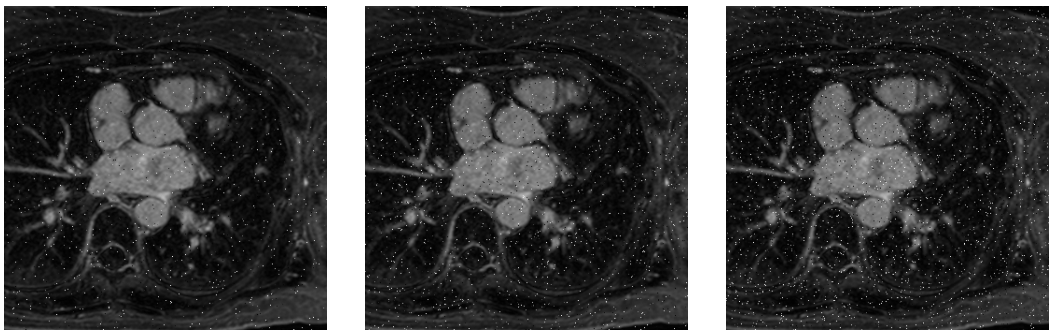


(a) s=0.01 (b) s=0.02 (c) s=0.03

(d) s=0.04 (e) s=0.05 (f) s=0.06

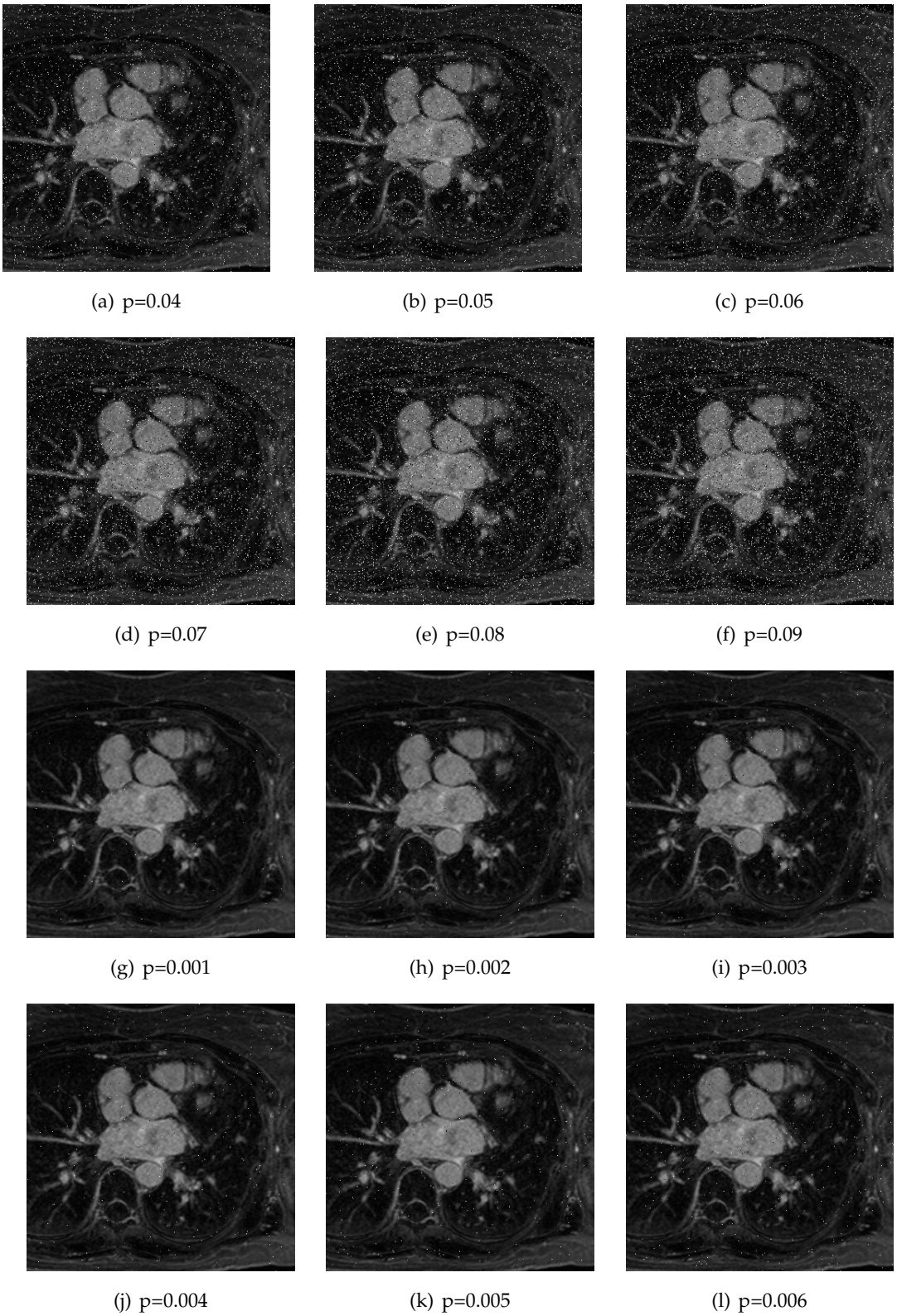
(g) s=0.07 (h) s=0.08 (i) s=0.09

FIGURE 3.2: Adding gauss noise to the original image



(a) p=0.01 (b) p=0.02 (c) p=0.03

FIGURE 3.3: Adding salt and pepper noise to the original image



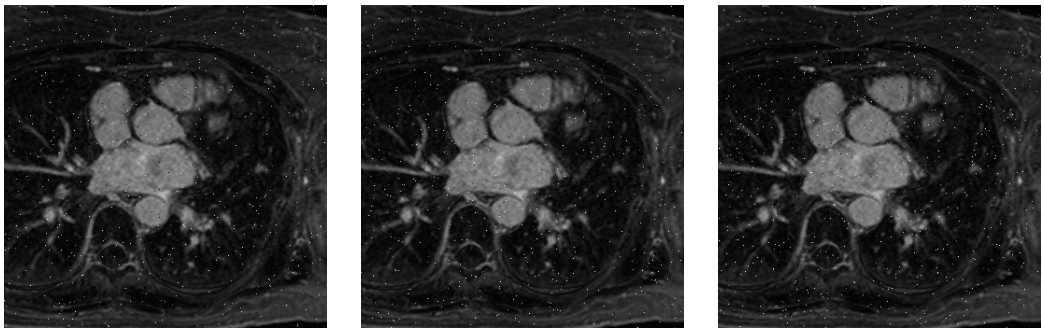
(a) p=0.04 (b) p=0.05 (c) p=0.06

(d) p=0.07 (e) p=0.08 (f) p=0.09

(g) p=0.001 (h) p=0.002 (i) p=0.003

(j) p=0.004 (k) p=0.005 (l) p=0.006

FIGURE 3.3: Adding salt and pepper noise to the original image

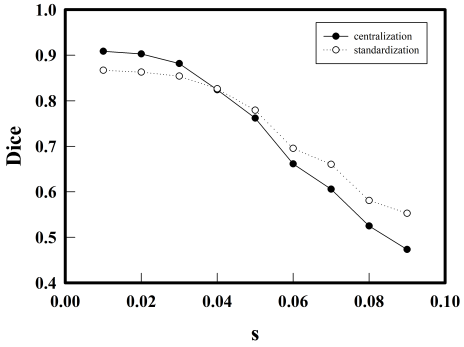


(a) p=0.007 (b) p=0.008 (c) p=0.009

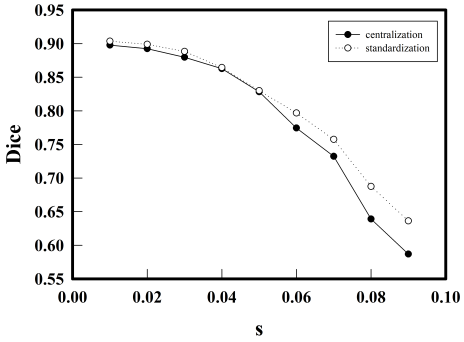
FIGURE 3.3: Adding salt and pepper noise to the original image

As shown in Fig. 3.4, it shows the sensitivity of networks to noise for different preprocessing methods on the 2018 atrial segmentation challenge. For the Fig. 3.4(a) and Fig. 3.4(b), facing gauss noise, there is a clear distinction when $s = 0.04$. If $s < 0.04$, the dice coefficient fluctuates less, but $s > 0.04$, the dice coefficient drops sharply, which is an overall trend no matter what kind of preprocessing method. However, from the overall trend, the dice coefficient based on standardization decreases more slowly than centralization. For the Fig. 3.4(c) and Fig. 3.4(d), facing salt and pepper noise, there is an obvious difference to the Gaussian noise. Their changing trends are the concave function, which denotes that the dice coefficient always drops sharply. In the Fig. 3.4(c), standardization is significantly better than centralization in term of anti-noise. In the Fig. 3.4(d), standardization is still better than centralization between p=0.02 and p=0.06, but finally, they approach a very low dice coefficient and the dice coefficient of standardization is always higher than the centralization.

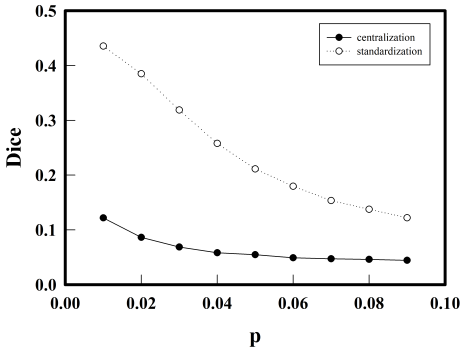
We continue to explore different network whether different networks have different sensitivity to noise. Compared Fig. 3.4(a) and Fig. 3.4(c) with Fig. 3.4(b) and Fig. 3.4(d), for the same parameter s and p, the dice coefficient based on UNet is always higher than the improved FCN framework [19], therefore, different networks have certain differences in sensitivity to noise.



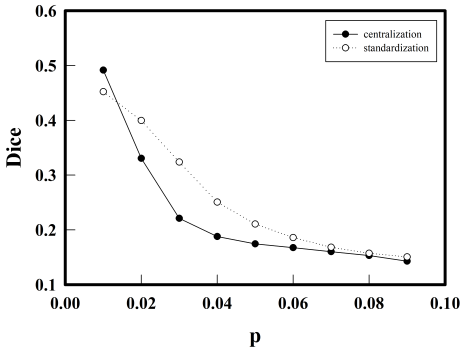
(a) Based on the improved FCN framework [19]



(b) Based on UNet



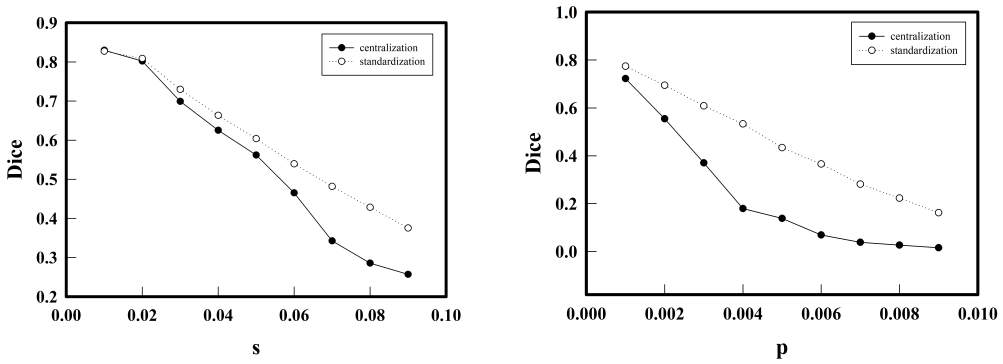
(c) Based on the improved FCN framework [19]



(d) Based on Unet

FIGURE 3.4: The sensitivity of networks to noise for different preprocessing methods on the 2018 atrial segmentation challenge

The 2018 atrial segmentation challenge dataset is only binary segmentation task, so we continue to choose one different dataset that belongs to multi-class segmentation task. As shown in Fig. 3.5, it shows the sensitivity of the improved FCN framework [19] to noise for different preprocessing methods on the MRbrains2018 dataset. For Fig. 3.5, to facilitate the summary of the trend, the dice coefficient denotes the mean value of segmentation results of 8 brain structures. No matter what kind of noise, the dice coefficient based on standardization is always higher than the centralization. For the multi-class segmentation task, facing to the standardization preprocessing method, the relationship between dice coefficient and noise is approximately linear.



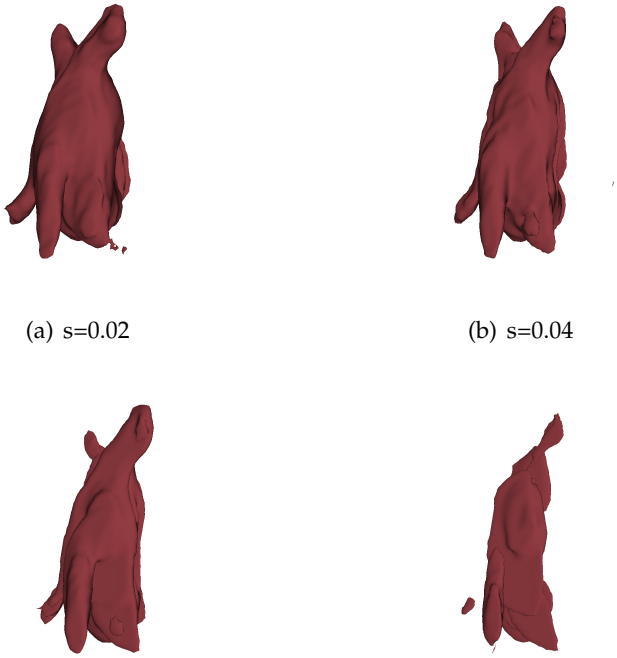
(a) (b)

FIGURE 3.5: The sensitivity of the improved FCN framework [19] to noise for different preprocessing methods on the MRbrains2018 dataset

According to Fig. 3.4 and Fig. 3.5, we can make one conclusion that, for different preprocessing methods, the sensitivity of networks to noise is different. The standardization makes the network more anti-noise as the noise increases than the centralization.

Fig. 3.6 and Fig. 3.7 denote the 3D segmentation results of one sample based on the improved FCN framework [19] for different preprocessing methods. For the same parameter s, the integrity of left atrial segmentation based on the standardization is better than the centralization.

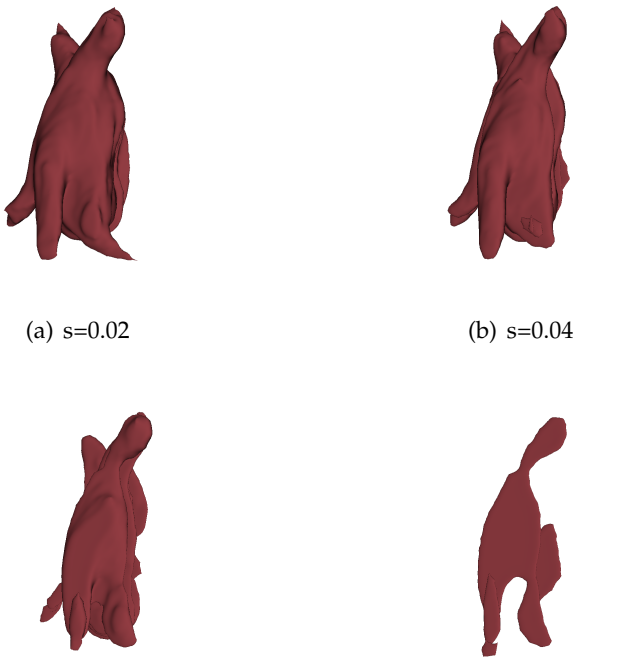
Fig. 3.8 and Fig. 3.9 illustrate the 3D segmentation results based on UNet for different preprocessing methods. There is one same phenomenon as Fig. 3.6 and Fig. 3.7 when the parameter s is same, which is that using the standardization can get better integrity of left atrial segmentation than the centralization. In Fig. 3.8, the surface of the segmentation results is smoother than the segmentation results of Fig. 3.9, which also explains that the standardization preprocessing method makes the network more robust to noise than the centralization.



(a) s=0.02 (b) s=0.04

(c) s=0.06 (d) s=0.09

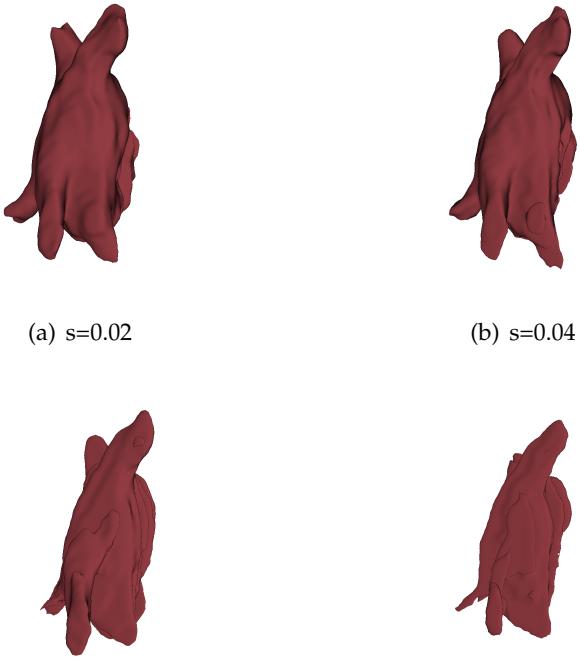
FIGURE 3.6: The 3D segmentation results based on the improved FCN framework [19] for standardization



(a) s=0.02 (b) s=0.04

(c) s=0.06 (d) s=0.09

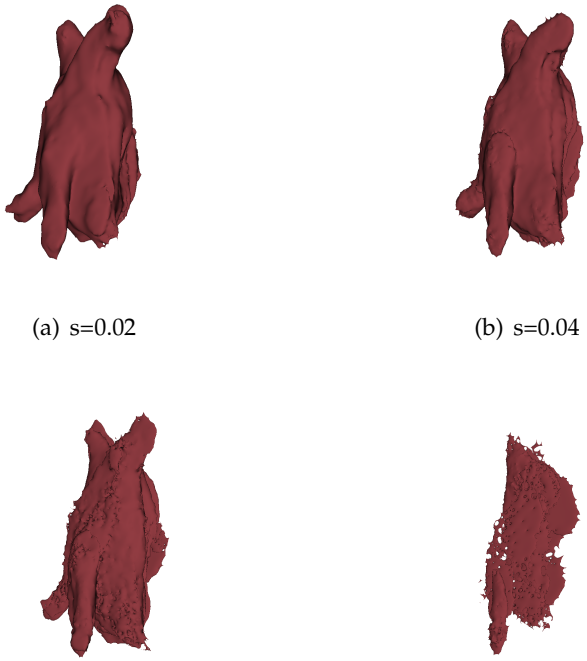
FIGURE 3.7: The 3D segmentation results based on the improved FCN framework [19] for centralization



(a) s=0.02 (b) s=0.04

(c) s=0.06 (d) s=0.09

FIGURE 3.8: The 3D segmentation results based on UNet for standardization



(a) s=0.02 (b) s=0.04

(c) s=0.06 (d) s=0.09

FIGURE 3.9: The 3D segmentation results based on UNet for centralization

3.3 Conclusion

In this chapter, we have explored the impact of different preprocessing methods on the output of network. We have seen that using the standardization preprocessing method makes the network more robust to noise than the centralization method, and different networks have different sensitivity to noise. Therefore, we will choose standardization as our preprocessing method to process the dataset.

Chapter 4

Two-stage Segmentation Method

For medical images, in addition to the object regions, there are a large number of background regions, which affects the segmentation accuracy. Therefore, in this chapter, we first localize roughly the object to reduce the influence of the background, and then crop the object regions to segment.

4.1 Methodology

4.1.1 Overview of Network Architecture

The global overview of our A^0Net consists of two parts (localization and segmentation) as depicted in Fig. 4.1, and the architecture of our networks in Fig. 4.2. The first part (the "localization network") is used to localize roughly the object position. The second part is devoted to segment the object (the "segmentation network").

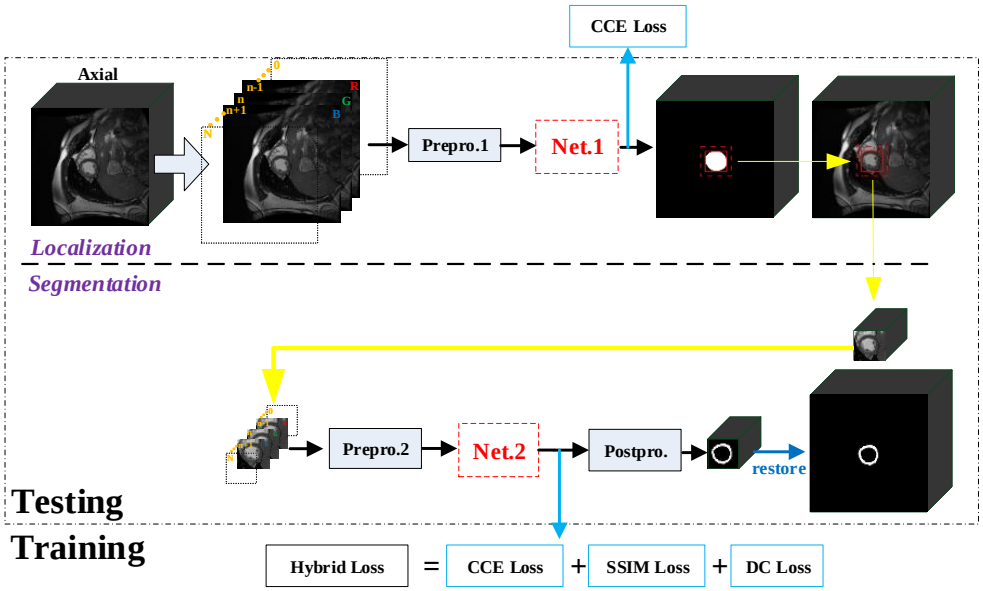


FIGURE 4.1: Global overview of the proposed method (A^0Net).

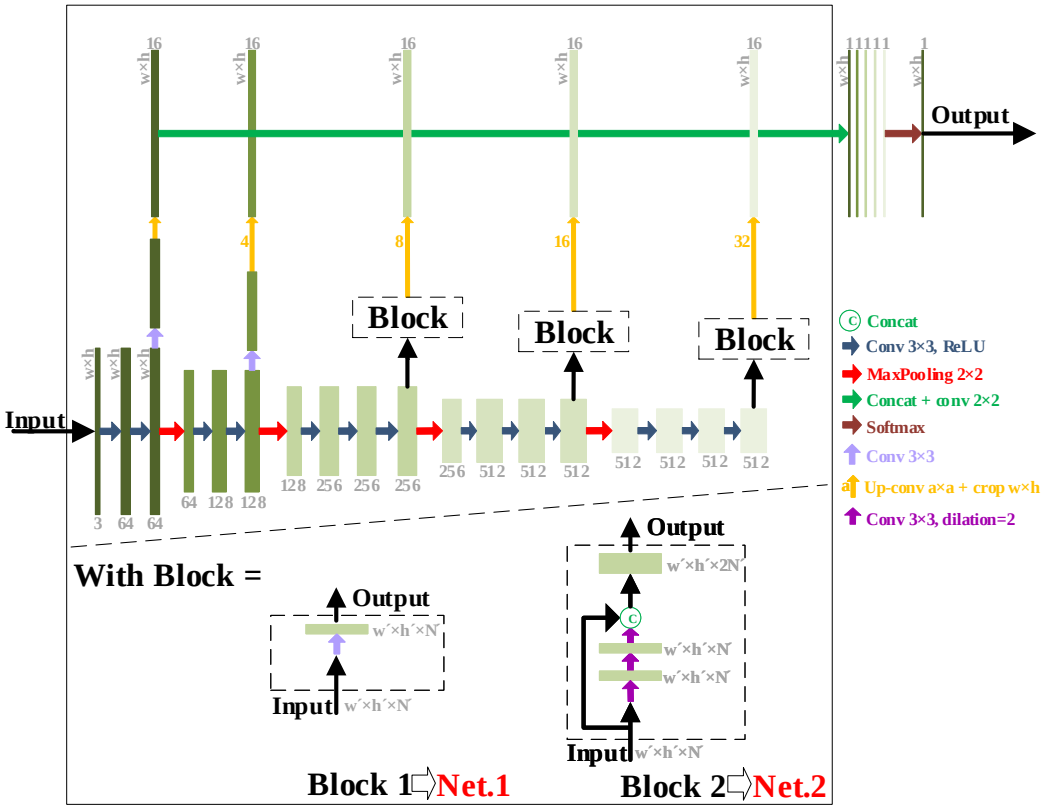


FIGURE 4.2: Architecture of our networks. **Block 1** and **Block 2** correspond to the components of **Net.1** and **Net.2** of Fig. 4.1, respectively. Because the role of **Net.1** is only to roughly locate the target, using **Block 1** instead of **Block 2** can both reduce model parameters and improve the speed of model prediction. N denotes the number of feature map.

4.1.2 Localization Network

The localization network (**Net.1**) is depicted in Fig. 4.2. The black dotted box **Block 1** is dedicated to the localization network, it can be replaced by **Block 2** to become the segmentation network (**Net.2**). For **Net.1** and **Net.2**, the difference concerns only **Block 1** and **Block 2** as shown in Fig. 4.2, while the other components of the architecture are the same. **Block 1** consists of one convolutional layers with 256 or 512. First, we rely on the original VGG16 [15] network architecture, pre-trained on millions of natural images of ImageNet for image classification [16]. We then discard its fully connected layers to keep only the sub-network made of five convolution-based "stages" (the base network). Each stage is made of two convolutional layers, a ReLU activation function, and a max-pooling layer. Since the max-pooling layers decrease the resolution of the input image, we obtain a set of fine to coarse feature maps (with 5 levels of features). Inspired by the works in [17–20], we added *specialized* convolutional layers (with a 3×3 kernel size) with K (*e.g.* $K = 16$) feature maps after the up-convolutional layers placed at the end of each stage. The outputs of the specialized layers show the same resolution than the input image, and are concatenated

together. We add a 1×1 convolutional layer at the output of the concatenation layer to linearly combine the fine to coarse feature maps [1].

4.1.3 Segmentation Network

As mentioned above, we replace **Block 1** of **Net.1** with **Block 2**, which becomes the segmentation network (**Net.2**). Because the role of **Net.2** is mainly to obtain accurate segmentation results, we use **Block 2** that is more complicated than **Block 1** in Fig. 4.2. It can capture the global information and decrease the effect of surrounding similar tissues. **Block 2** consists of three convolutional layers with 256 or 512 dilated (dilation = 2) [21] 3×3 filters, and one layer of concatenation.

4.1.4 Hybrid Loss

To obtain high quality regional segmentation and nice boundaries, we define ℓ as a hybrid loss: $\ell = \lambda_1\ell_{\text{CCE}} + \lambda_2\ell_{\text{SSIM}} + \lambda_3\ell_{\text{DC}}$, where ℓ_{CCE}, ℓ_{SSIM} and ℓ_{DC} respectively denote CCE loss [22], SSIM loss [23] and DC loss [24] respectively, $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

CCE [22] loss is commonly used for multi-class classification and segmentation. It is defined as:

$$\ell_{\text{CCE}} = -\sum_{i=1}^{C}\sum_{a=1}^{H}\sum_{b=1}^{W} y^i_{(a,b)} \ln {y_*}^i_{(a,b)}, \tag{4.1}$$

where C is the number of classes of each image, H and W are the height and width of image, $y^i_{(a,b)} \in \{0,1\}$ is the ground truth one-hot label of class i in the position (a,b) and ${y_*}^i_{(a,b)}$ is the predicted probability of class i.

SSIM loss can assess image quality [23], and can be used to capture the structural information, which will decrease the mis-segmentation rate of surrounding similar tissues. Therefore, we integrated it into our training loss to learn the differences between the segmented domain and similar tissues around the segmented domain. Let **S** and **G** be the predicted probability map and the ground truth mask respectively, the SSIM of **S** and **G** is defined as:

$$\ell_{\text{SSIM}} = 1 - \frac{(2\mu_{\text{S}}\mu_{\text{G}} + C_1)(2\sigma_{\text{SG}} + C_2)}{(\mu_{\text{S}}^2 + \mu_{\text{G}}^2 + C_1)(\sigma_{\text{S}}^2 + \sigma_{\text{G}}^2 + C_2)}, \tag{4.2}$$

where μ_{S}, μ_{G} and σ_{S}, σ_{G} are the mean and standard deviations of **S** and **G** respectively, σ_{SG} is their covariance, $C_1 = 0.01^2$ and $C_2 = 0.03^2$ are used to avoid a division by zero.

DC [24] loss is used to measure the similarity between two sets as defined in Eq. 2.36. But for the multi-class segmentation task, Eq. 4.3 is not suitable due to the class imbalance problem in such cases. Therefore, we extend the definition of the DC loss to multi-class segmentation in the following manner:

$$dice_i = \left(\epsilon + 2\sum_{n=1}^{N_i} y^i_n\, {y_*}^i_n\right) / \left(\epsilon + \sum_{n=1}^{N_i}\left(y^i_n + {y_*}^i_n\right)\right) \tag{4.3}$$

[1] Note that we designed our network's architecture to work with any input shape.

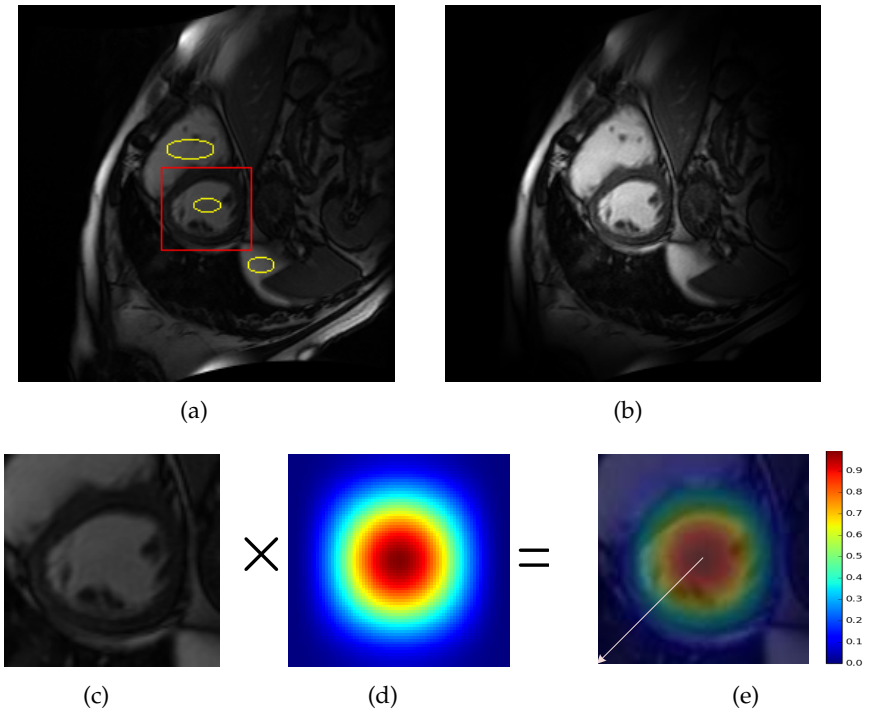


FIGURE 4.3: Gaussian-like attention (GA). (a) Original image. Red rectangle denotes segmented object, and yellow ellipse denotes similar tissues. (b) Gaussian-like attention image of (a) by using Eq. 4.5. (c) The cropped image after locating the segmented object (red rectangle). (d) The image of the Gaussian-like weighted function (ω_{GA}). (e) The image after blending (c) and (d).

$$\ell_{\text{DC}} = 1 - \sum_{i=1}^{C} dice_i / (N_i + \epsilon), \tag{4.4}$$

where N_i denotes the numbers of class i and ϵ is a smooth factor.

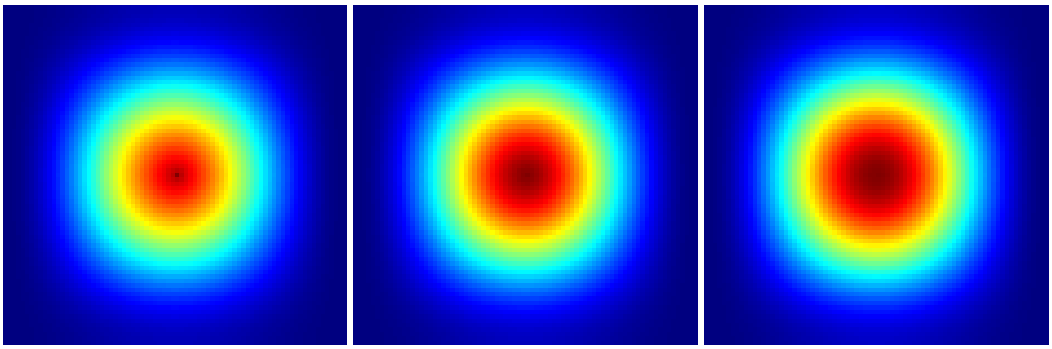
4.1.5 Gaussian-like Attention (GA)

Fig. 4.3(a) is an example from the MICCAI 2019 left ventricle (LV) Full Quantification Challenge dataset[2] (LVQuan19) [74, 75]. The red box denotes the object domain, here the left ventricle. A large number of similar tissues are around it, highlighted by the yellow ellipses. As shown in Fig. 4.7, the similar tissues can lead to mis-segmentation. Even after a localization procedure, these tissues are still present. An idea to decrease their impact on segmentation results is to get inspired by the biological visual system, which concentrates on certain image regions requiring detailed analysis [121]. We define the GA as: $I_{\text{GA}}(a,b) = I(a,b)\omega_{\text{GA}}(a,b)$, where $I(a,b)$ denotes the image intensity at location (a,b) and $\omega_{\text{GA}}(a,b)$ is a Gaussian-like weighted function defined by

$$\omega_{\text{GA}}(a,b) = \alpha \exp^{-|\frac{(a,b)-(a^*,b^*)}{\delta}|^{\beta}} \tag{4.5}$$

where (a^*,b^*) denotes the object center, α is a normalization constant, δ is a scale parameter, and β is a shape parameter. As shown in Fig 4.4, when we keep the

[2] https://lvquan19.github.io



(a) β=0.5 (b) β=1 (c) β=2

FIGURE 4.4: Different β.

rest of the parameters unchanged, the change in β leads to a change in the range of attention.

As shown in Fig 4.5, we use temporal information, therefore, in order to cooperate with the operation of temporal information, we refine the normalization constant α at Eq. 4.12 by Hamming window at Eq. 4.6 and Hanning window at Eq. 4.8.

$$\kappa(p) = a_0 - (1 - a_0) \cdot cos\left(\frac{2\pi \cdot p}{H-1}\right), \quad 0 \leq p \leq H-1 \tag{4.6}$$

where $a_0 = 0.53836$ because of Hamming windows. H denotes height.

$$\mathbf{A}(p) = \Big[\kappa(0), \quad \cdots, \quad \kappa(p), \quad \cdots, \quad \kappa(H-1)\Big]_{H\times 1} \tag{4.7}$$

$$\nu(q) = 0.5 \cdot cos\left(\frac{2\pi \cdot q}{W-1}\right), \quad 0 \leq q \leq W-1 \tag{4.8}$$

where W denotes width.

$$\mathbf{B}(q) = \begin{bmatrix} \nu(0) \\ \vdots \\ \nu(q) \\ \vdots \\ \nu(W-1) \end{bmatrix}_{1\times W} \tag{4.9}$$

$$\Theta(a,b) = \exp^{-|\frac{(a,b)-(a^*,b^*)}{\delta}|^\beta} \tag{4.10}$$

$$\mathbf{C}(a,b) = \begin{bmatrix} \Theta(1,1) & \cdots & \Theta(1,W) \\ \vdots & \ldots & \vdots \\ \Theta(H,1) & \cdots & \Theta(H,W) \end{bmatrix}_{H\times W} \tag{4.11}$$

$$\alpha = \frac{\mathbf{A}(p)\mathbf{B}(q)}{\sum_{a=1}^{H}\sum_{b=1}^{W}\mathbf{A}(p)\mathbf{B}(q) \otimes \mathbf{C}(a,b)} \tag{4.12}$$

$$\delta = \frac{H + W}{2} \tag{4.13}$$

Because the cropped image size constantly changes after locating, δ is used for decreasing the impact of image size.

If $I_{\mathrm{GA}}(a, b)$ is directly applied on each original image, the object of interest would probably be missed. Therefore, we first must find the region of interest; and then use $I_{\mathrm{GA}}(a, b)$ to focus on the object. This procedure is depicted in Fig. 4.3(e), where similar tissues are clearly less important compared to Fig. 4.3(c).

4.2 Experimental Results

4.2.1 Dataset Description

We evaluated our method on four datasets: LVQuan19, the MICCAI Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease[3] (HVSMR16), Multi-Modality Whole Heart Segmentation [4] (MM-WHS2017) and MICCAI 2018 Atrial Segmentation Challenge [5] (AtriaSeg18). The aim of **LVQuan19** is to segment the myocardium of the left ventricle and estimate a set of clinical significant LV indices such as regional wall thicknesses, cavity dimensions, and cardiac phase and so on. It contains the processed SAX MR sequences of 56 patients. For each patient, 20 temporal frames are given and cover a whole cardiac cycle. All ground truth (GT) values of the LV indices are provided for every single frame. The pixel spacings of the MR images range from 0.6836 mm/pixel to 1.5625 mm/pixel, with mean values of 1.1809 mm/pixel. The LV dataset includes two different image sizes: 256×256 or 512×512 pixels. The aim of **HVSMR16** [71] is to segment myocardium and blood pool, it contains 10 training cardiovascular magnetic resonance (CMR) scans. For each patient, three kinds of images were provided: the full-volume axial images, the cropped axial images around the heart and thoracic aorta, and the cropped short axis reconstruction. In the current work, we only use the full-volume axial images. The slice spacings of the full-volume axial images range from 0.65 mm/pixel to 1.15 mm/pixel, while in-plane resolution ranged from 0.73 mm/pixel to 1.15 mm/pixel. The average sizes: $387 \times 387 \times 165$ pixels. **MM-WHS2017** [72] aims to segment 7 substructures of the whole heart. Although it contains 20 cardiac MRI and 20 CT images, we only use the MRI modality. The slice spacings of MRI volume range from 0.899 mm/pixel to 1.60 mm/pixel, while in-plane resolution ranged from 0.78 mm/pixel to 1.2 mm/pixel. The average sizes: $324 \times 325 \times 171$ pixels. **AtriaSeg18** aims to segment the left atrium and contains 100 annotated 3D MRIs from patients with atrial fibrillation. The voxel

[3] http://segchd.csail.mit.edu/index.html
[4] http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs17/index.html
[5] http://atriaseg2018.cardiacatlas.org/

size of the MR images is $0.625 \times 0.625 \times 0.625$ mm. The dataset includes two different image sizes: $88 \times 576 \times 576$ pixel and $88 \times 640 \times 640$ pixel.

4.2.2 Preprocessings

Since the VGG-16 network's input is an RGB image, we propose to take advantage of the temporal information by stacking 3 successive 2D frames: to segment the n^{th} slice, we use the n^{th} slice of the MR volume, and its neighboring $(n-1)^{th}$ and $(n+1)^{th}$ slices, as green, red and blue channels, respectively. This new image, named "temporal-like" image, enhances the area of motions, here the heart, as shown in Fig. 4.5.

Let us remind what we call *Gauss normalization*: for each $(2D+t)$-image I corresponding to a given patient, we compute $I := (I-\mu)/\sigma$ where μ is the mean of I and σ its standard deviation (σ is assumed not to be equal to zero). There are then two different pre-processing steps as depicted in Fig. 4.1.

1) The first pre-processing (see **Prepro.1** in Fig. 4.1) begins with a Gauss normalization. Then, for each n, we created the $width \times height \times 3$ pseudo-color ("temporal-like") image where R, G, B correspond respectively to the $n-1, n, n+1$ frames and we concatenate them.

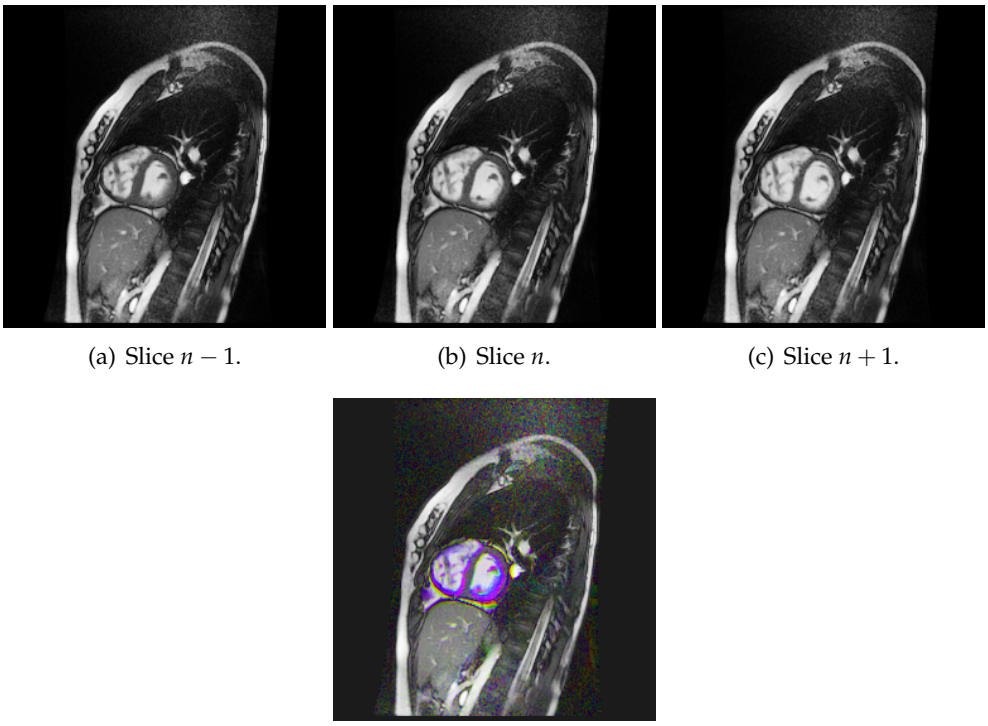
2) The second pre-processing (**Prepro.2** in Fig. 4.1) follows five steps: (1) data augmentation using rotations and flips for the LVQuan19 dataset (only for the training phase), but it is not used on the HVSMR16, MM-WHS2017 and AtriaSeg18 dataset, (2) resizing with a fixed pixel-spacing ($0.65mm$), (3) GA, (4) Gauss normalization, and (5) pseudo-color concatenated image like above. Such a use of a pseudo-color image in the context of 3D medical imaging has been proven effective in [122] to segment brain structures and in [123] to extract white matter hyperintensities in brain volumes.

4.2.3 Postprocessing

Let us assume that we crop an initial volume of T frames of size $T \times W \times H$ into an image of size $T \times w \times h$ (where the crop is due to the localization procedure, and W and H are the initial width and height of a slice). After **Prepro.2** we obtain a $T \times w \times h \times 3$ image. Then we filter the ouput of the segmentation network, of size $T \times w \times h$, by keeping only the greatest connected component, in order to get back the initial pixel-spacing. Finally, we add a padding of zeros to get back a $T \times W \times H$ image.

4.2.4 Implementation and Experimental Setup

We implemented our experiments on Keras/TensorFlow using a NVidia Quadro P6000 GPU. For the localization network, we used the multinomial logistic loss function for a one-of-many classification task, passing real-valued predictions through a softmax to get a probability distribution over classes. We used an Adam optimizer



(a) Slice $n-1$. (b) Slice n. (c) Slice $n+1$.

(d) RGB concatenation at n.

FIGURE 4.5: Illustration of our "temporal-like" procedure.

(batchsize = 1, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 0.001$, lr $= 0.002$) and we did not use learning rate decay. We trained the network during 10 epochs. For this step, we merged all the classes into the object class to obtain a binary segmentation. For the segmentation network, we used the same optimizer and parameters detailed previously. We used the hybrid loss as loss function. For this task, we considered three different classes (background, myocardium, cavity) for LVQuan19, three different classes (background, myocardium, blood pool) for HVSMR2016 and eight different classes (background, myocardium, left atrium, left ventricle, right atrium, right ventricle, ascending aorta and pulmonary artery) for MM-WHS2017.

4.2.5 Evaluation Methods

Three measures are used to evaluate our method: DC given in Eq. 2.36, 95% in the Hausdorff distance (95HD) [124] and Boundary of Dice Coefficient (BDC) to quantitatively evaluate the boundaries. As many diseases appear in the myocardium wall, we chose to quantitatively evaluate the precision of the segmentation on boundaries.

Fig. 4.6 shows the illustration of BDC procedure. For the BDC evaluation method, given a segmentation map M, we first convert the class i to a binary mask, M^i_{bm}. Then, we obtain the mask of class i of its one pixel wide boundary by conducting an XOR(M^i_{bm}, M^i_{erd}) operation where M^i_{erd} is the eroded binary mask of M^i_{bm}. The same method is used to get the GT mask boundaries, M^i_{g}. Then the DC is calculated on the boundaries of the GT and segmentation masks to obtain the BDC.

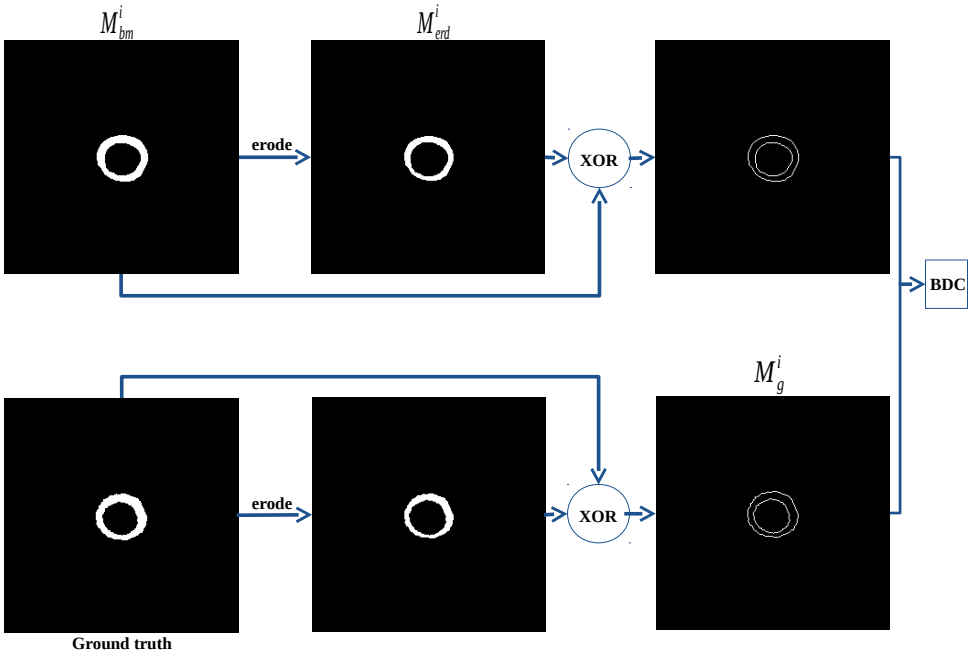


FIGURE 4.6: Illustration of BDC procedure.

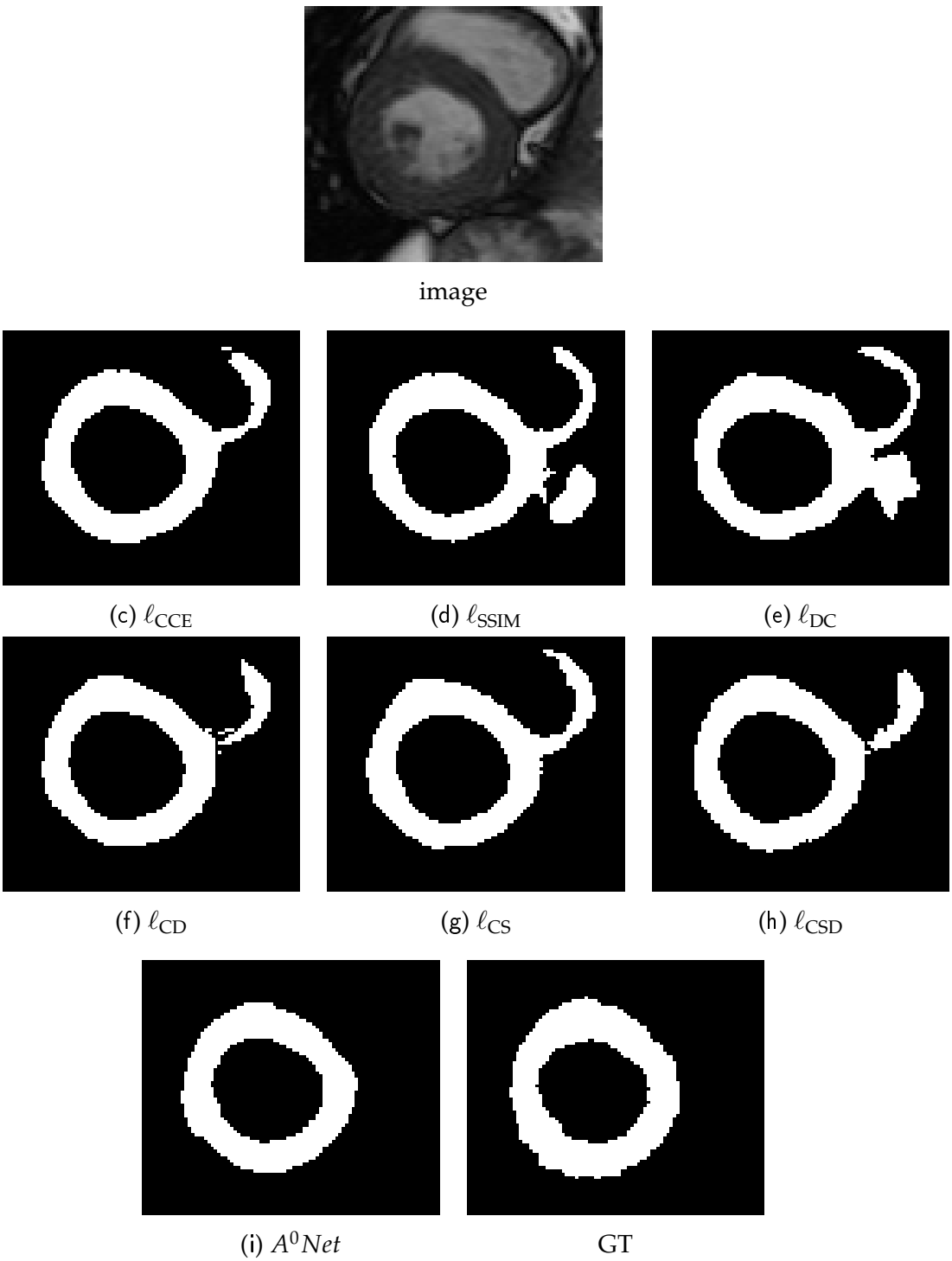
4.2.6 Ablation Study

TABLE 4.1: Ablation study; Dice values are for the myocardium.

Ablation	Configurations	DC/%	95HD	BDC
Architecture	a: B. + ℓ_{CCE}	84.15	3.186	0.269
	b: B. + L. + ℓ_{CCE} [20]	86.68	2.209	0.281
	c: BLP + ℓ_{CCE}	87.74	2.019	0.303
Loss	d: BLP + ℓ_{SSIM}	87.30	2.094	0.297
	e: BLP + ℓ_{DC}	87.11	2.193	0.295
	f: BLP + ℓ_{CD}	87.53	2.071	0.300
	g: BLP + ℓ_{CS}	87.77	2.043	0.303
	h: BLP + ℓ_{CSD}	87.87	1.912	0.305
A^0Net (our method)	i: BLP + GA + ℓ_{CSD}	**87.93**	**1.826**	**0.306**
UNet [53]	-	86.20	3.976	0.291

"B." means "baseline" (**Net.1**) [125, 126]; "L." means "localization"; "P." means "Block 2"(**Net.2**); "BLP" means "baseline + localization + P".

Note: $\ell_{CD} = \ell_{CCE} + \ell_{DC}$; $\ell_{CS} = \ell_{CCE} + \ell_{SSIM}$; $\ell_{CSD} = \ell_{CCE} + \ell_{SSIM} + \ell_{DC}$.



image

(c) ℓ_{CCE} (d) ℓ_{SSIM} (e) ℓ_{DC}

(f) ℓ_{CD} (g) ℓ_{CS} (h) ℓ_{CSD}

(i) A^0Net GT

FIGURE 4.7: The comparative results trained with our A^0Net on different losses.

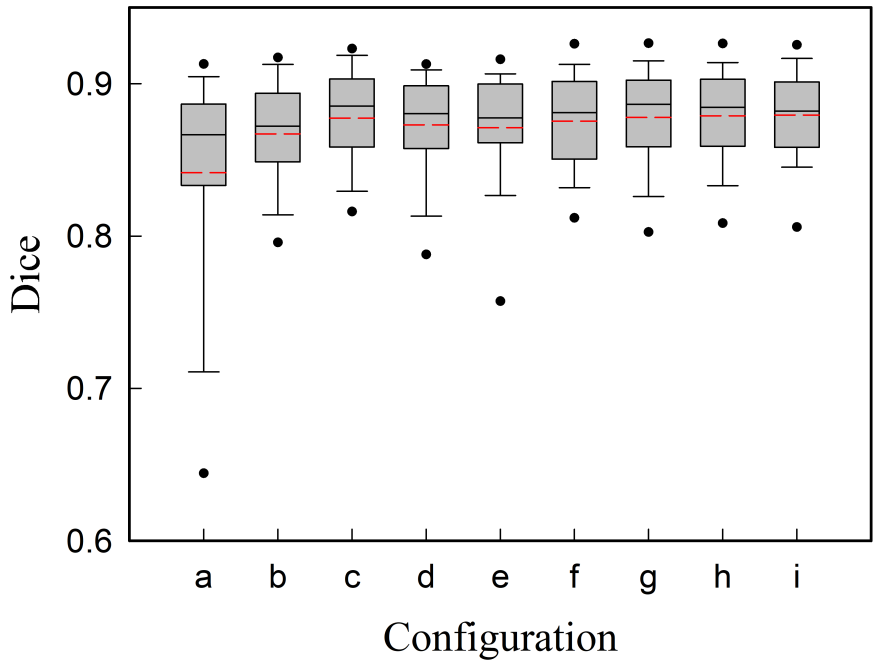


FIGURE 4.8: Box plots of dice scores for the 56 patients. The red dotted line represents the average value, and a, b, c, etc. on the abscissa correspond to the methods of Tbl. 4.1

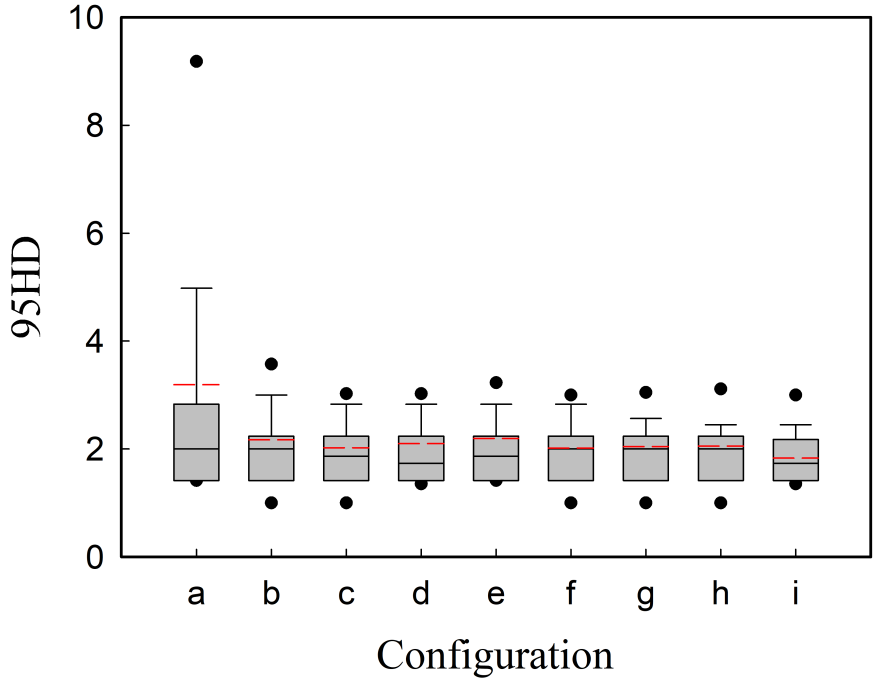


FIGURE 4.9: Box plots of 95HD for the 56 patients. The red dotted line represents the average value, and a, b, c, etc. on the abscissa correspond to Tbl. 4.1

To validate the influence of each component used in our method, we conducted the ablation study that includes three parts (architecture, loss and GA) on the LVQuan19 dataset with 5-fold cross-validation. Results are shown in Tbl. 4.1. **Architecture ablation**: To demonstrate the effects of our A^0Net, we compared the results of our method with other related frameworks. We took a network used in our previous works [125, 126] as baseline network (**Net.1**). First, we added a localization module (as shown in Fig. 4.1) based on the baseline; with this module, we obtained a mean improvement of 1.89% in terms of DC, 0.9772 on 95HD, which meant that reducing the proportion of the background in the image is beneficial to improve segmentation accuracy. This architecture was the one we presented for the Challenge LVQUAN19 [20]. Further, we added the **Block 2** module, so **Net.1** was changed to **Net.2** (Baseline+Block2) as shown in Fig. 4.2. We learned from our comparison results that, when using dilated convolution and capturing the global information in the feature maps of high level, we could refine the segmentation results, which meant further improvement of 1.70% in terms of DC, 0.1893 on 95HD. **Loss ablation**: To prove the effects of our hybrid loss, we conducted comparative experiments over different losses based on our method. The results in Tbl. 4.1 illustrate that the proposed hybrid loss helps to improve the performance, and, compared with other combinations, that loss function based on three-level hierarchy (pixel-, patch- and map-level) can fully guide the network to study the transformation relationship between the input image and the corresponding label. **GA ablation**: As shown in Fig. 4.7, without GA, the surrounding similar tissues are mis-segmented, meaning that the segmentation results are disturbed by these similar tissues, and mis-segmented parts are connected to the ground truth, which is very difficult to remove. Therefore, by using our GA module, we decrease the impact of the surrounding similar tissues, and the segmentation results are better.

λ **ablation** To explore the influence of the proportionality coefficient λ of the hybrid loss on the segmentation results, we continued to conduct the λ ablation study, and its results were shown in Tbl. 4.2. As shown in the Tbl. 4.2, if the proportionality coefficient $\lambda_1 = \lambda_2 = \lambda_3 = 1$, the segmentation results will be best. Compared with the Tbl. 4.1, no matter what the proportionality coefficient is, the 95HD is still around 1.85, which verifies that the proposed hybrid loss can preserve more boundary details than a single loss. If added a higher weight to any one of the three loss functions of the hybrid loss, or added a higher weight to any two of the three loss functions, the 95HD is lower than unweighted hybrid loss, which meant that in different level hierarchy, it is best not to bias against a certain level hierarchy.

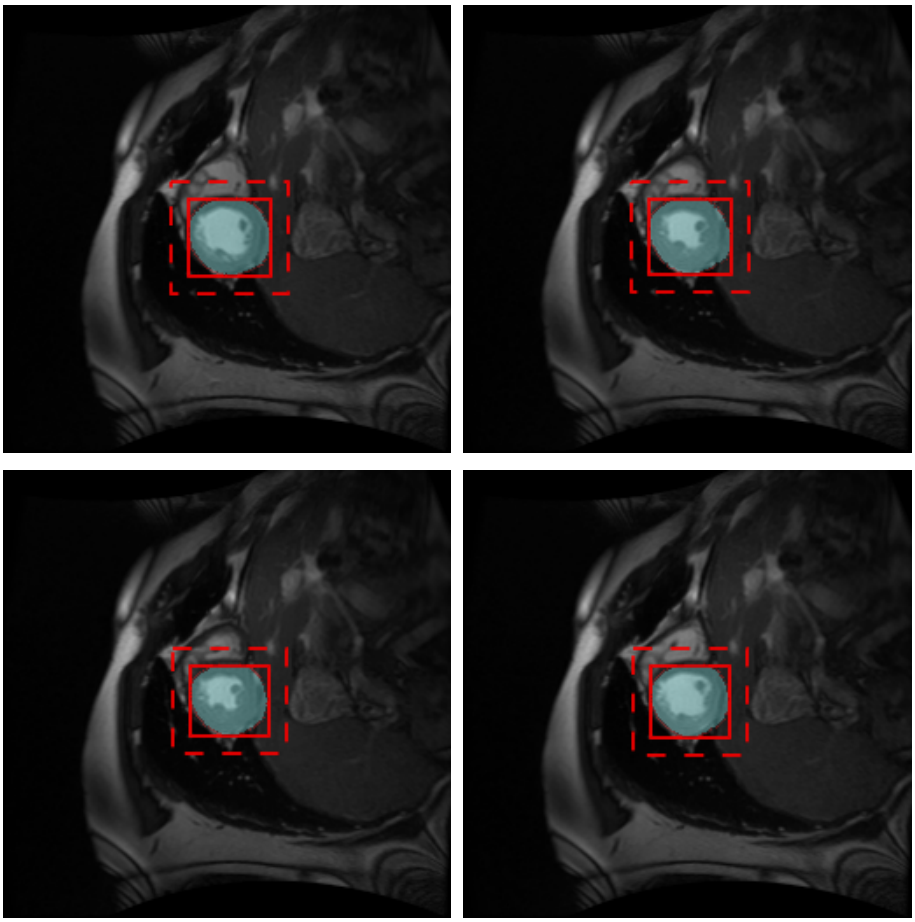
Statistical analysis Fig. 4.8 shows the box plots of the evaluation on different framework configurations for dice scores. Compared with others configurations, the segmentation results obtained by our method (configuration:i) have a small standard deviation, which shows that our method is more stable on region segmentation. Fig. 4.9 shows the box plots of the evaluation for 95HD. Compared with others configurations, based on the median quantile of box plots and the average of 56 patients,

TABLE 4.2: Ablation study on the proportionality coefficient λ of the hybrid loss

λ_1	λ_2	λ_3	DC/%	95HD	BDC
1	1	1	**87.93**	**1.826**	**0.306**
1	1	2	87.34	1.839	0.290
2	1	1	87.18	1.912	0.291
1	2	1	87.32	1.848	0.294
Average			87.28	1.866	0.292
1	2	2	87.20	1.876	0.285
2	1	2	87.13	1.879	0.287
2	2	1	87.04	1.891	0.288
Average			87.12	1.882	0.287
1	2	3	87.25	1.898	0.289
1	3	2	87.24	1.892	0.292
2	1	3	87.28	2.064	0.294
2	3	1	87.16	1.889	0.293
3	1	2	87.09	1.928	0.289
3	2	1	87.20	1.851	0.289
Average			87.20	1.920	0.291

most of the values of our method are low, which shows that our method optimizes the boundary quality.

Fig. 4.10 shows several localization and segmentation results of our A^0Net on LVQuan19. Fig. 4.10a indicates that we started with finding the smallest rectangular box for each slice of the patient's heart, ensuring that each box contained the segmentation object. Then we found the biggest rectangular box on the basis of these smallest rectangular boxes; and based on its shape, we cropped a new 3D volume from the original 3D volume as shown in the segmentation module of Fig. 4.1. Thanks to the localization results of Fig. 4.10a, we knew that the object was contained in/by the box, which greatly increased the proportion of objects in the image and reduced class imbalance. Fig. 4.10b compares ground truth and prediction, and we can see that the differences mainly are near the boundaries.



(a) Some localizations of the LV (in blue) of the 9^{th} patient. The red dotted box denotes that we extend next to the box by a size equal to 10 pixels to ensure that the whole LV is included into the bounding box.

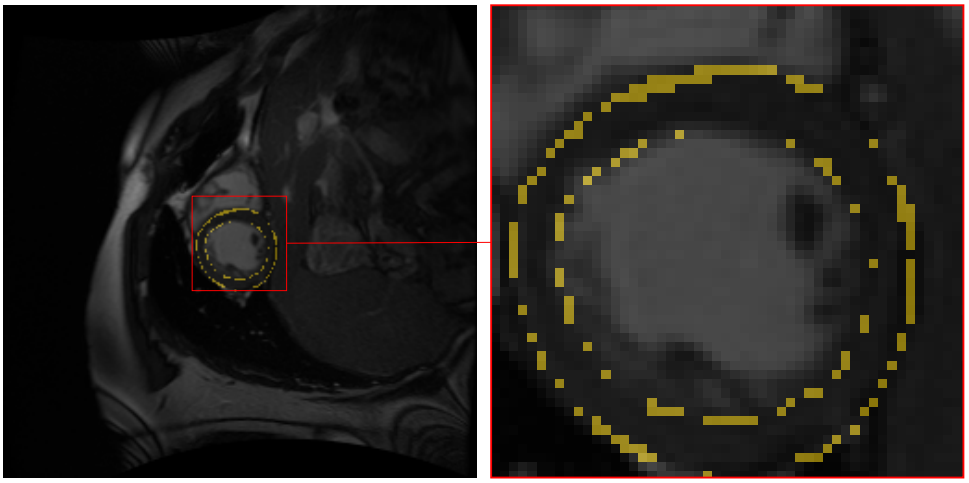
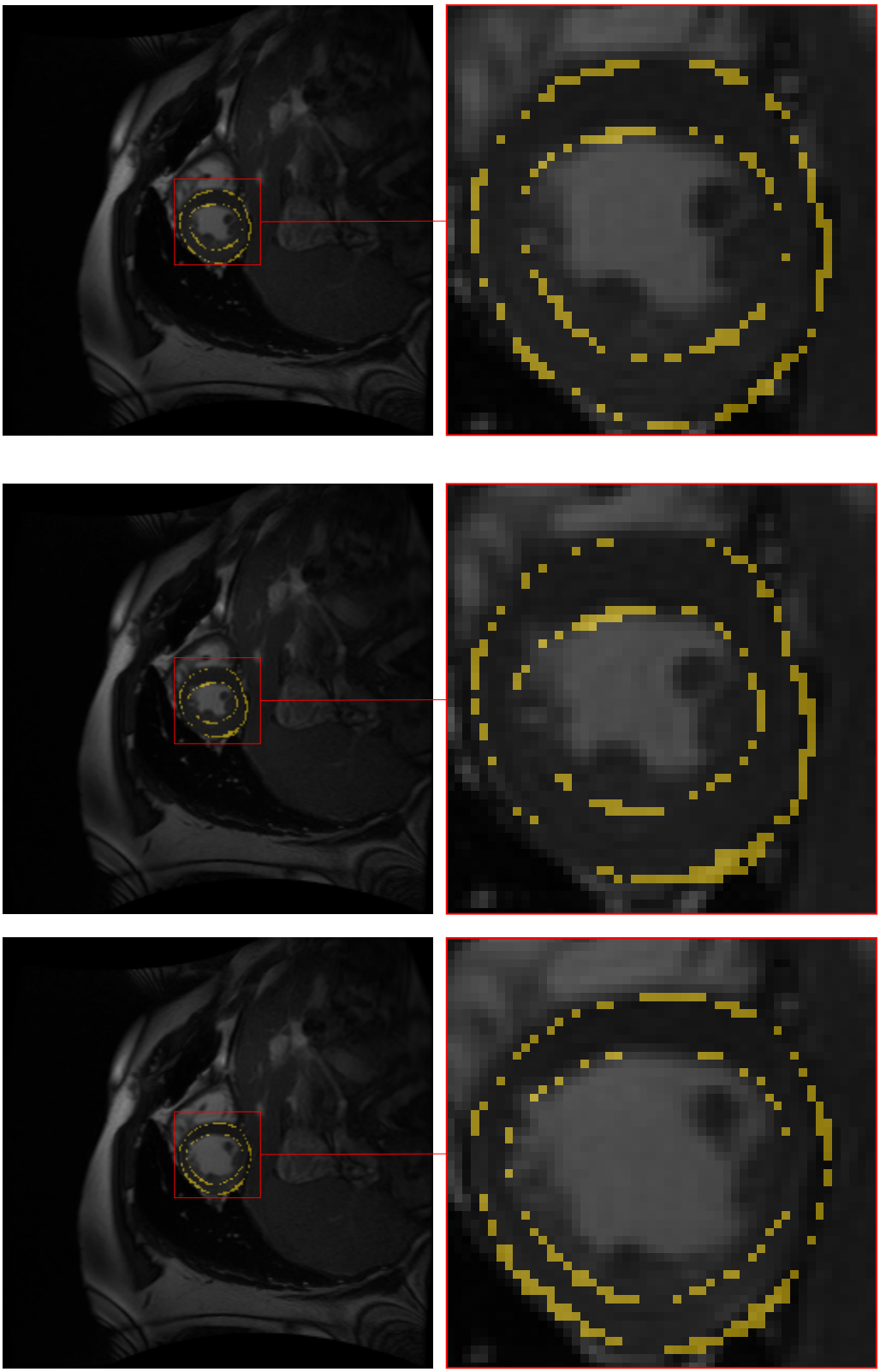


FIGURE 4.10: Localization and segmentation of our A^0Net on LVQuan19.



(b) Different comparisions between ground truth and prediction corresponding to (a); yellow denotes the difference.

FIGURE 4.10: Localization and segmentation of our A^0Net on LVQuan19.

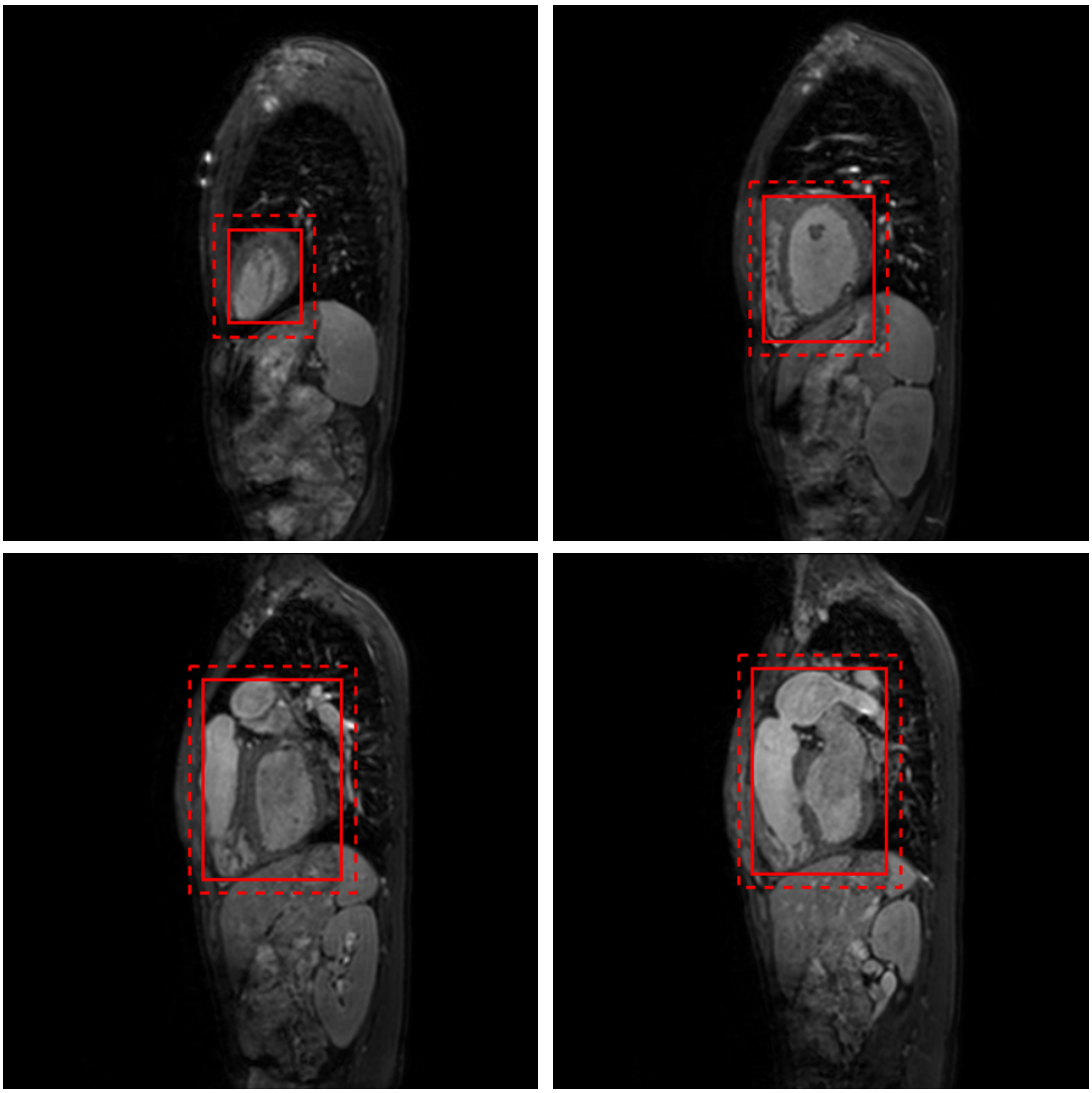
4.2.7 Comparison with State-of-the-Art Methods

TABLE 4.3: Comparison of the proposed method and other challengers on the HVSMR16 training dataset.

Method	DC of myocardium	Computation time	Datasets	Data augm.
A^0Net (our method)	0.826±0.038	< 2s	Full-volume	No
Best [127]	0.825±0.042	> 12s	Cropped axial	Yes
Second-best [128]	0.80±0.06	41.5±14.7s	Cropped axial	Yes
D.-S. 3D FCN [129]	0.726	2.5min	Cropped axial	Yes
Rahil [130]	0.69	-	Cropped axial	No
Maria [131]	0.74±0.09	-	Full-volume	No
3D U-Net [132]	0.694±0.076	-	Cropped axial	Yes
VoxResNet [133]	0.774±0.067	-	Cropped axial	Yes

Note: Red: best; Blue: second-best.

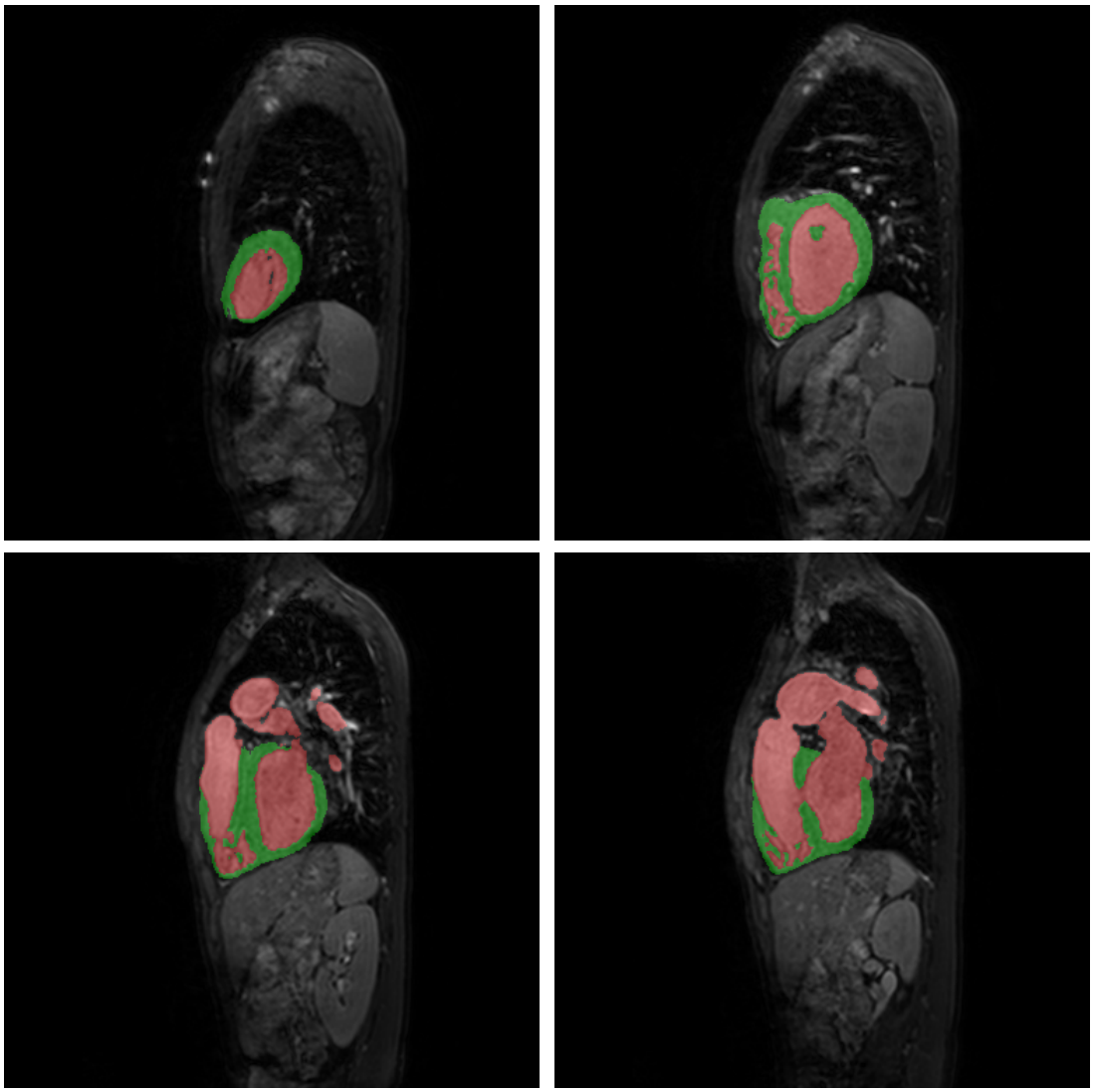
We compare the proposed method with other challengers on HVSMR16 training dataset with 10-fold-cross-validation. The best and second-best methods of HVSMR16 challenge both use only the cropped axial or cropped short axis reconstruction images rather than full-volume axial images as in the training dataset. The cropped axial images are equivalent to our localization results. To ensure that the whole segmented domain is included in our localization result, we enlarged the crop area by taking 10 supplementary pixels. Our segmentation results are obtained based on the full-volume axial images without data augmentation unlike the two first methods. As shown in Tbl. 4.3, our segmentation results on myocardium are better than the best method of the HVSMR16 challenge. For an entire 3D volume, our computational time for the entire pipeline is clearly less than other methods.



(a) Some localization results in one scan.

(b) 3D visualization of the segmentation results in one scan. Left: ground truth, Right: prediction.

FIGURE 4.11: Localization and segmentation of our A^0Net on HVSMR16. Green denotes the segmentation results of myocardium.

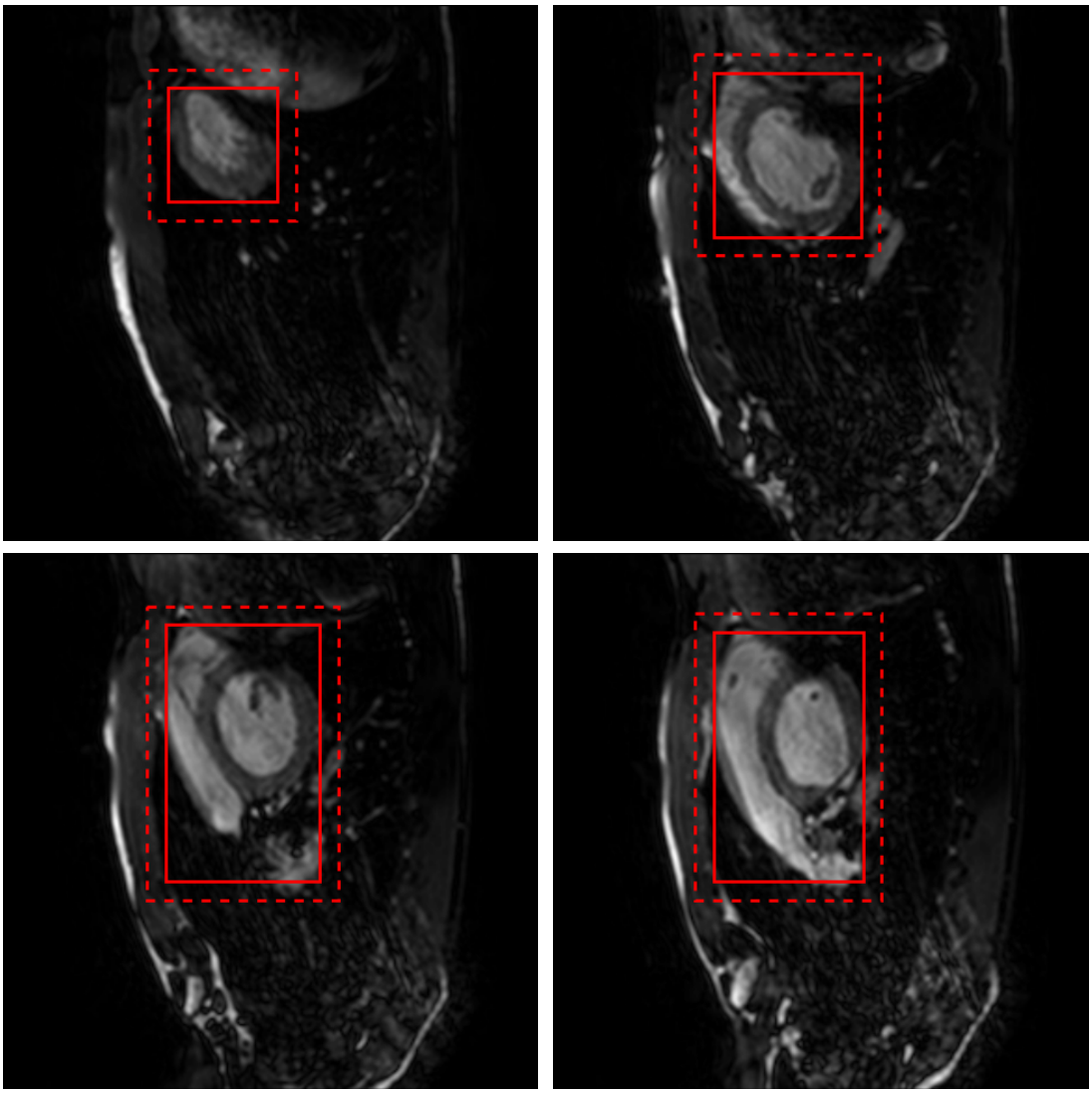


(c) Some segmentation results in one scan corresponding to (a).

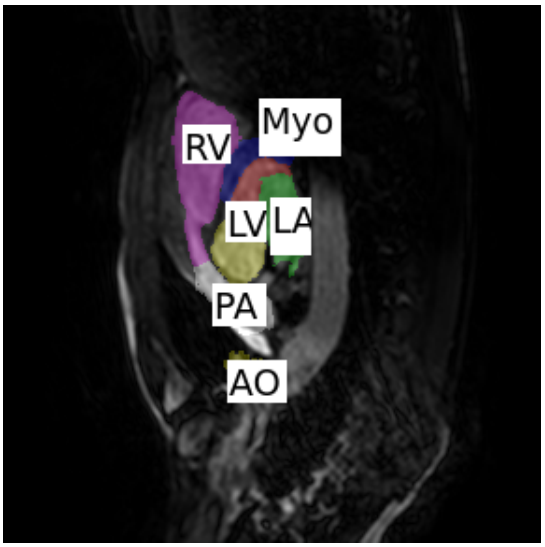
FIGURE 4.11: Localization and segmentation of our A^0Net on HVSMR16. Green denotes the segmentation results of myocardium.

TABLE 4.4: Comparison of our method and other challengers on the MM-WHS2017 MRI training dataset for segmenting the myocardium.

Method	DC (train)	DC (test)	Computation time	Data augmentation
Our (best)	0.851	-	$<$ 2s	No
Best [134]	0.796	0.781	$<$ 2min & $>$ 2s	No
Second-best [135]	0.752	0.778	-	Yes
UB2 [136]	-	0.811	-	Yes
3D U-Net [137]	0.720	0.791	-	Yes

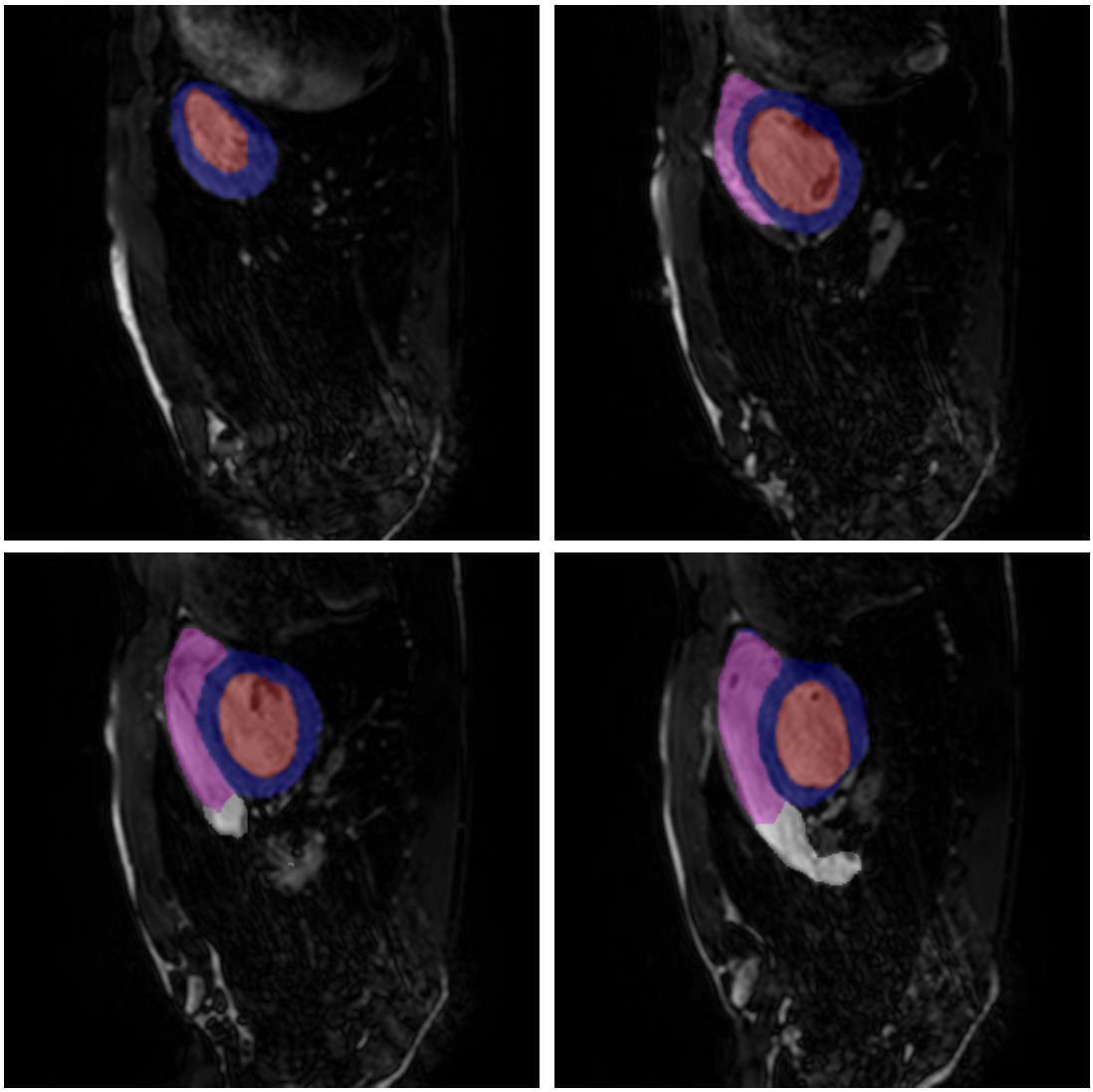


(a) Some localization results in one patient.



(b) Seven structures of the whole heart. Myo: myocardium, LA: left atrium, LV: left ventricle, RA: right atrium, RV: right ventricle, AO: ascending aorta, PA: pulmonary artery.

FIGURE 4.12: Localization and segmentation of our A^0Net on MM-WHS2017.



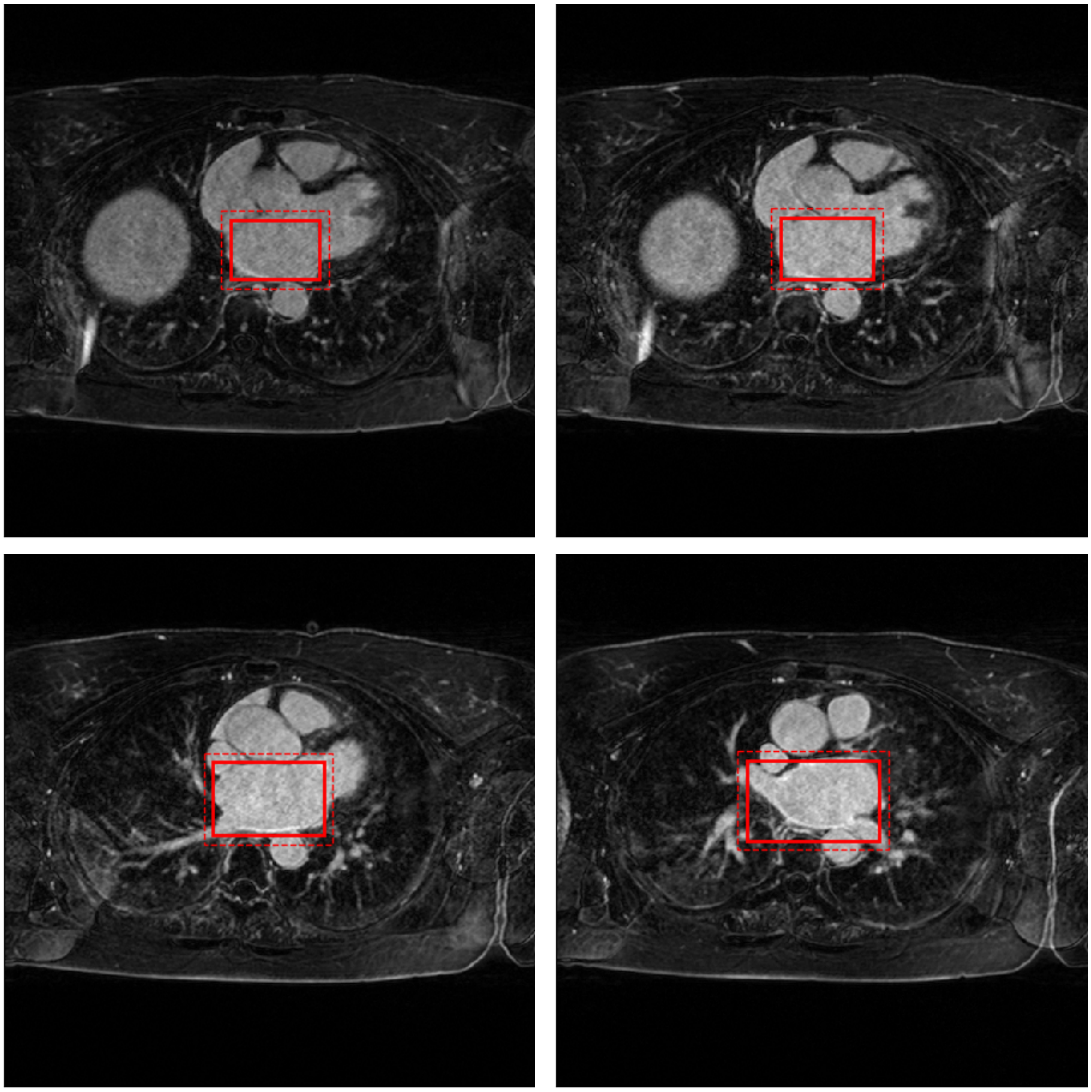
(c) Some segmentation results in one patient corresponding to (a).

FIGURE 4.12: Localization and segmentation of our A^0Net on MM-WHS2017.

We continued to test our method on the MM-WHS2017 challenge with 5-fold cross-validation and we obtained segmentation results for each class. As we focus on the myocardium segmentation, we will only present our results for this structure. For the comparison with state-of-the-art methods, we choose to compare our results with the results of the first and second prizes of the challenge, who respectively get dices of 0.87 and 0.863 in average for all classes. We reported their results on the training and on the testing sets. We also add a comparison with a late submission on the testing set only (scores on the training dataset are not available), having the best actual score of the challenge [136, 138]. As shown in Tbl. 4.4, compared with the first and second prizes of the MM-WHS2017 challenge, without using data augmentation, our method outperformed them for the segmentation of the myocardium of the left ventricle. Furthermore, our method needs less time to compute the prediction, which further validates the results in LVQuan19. Fig. 4.12 shows some localization

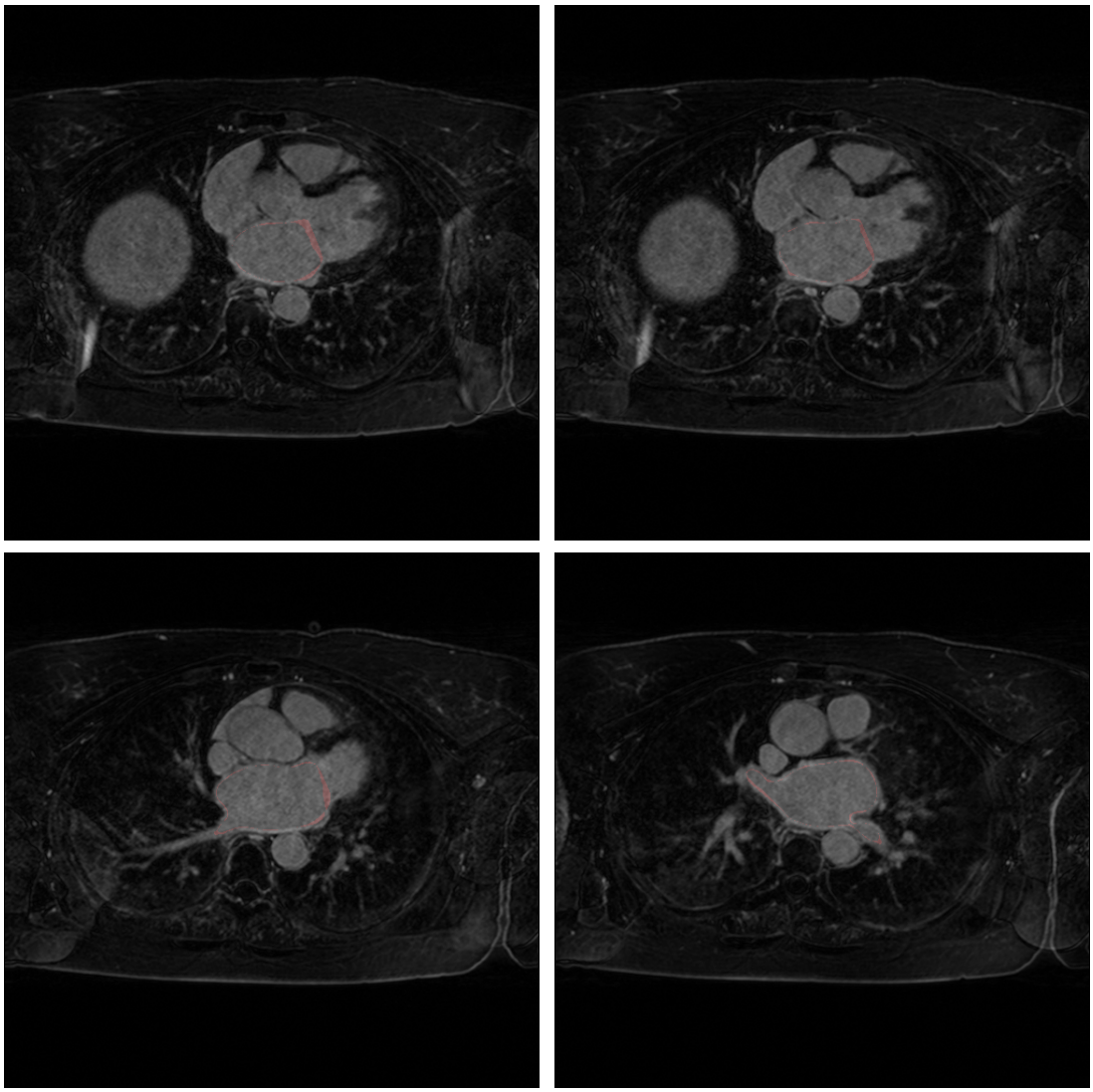
and segmentation results. Concerning the whole heart segmentation, the class imbalance causes a lot of damage without the localization module, because the seven structures of the heart do not always appear at the same time in a slice of the same 3D volume of a same patient. Without the GA module and **Block 2**, the network can confuse one class with another: the RA can be confused with the RV, the LV can be confused with the LA, and so on. Accordingly, a good segmentation requires to capture the global information by dilated convolutions and to enhance contrast using the GA module.

Finally, we tested our method on the AtriaSeg18 challenge with 5-fold cross-validation and we obtained segmentation results for left atrium (see Fig. 4.13). We got dice of 0.894 for localization and 0.904 for segmentation. Compared with the results of localization, the segmentation results only obtained a mean improvement of 1%. Therefore, our method is better on small dataset.



(a) Some localization results in one scan.

FIGURE 4.13: Localization and segmentation of our A^0Net on AtriaSeg18.



(b) Some segmentation results in one patient corresponding to (a). Red denotes the difference between ground truth and prediction

FIGURE 4.13: Localization and segmentation of our A^0Net on AtriaSeg18.

4.3 Conclusion

In this chapter, we propose a new single-minded attention network framework, A^0Net, and present a new hybrid loss for boundary-aware segmentation. A^0Net is able to prevent the interferences of surrounding similar tissues, while the hybrid loss guides it at several levels. Both generate a better capture not only of large-scale information but also of fine structures to produce segmentations with nice boundaries. The computation time of the entire pipeline is less than 2 seconds on Quadro P6000 GPU for an entire 3D volume and the proposed model size is about 122 MB, making it usable for clinical practice. However, the proposed two-stage segmentation method is not one end-to-end segmentation method, we need to train the localization network and the segmentation network separately. Otherwise, the localization

accuracy would affect the segmentation accuracy. Therefore, one end-to-end segmentation is required.

Chapter 5

End-to-end Segmentation Method

In the chapter 4, we have proposed two-stage method to segment heart, but it is not end-to-end segmentation method. In this chapter, we want to replace the localization of the chapter 4 with an attention module, in order to achieve end-to-end trainable segmentation method to obtain higher segmentation accuracy.

5.1 Methodology

5.1.1 Architecture of Network

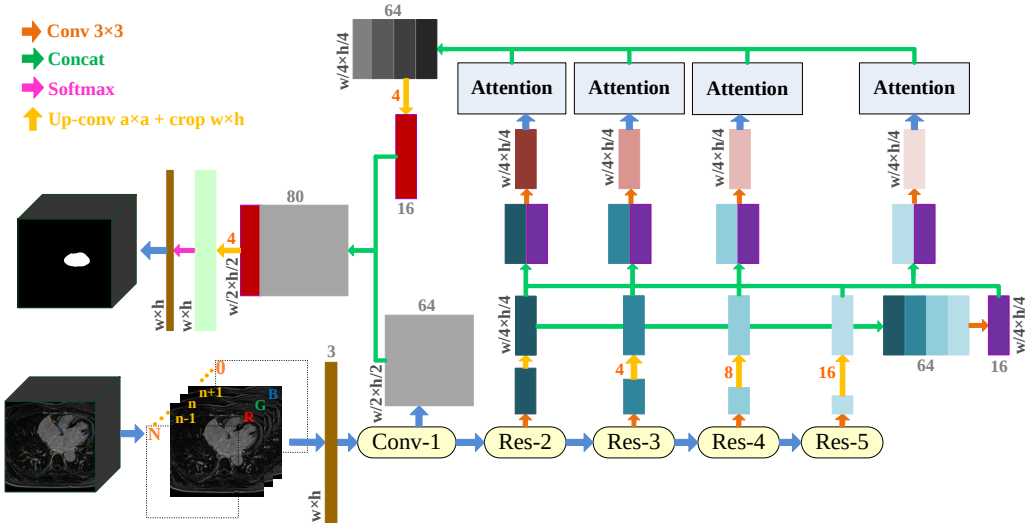


FIGURE 5.1: Architecture of our networks.

We propose a new attention network (see Fig. 5.1) using ResNet-101 pretrained on ImageNet [25] to compute feature maps. We discard its average pooling and fully connected layers, and keep only the sub-network made of one convolution-based and four residual-based "stages". Since the resolution decreases at each stage, we obtain a set of fine to coarse feature maps (with five levels of features).We add *specialized* convolutional layers (with a 3×3 kernel size) with K (*e.g.* $K = 16$) feature maps placed at the end of four residual-based "stages". They are concatenated together after up-convolutional layers. These last feature maps are combined with each of the outputs of the specialized layers, and then fed into the attention module

to generate the attention features. Finally, we concatenate the attention features with the outputs of *Conv1* and we fed them into the softmax layer.

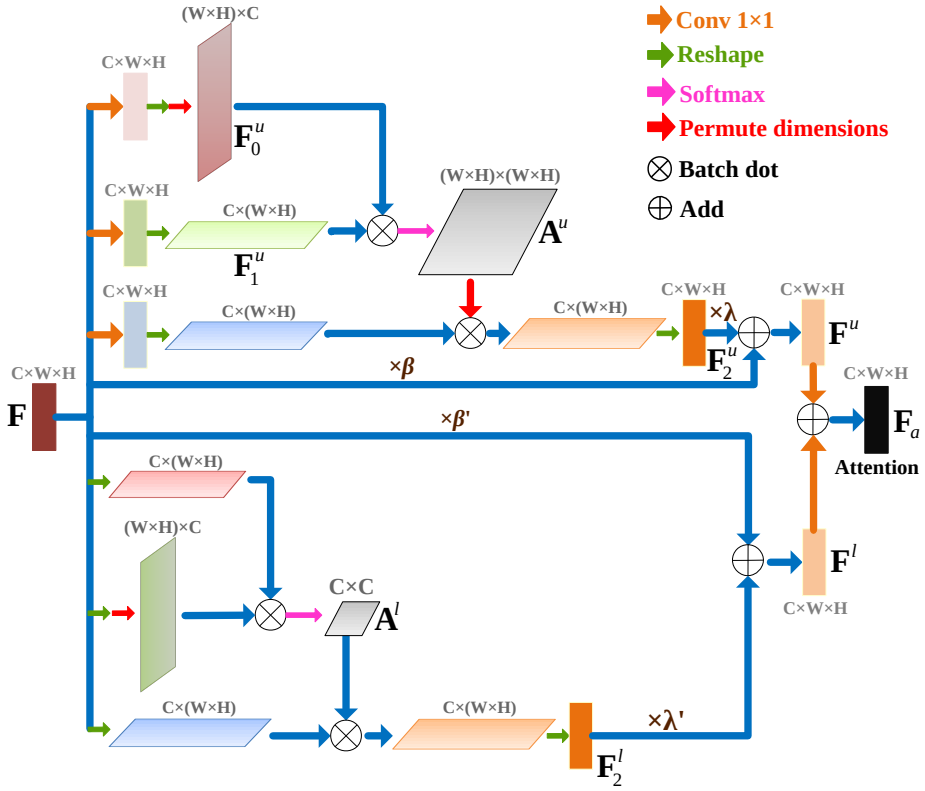


FIGURE 5.2: Attention Module. λ, λ', β and β' as hyperparameters, which is trained like the convolutional kernel. They decrease the weight of the unimportant feature maps.

5.1.2 Attention Module

As mentioned before, in a traditional segmentation model, the usual issue is that receptive fields are too small, which leads to poor contextual representations. Furthermore, the relationship between the different channels should be explored since each channel map represents one feature-specific response. Therefore, improving the dependencies among channel maps can lead to richer features. To solve these issues, we use an attention module inspired by [26]. As shown in Fig. 5.2, $\mathbf{F} \in \mathbb{R}^{C\times W\times H}$ acts as an input feature map for the attention module, where C, W, H are the channel, the width and the height of the feature map respectively. The upper branch $\mathbf{F}$ is fed into a convolutional, a Reshape and then a Transpose layers, resulting in a feature map $\mathbf{F}_0^u \in \mathbb{R}^{(W\times H)\times C}$. In the second branch (consider the order from top to bottom), the input feature map $\mathbf{F}$ follows the same operations minus the Transpose layer, resulting in $\mathbf{F}_1^u \in \mathbb{R}^{C\times(W\times H)}$. Then, the Multiply and the Softmax layers follow; they are applied on $\mathbf{F}_0^u$ and $\mathbf{F}_1^u$ to obtain the spatial attention map $\mathbf{A}^u \in \mathbb{R}^{(W\times H)\times(W\times H)}$. The input $\mathbf{F}$ is fed into a different convolutional layer in the third branch, and is then multiplied by $\mathbf{A}^u$ fed into the Transpose layer, resulting in $\mathbf{F}_2^u$. Therefore the output $\mathbf{F}^u$ of the upper branch can be formulated as follows: $\mathbf{F}^u = \lambda \times \mathbf{F}_2^u + \beta \times \mathbf{F}$, where

$\lambda \in \mathbb{R}^C$ is initialized to [0,..,0], and $\beta \in \mathbb{R}^C$ is initialized to [1,..,1]. The values λ and β are used to gradually learn the importance of the spatial attention map.

In the lower branch, the attention module mainly focuses on the most important channels. The channel attention map $\mathbf{A}^l$ can be obtained by different combinations of convolutional, Reshape and Transpose layers as shown at the bottom of Fig. 5.2. Finally, the output $\mathbf{F}^l$ of the lowest branch can be defined as follows: $\mathbf{F}^l = \lambda' \times \mathbf{F}_2^l + \beta' \times \mathbf{F}$, where $\lambda' \in \mathbb{R}^C$ is initialized to [0,..,0], and $\beta' \in \mathbb{R}^C$ is initialized to [1,..,1]. The feature map $\mathbf{F}_2^l$ denotes the results of the product of the input $\mathbf{F}$ with $\mathbf{A}^l$ fed into a convolutional passing through the transpose block. Therefore, the attention feature map $\mathbf{F}_a$ is defined as:

$$\mathbf{F}_a = Conv\left(\mathbf{F}^u\right) + Conv\left(\mathbf{F}^l\right). \tag{5.1}$$

Compared with [26], our proposed attention module is different with it. Firstly,the final outputs of Position Attention Module (PAM) and Channel Attention Module (CAM) are differently defined. In our method, as shown in Fig. 5.2, the final output of PAM is defined as: $\mathbf{F}^u = \lambda \times \mathbf{F}_2^u + \beta \times \mathbf{F}$, where $\lambda \in \mathbb{R}^C$ is initialized to [0,..,0], and $\beta \in \mathbb{R}^C$ is initialized to [1,..,1]. The values λ and β are used to gradually learn a weight during the training process, therefore $\mathbf{F}_2^u$ and $\mathbf{F}$ both are assigned more weight for important feature maps, which highlights more important features. However, in [26], the final output of PAM is only defined as: $\mathbf{F}^u = \lambda \times \mathbf{F}_2^u + \mathbf{F}$, where λ is initialized to [0,..,0], if only considering to assign more weight to $\mathbf{F}_2^u$, ignoring the effect of $\mathbf{F}$ and assigning same weight to $\mathbf{F}$, the redundant information of $\mathbf{F}$ will be transfered directly the output of PAM, which will have an diminished effect on attention. CAM is also like PAM that the final output is differently defined. Therefore, the improved attention module (our attention module) pays more attention to the important feature. Secondly, for the CAM in [26], employing convolution layers before the input of CAM, which leads to that the relationship between different channel maps has been destroyed in advance, but we do not employ convolution layers to embed features before computing relationships of two channels in our CAM module, which can maintain relationship between different channel maps. Finally, [26] only is used in the output of network, not adopted cascading operation because the feature map of huge shape (H×W)×(H×W) in the PAM needs to huge GPU memory. However, the higher-level feature maps as the input of attention module will lose more detailed information of targets. Therefore, our network applies the improved attention module to different cascades, which not only reduces the redundant use of information, but also makes full use of different levels feature maps.

5.1.3 Hybrid Loss

The hybrid loss consists of two parts: region loss and boundary one. It is defined as: $\ell_{\mathrm{H}} = \ell_{\mathrm{R}} + \ell_{\mathrm{B}}$, where ℓ_{R} denotes the region loss and ℓ_{B} denotes the boundary loss. The region loss is same with the hybrid loss in section 4.1.4 of chapter 4. Based on

the region loss, we add the boundary loss into the hybrid loss, which can optimize the segmentation result. They are explained hereafter.

5.1.3.1 Region Loss

To obtain high quality regional segmentation, we define ℓ_{R} as a region loss: $\ell_{\mathrm{R}} = \ell_{\mathrm{CCE}} + \ell_{\mathrm{SSIM}} + \ell_{\mathrm{DC}}$, where ℓ_{CCE}, ℓ_{SSIM} and ℓ_{DC} denote Categorical Cross Entropy (CCE) loss [22], Structural Similarity (SSIM) loss [23] and Dice Coefficient (DC) loss [24] respectively.

CCE [22] loss is commonly used for multi-class classification and segmentation. It is defined as

$$\ell_{\mathrm{CCE}} = -\sum_{i=1}^{C}\sum_{a=1}^{H}\sum_{b=1}^{W} y^{i}_{(a,b)} \ln {y_*}^{i}_{(a,b)}, \tag{5.2}$$

where C is the number of classes of each image, H and W are the height and width of image, $y^{i}_{(a,b)} \in \{0,1\}$ is the ground truth one-hot label of class i at position (a,b) and ${y_*}^{i}_{(a,b)}$ is the predicted probability that (a,b) belongs to class i.

SSIM loss can assess image quality [23], and can be used to capture the structural information, which will decrease the mis-segmentation rate of surrounding similar tissues. Therefore, we integrated it into our training loss to learn the differences between the segmented domain and similar tissues around the segmented domain. Let $\mathbf{S}$ and $\mathbf{G}$ be the predicted probability map and the ground truth mask respectively, the SSIM loss function of $\mathbf{S}$ and $\mathbf{G}$ is defined as

$$\ell_{\mathrm{SSIM}} = 1 - \frac{(2\mu_{\mathrm{S}}\mu_{\mathrm{G}} + \varepsilon_1)(2\sigma_{\mathrm{SG}} + \varepsilon_2)}{(\mu_{\mathrm{S}}^2 + \mu_{\mathrm{G}}^2 + \varepsilon_1)(\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{G}}^2 + \varepsilon_2)} \tag{5.3}$$

where μ_{S}, μ_{G} and σ_{S}, σ_{G} are the means and standard deviations of $\mathbf{S}$ and $\mathbf{G}$ respectively, σ_{SG} is their covariance, $\varepsilon_1 = 0.01^2$ and $\varepsilon_2 = 0.03^2$ are used to avoid a division by zero.

DC [24] loss is used to measure the similarity between two sets as defined in Eq. 2.36. But for the multi-class segmentation task, Eq. 2.36 is not suitable due to the class imbalance problem in such cases. Therefore, we extend the definition of the DC loss to multiclass segmentation in the following manner:

$$dice_i = \left(\epsilon + 2\sum_{n=1}^{N_i} y^{i}_{n}\, {y_*}^{i}_{n}\right) / \left(\epsilon + \sum_{n=1}^{N_i}\left(y^{i}_{n} + {y_*}^{i}_{n}\right)\right) \tag{5.4}$$

$$\ell_{\mathrm{DC}} = 1 - \sum_{i=1}^{C} dice_i / \left(N_i + \epsilon\right), \tag{5.5}$$

where N_i denotes the numbers of class i and $\epsilon > 0$ is a smooth factor.

5.1.3.2 Boundary Loss

The loss functions mentioned before are mainly for region segmentation, so we propose a multi-class boundary loss function based on Kervadec's distance [27] to be able to refine the segmentations. As shown in Fig. 5.3, ΔA denotes the difference between the boundary $\mathbf{G}^{i}_{\mathbf{B}}$ of the ground truth of class i and the boundary $\mathbf{S}^{i}_{\mathbf{B}}$ of the

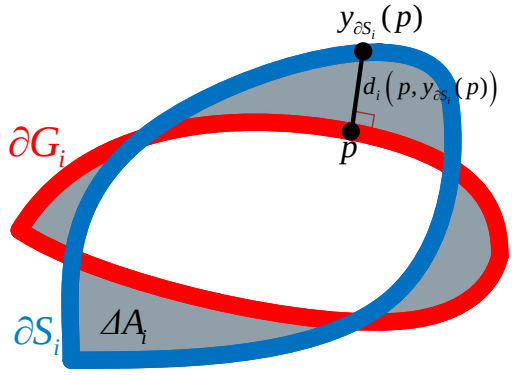


FIGURE 5.3: Illustration of calculating boundary loss

prediction of class i. When ΔA tends to zero, it means that the segmentation results are becoming better around the boundaries. Therefore, for a given class i, when the prediction and the ground truth are close enough, which is easily obtained thanks to our regional loss, minimizing the difference between their boundaries can be obtained by minimizing Kervadec's distance [27]:

$$\ell_{\mathrm{B}}^{i} = \int_{\partial G_i} \left\| y_{\partial S_i}(p) - p \right\|^2 \mathrm{d}p \tag{5.6}$$

where ∂G_i and ∂S_i denotes the boundaries of $\mathbf{G}_{\mathbf{B}}^{i}$ and (binarized) $\mathbf{S}_{\mathbf{B}}^{i}$ and $\|\cdot\|$ denotes the L2 norm. When p is a point in ∂G_i, $y_{\partial S_i}(p)$ denotes the corresponding point on boundary ∂S_i along the direction normal to ∂G_i (see Fig. 5.3). It can be shown [27] that minimizing ℓ_{B}^{i} is equivalent to minimize the area of the surface $\Delta A_i = (\mathbf{G}_{\mathbf{B}}^{i} \backslash \mathbf{S}_{\mathbf{B}}^{i}) \bigcup (\mathbf{S}_{\mathbf{B}}^{i} \backslash \mathbf{G}_{\mathbf{B}}^{i})$ (see Fig. 5.3). Thus, our multi-class boundary loss naturally follows:

$$\ell_{\mathrm{B}} = \sum_{i=1}^{C} \int_{\partial G_i} \left\| y_{\partial S_i}(p) - p \right\|^2 \mathrm{d}p \tag{5.7}$$

Fig. 4.7 shows the prediction results with different loss functions. Fig. 4.7(image) and Fig. 4.7(GT) are the input image and its corresponding ground truth. Fortunately, after several iterations of the network, segmentation results can be obtained based on a single loss function such as CCE, SSIM and DC loss. However, their segmentation results all have wrong segmentation connected to the region of ground truth as shown in Fig. 4.7(c), 4.7(d), 4.7(e), which can not be removed in post-processing.

According to Eq. 5.2, the CCE loss is calculated on a pixel-by-pixel basis (pixel-wise level), therefore, it does not consider using the information of surrounding pixels. Although this helps to ensure the convergence of all pixels and obtain a relatively good local optimum, the loss function will choose to give one neutral prediction probability such as 0.5 at boundaries of the target in order to avoid large losses, which often leads to ambiguities in the boundary. As shown in Fig. 4.7(c), compared with Fig. 4.7(d) and Fig. 4.7(e), the segmentation results are fine structures, which

also verifies that CCE loss makes all pixels converge.

When SSIM loss is used as the loss function of networks, a 7×7 sliding window is used on the image and its corresponding ground truth, and then the 7×7 image patch is taken out to calculate the SSIM loss, so the SSIM loss is computed based on the patch-level. The SSIM loss makes up for the lack of CCE loss and fully considers the surrounding information of each pixel. The SSIM loss assigns higher weights to the pixels in the transition area between each class, so even if the prediction probability of each class is the same at the boundary, the loss around the boundary will be higher. However, if the 7×7 image patch belongs to the background region, the μ_G, σ_G and σ_{SG} will be equal to zero, Eq. 5.3 is simplified as

$$\ell_{\text{SSIM}}^{\text{background}} = 1 - \frac{\varepsilon_1 \varepsilon_2}{(\mu_S^2 + \varepsilon_1)(\sigma_S^2 + \varepsilon_2)} \tag{5.8}$$

When the predicted probability map $\mathbf{S}$ is very close to the ground truth (background), $\ell_{\text{SSIM}}^{\text{background}}$ will be dropped sharply from 1 to 0, and then it does not contribute to the training, so the network can neglect the background accuracy in the beginning phase of the training process, which is very important for medical images with a large number of background areas. As shown in Fig. 4.7(d), there is also one problem if using the SSIM loss as the loss function of network, the network incorrectly predicts a small part of the segmentation results belonging to the background region, but this part is not connected to the ground truth. This phenomenon can be well explained, because during the training process, the SIMM loss of each class continues to decrease until its fluctuation range is minimal. At this time, the background loss that is ignored from the beginning of the training becomes the dominant, therefore, the network is easy to predict the wrong segmentation belonging to the background region.

The calculation method of DC is different between the training phase and the evaluation phase. For the evaluation phase, it is calculated based on 3D volume. However, in the training phase, its calculation is slice-by-slice for medical images. Because y_n^i is ground truth one-hot label of class i and ${y_*}_n^i$ is corresponding predicted probability map. $y_n^i {y_*}_n^i$ denotes their difference map, which means calculating the difference from a global perspective, so the DC loss is computed based on the map-level. But it can be seen from Eq. 5.4 that the class that occupies a large area plays a leading role for the loss, which is not good for medical images with a lot of background. Therefore, Eq. 5.5 is used to redistribute the weight of the loss of each class, and finally make the segmentation result of each class more uniform.

The three loss functions mentioned above are all region-based. But the boundary loss function is defined at Eq. 5.7, which takes the form of a distance metric on the space of contours, not regions. For the problem of class imbalance, we can reduce the problems related to region loss through boundary loss. As shown from Eq. 5.7, the boundary loss mainly uses the integral on the difference area between ground truth and the prediction, which also supplements the information for the region loss.

Therefore, we use the respective advantages of the above four loss functions and combine them to propose a new hybrid loss function. CCE loss pays attention to the reasonable classification of all pixels. SSIM loss compensates for the ambiguity of CCE loss at the boundary and gives the boundary a larger loss value. DC stands in the overall perspective to guide the correctness of the general direction. As shown in Fig. 4.7(h), through this region loss function, the wrong segmentation part is no longer connected to the ground truth. Finally, the boundary loss refines details.

5.2 Experimental Results

5.2.1 Dataset Description

We evaluate our method on the MICCAI 2018 Atrial Segmentation Challenge [1] (AtriaSeg18). Its aim is to segment the left atrium. It contains 100 annotated 3D MRIs from patients with atrial fibrillation. The pixel spacing of the MR images is 0.625 $\times$ 0.625 $\times$ 0.625 mm/pixel. The dataset includes two different image sizes: $88\times576\times576$ and $88\times640\times640$.

5.2.2 Preprocessing

Fig. 5.5(a) shows the histograms of the original volumes have various shapes, according to their histograms, after cropping each slice to 346×346 pixels as shown in Fig. 5.4(c), we map the gray-level range to [0,255] by histogram equalization. Fig. 5.5(b) shows the gray-level scale of each volume after histogram equalization based on the 1/4 of the pixels in red region as shown in Fig. 5.4(a). The pre-processing begins with a Gaussian normalization. Because ResNet-101 network's input is an RGB image, we propose to take advantage of the 3D information by stacking 3 successive 2D frames: to segment the n^{th} slice, we use the n^{th} slice of the MR volume, and its neighboring $(n-1)^{th}$ and $(n+1)^{th}$ slices, as green, red and blue channels, respectively. This new image, named "3D-Like" image, enhances the boundaries of objects, as shown in Fig. 5.4.

5.2.3 Postprocessing

We crop the initial volume of size $88\times W\times H$ into an image of size $88\times w\times h$ (where W and H are the initial width and height of a slice). We keep only the greatest connected component of the output segmentation and pad with zeros to get back a $T\times W\times H$ image.

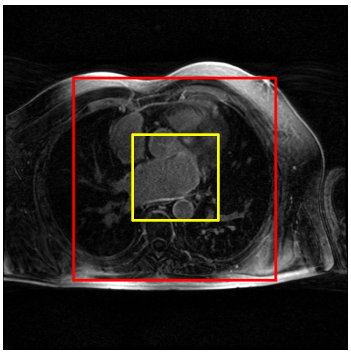
[1] http://atriaseg2018.cardiacatlas.org/

5.2.4 Implementation and Experimental Setup

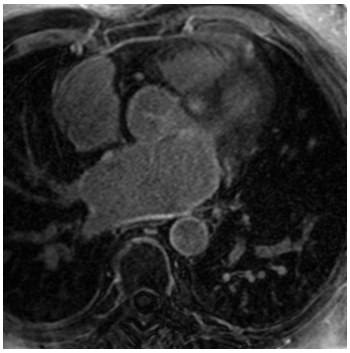
We implemented our experiments on Keras/TensorFlow using a NVidia Quadro P6000 GPU. We used the hybrid loss function, softmax to get a probability distribution over classes, Adam optimizer (batchsize = 3, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 0.001$, lr $= 0.01$) and did not use learning rate decay. We trained the network during 30 epochs.

5.2.5 Evaluation Methods

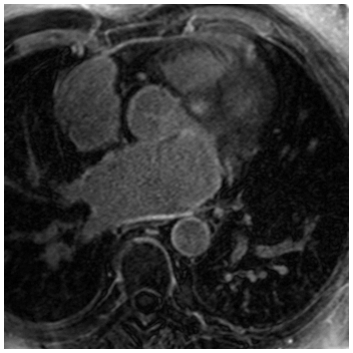
Three metrics are used to evaluate our method: dice to evaluate the regions, and 95% Hausdorff distance (95HD) and Average Hausdorff distance (AHD) to quantitatively evaluate the boundaries.



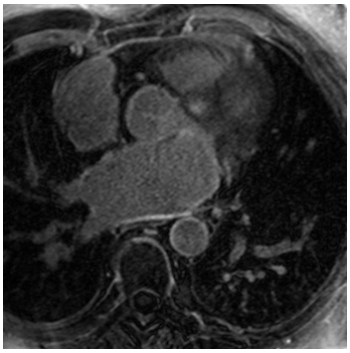
(a) Initial image



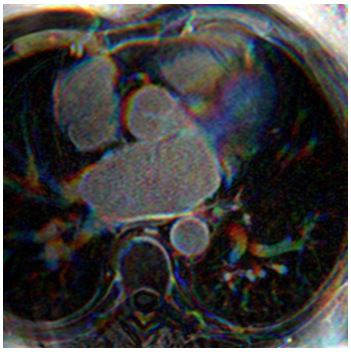
(b) $n-1$



(c) n

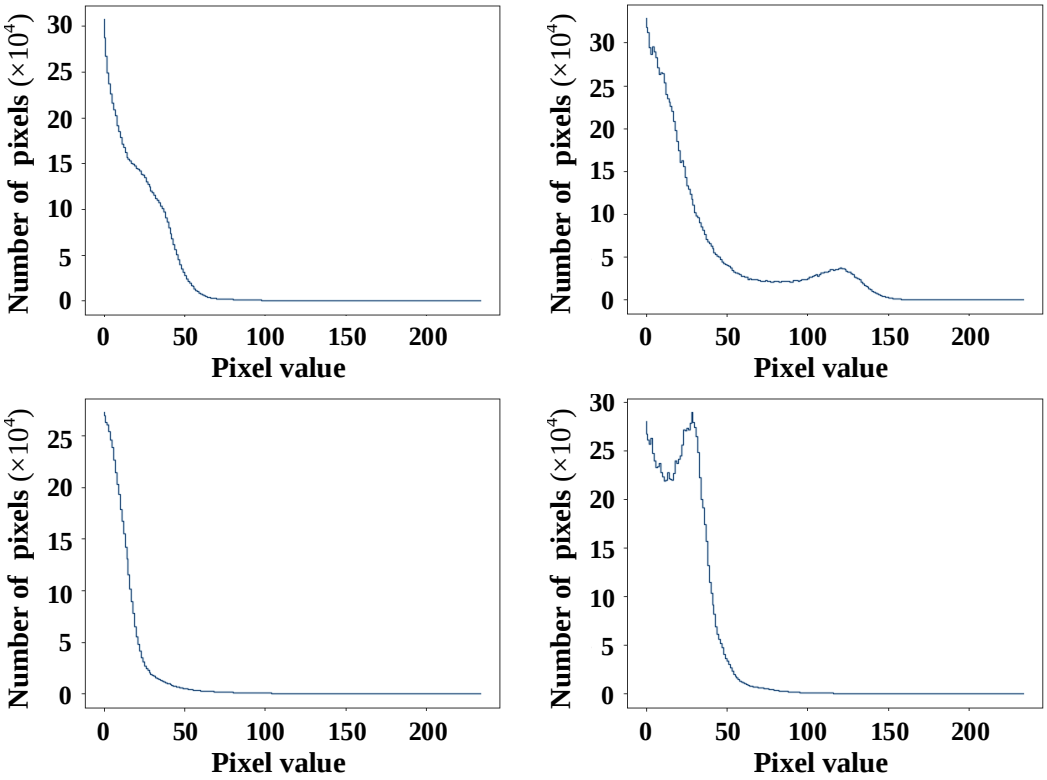


(d) $n+1$

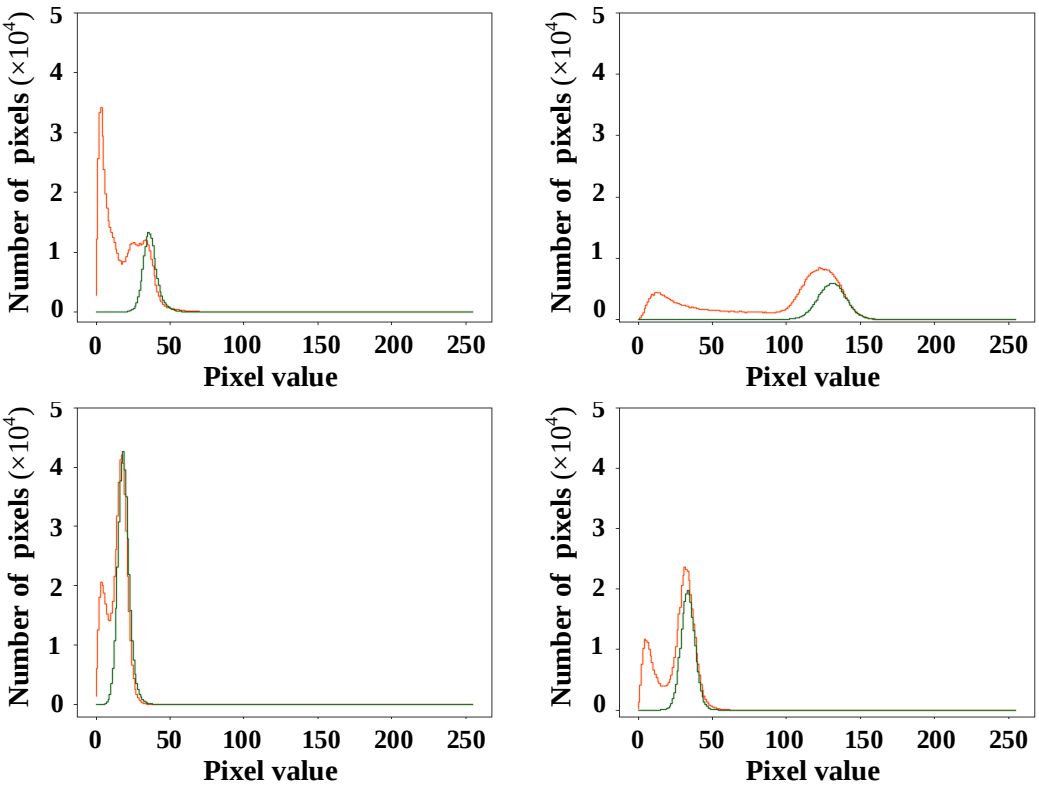


(e) "3D-Like"

FIGURE 5.4: Illustration of our "3D-Like" procedure. The red box depicts the boundary of the cropped input image. Three successive cropped slices (b-d) are used to build a "3D-Like" image (e).



(a) Whole volume histograms



(b) Partial volume histograms

FIGURE 5.5: (a): The histograms of the original volumes have various shapes; (b): to normalize the gray-level scale of each volume, we consider the histogram of their central sub-volume (in orange; see also Fig. 5.4(a)), which has the same dynamic than the oneof the left atrial region given by the ground-truth (in green).

5.2.6 Comparison with State-of-the-arts Methods

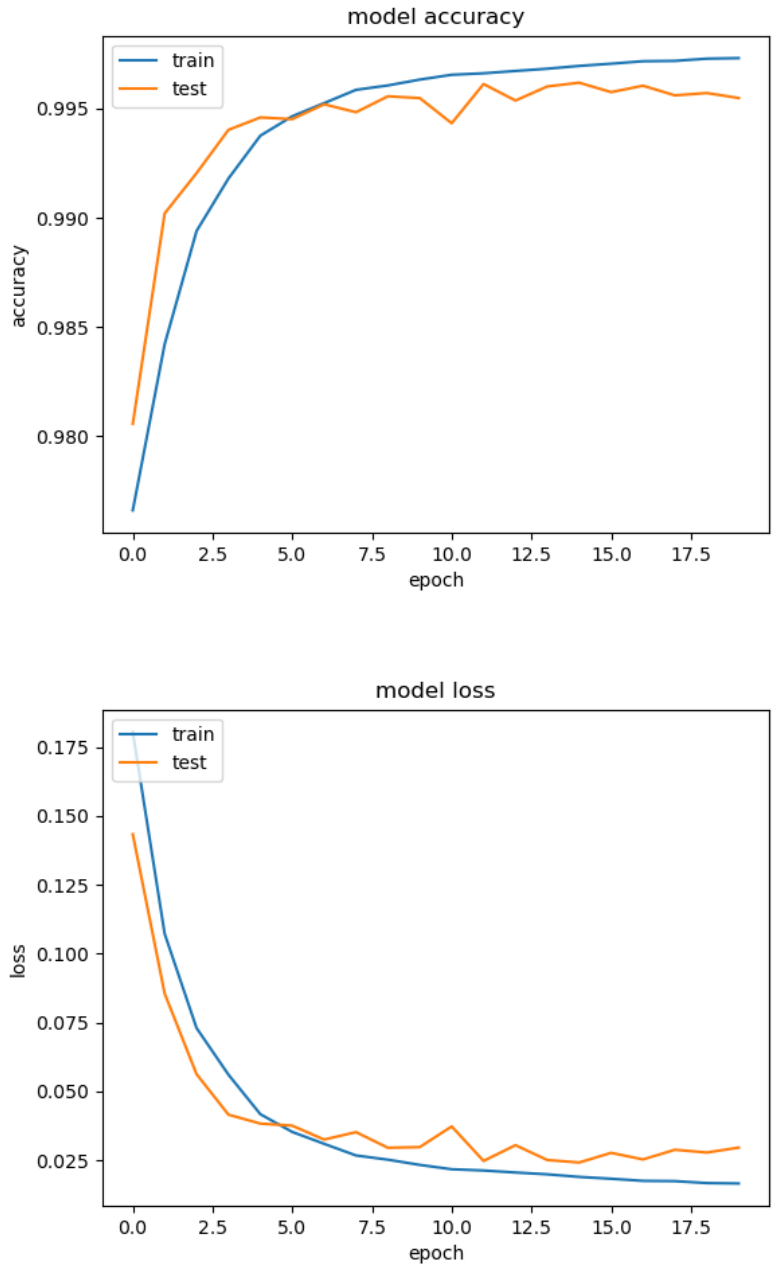
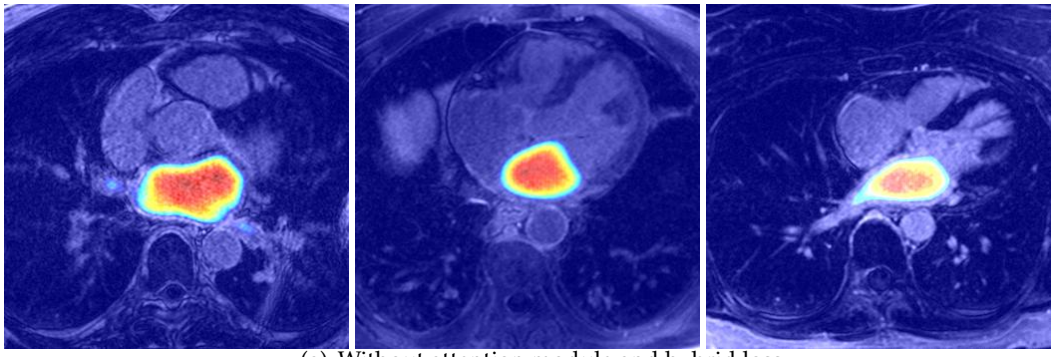


FIGURE 5.6: Evolution of the loss and accuracy with the number of epochs.

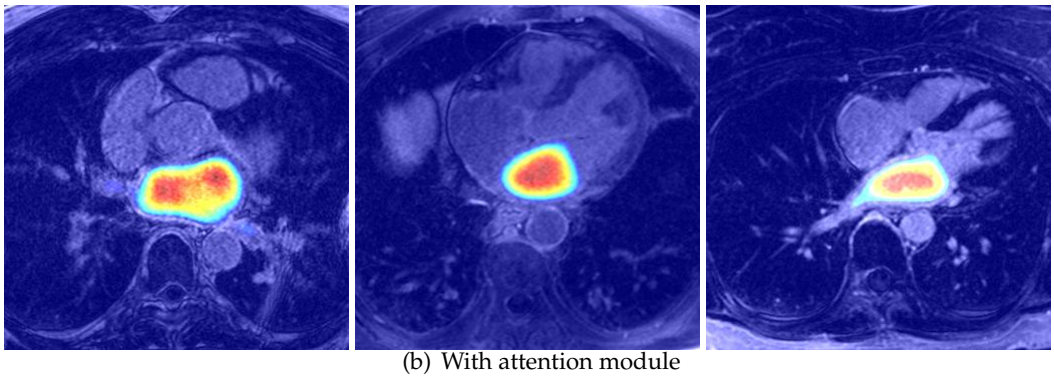
Fig. 5.6 shows the evolution of the loss and accuracy with the number of epochs. For the model accuracy, when the epoch reaches the fifth epoch, the training accuracy of network have arrived 99%. For medical images, there is a lot of redundant information in the image, so the accuracy can be higher in a short time. For the model loss, the loss is drop sharply before fifth epoch, and there is little fluctuation around 0.025 after that.

TABLE 5.1: Comparison of our method and other state-of-the-art architectures using a 5 fold cross-validation.

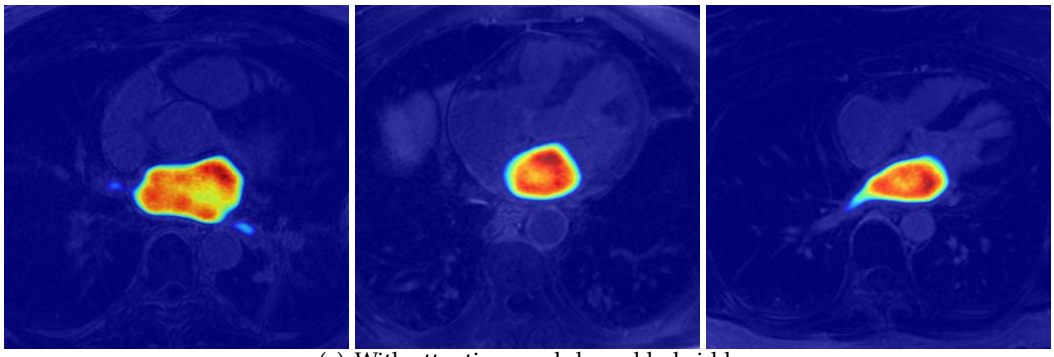
Method	Att. Module	Hyb. Loss	DC/%	95HD/mm	AHD/mm
U-Net [53]			88.556(±2.586)	4.447(±0.996)	0.212(±0.077)
		✔	89.613(±2.257)	4.169(±0.960)	0.210(±0.118)
DANet [26]			84.229(±3.774)	6.145(±2.341)	0.514(±0.477)
		✔	87.584(±2.765)	4.903(±1.448)	0.280(±0.179)
Deeplabv3+ [139]			85.444(±3.079)	5.872(±2.345)	0.504(±0.614)
		✔	87.556(±1.155)	5.210(±1.087)	0.273(±0.074)
Our Method			90.774(±1.568)	3.312(±1.277)	0.158(±0.092)
	✔		91.326(±1.174)	3.097(±0.810)	0.143(±0.055)
	✔	✔	**91.792(±1.065)**	**2.868(±0.667)**	**0.130(±0.042)**



(a) Without attention module and hybrid loss

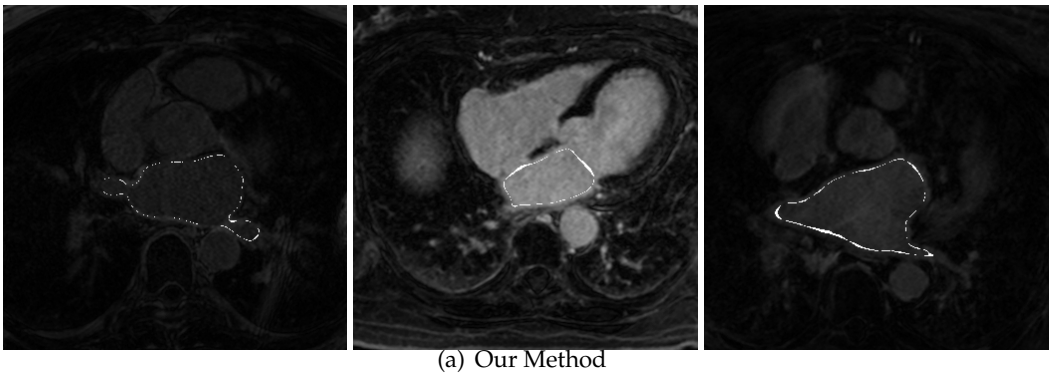


(b) With attention module

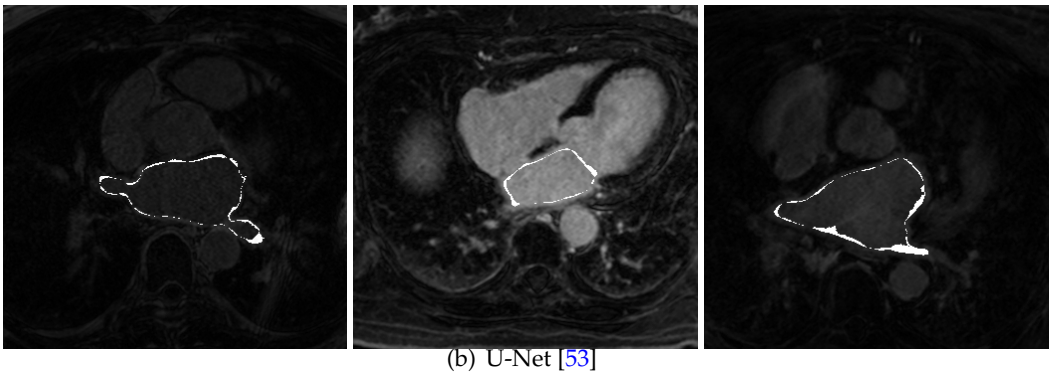


(c) With attention module and hybrid loss

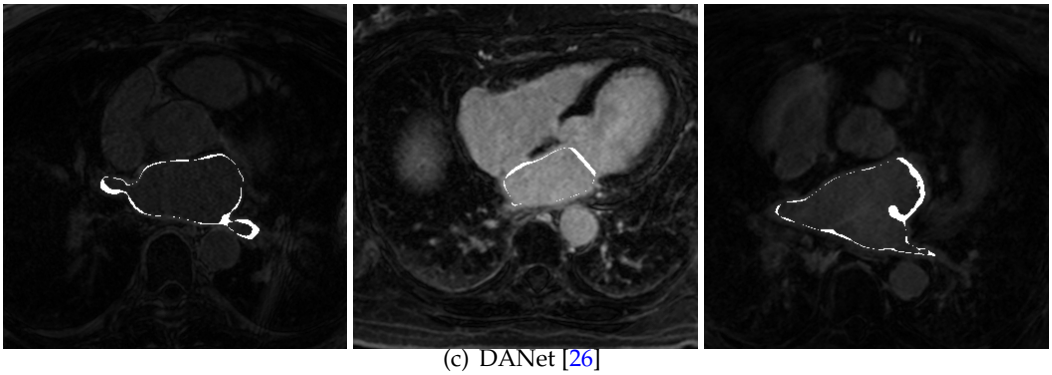
FIGURE 5.7: Ablation study for our method; red color denotes highest weight



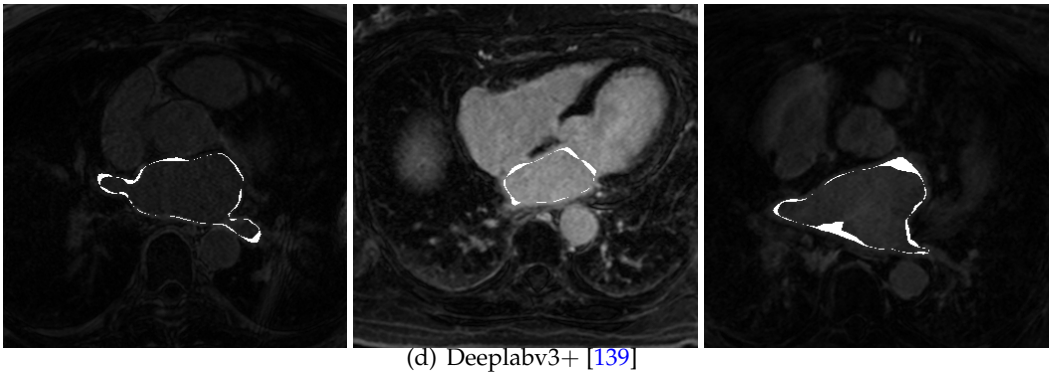
(a) Our Method



(b) U-Net [53]



(c) DANet [26]



(d) Deeplabv3+ [139]

FIGURE 5.8: Comparison of the proposed method and other state-of-the-art architectures. The white pixels are the differences between the prediction and the GT.

The experimental results obtained by several state-of-the-art segmentation networks are reported in Table 5.1. Compared to other networks proposed in the context of medical image segmentation ,i.e., U-Net [53], DANet [26] and Deeplabv3+ [139],

our network achieves a mean improvement of 3.236%, 7.563% and 6.348% (in terms of DC), 1.579 mm, 3.277 mm and 3.004 mm (on 95HD) and 0.082 mm, 0.384 mm and 0.374 mm (on AHD), respectively. For the proposed method, the improved performance could be explained by the fact that the attention module and hybrid loss. Fig. 5.7 shows the ablation study for our method. Compared Fig. 5.7(a) with Fig. 5.7(b), the attention module imitates the human visual system to decrease the impact of surrounding similar structures and obtain more detail informations about left atrial (LA). It increases segmentation performance by 0.552% (DC), 0.215 mm (95HD), and 0.015 mm (AHD), respectively as shown in Table 5.1. If combining hybrid loss with attention module, according to Fig. 5.7(c), the hybrid loss guides the attention module to pay attention to regions and boundaries of LA, which makes the red region closer to the boundary than Fig. 5.7(b). Fig. 5.8 shows the comparison of the proposed method and other state-of-the-art architectures. The white pixels of our method are the least in all state-of-the-art methods. Therefore, the proposed hybrid loss is applied to treat regions and boundaries of LA fairly during the training process, yielding to better results for all the networks.

5.2.7 Ablation Study

To explain the advantages of the proposed hybrid loss, we conduct an ablation study. We compare the segmentation results with and without hybrid loss (see Table 5.1). Segmentation performance increases for DC, 95HD and AHD for the 4 architectures, proving the benefits of the proposed hybrid loss.

We continue to test the hybrid loss on other datasets, and choose one multi-class segmentation task such as the MICCAI Workshop on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease[2] (HVSMR16). The aim of **HVSMR16** [71] is to segment myocardium and blood pool, it provides 10 training cardiovascular magnetic resonance (CMR) scans. For each patient, three kinds of images were provided: the full-volume axial images, the cropped axial images around the heart and thoracic aorta, and the cropped short axis reconstruction. In the current work, we only use the full-volume axial images. The slice spacings of the full-volume axial images range from 0.65 mm/pixel to 1.15 mm/pixel, while in-plane resolution ranged from 0.73 mm/pixel to 1.15 mm/pixel. The average sizes: $387 \times 387 \times 165$ pixels.

For the HVSMR16 dataset, we resize with a fixed pixel-spacing ($0.65mm$) and then crop to $250 \times 384 \times 384$, finally, use z-score normalization before inputting the network. We choose a previously proposed framework [19] as shown in Fig. 3.1 to complete experiments. Table 5.2 shows the segmentation results using a 5 fold cross-validation on HVSMR16 dataset. Compared without hybrid loss, adding the hybrid loss improves the segmentation accuracy of myocardium and blood pool from 75.15% to 78.36% and from 82.33% to 85.07% in term of dice, respectively.

[2] http://segchd.csail.mit.edu/index.html

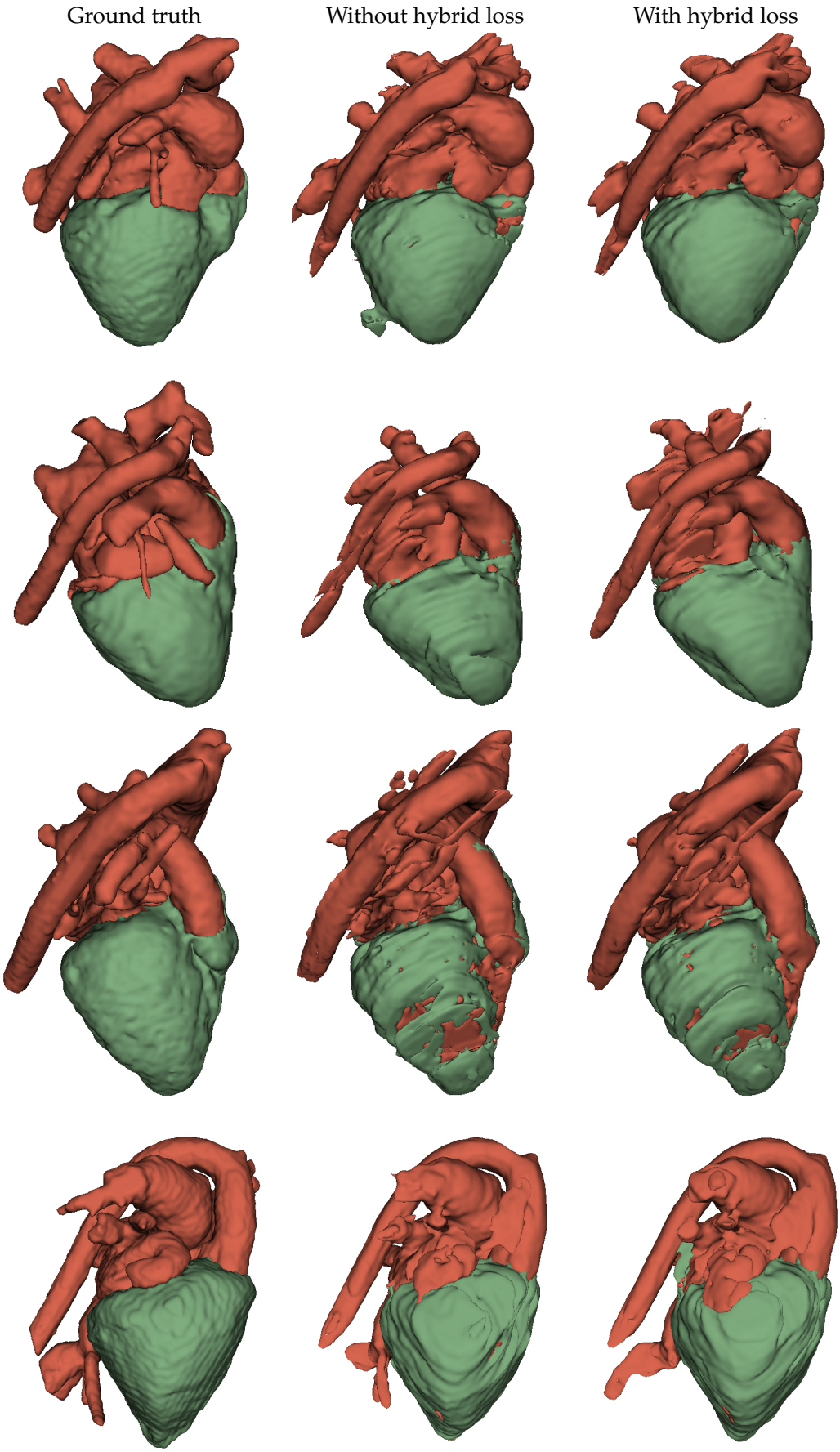


FIGURE 5.9: The 3D view of segmentation results based on HVSMR16 dataset; red color denotes blood pool, green color denotes myocardium

Fig. 5.9 shows the 3D view of segmentation results based on HVSMR16 dataset. We do not use post-precessing for the segmentation results, which is to give intuitive comparison. Compared without hybrid loss, the segmentation results that is obtained by using hybrid loss are more complete for blood pool part.

TABLE 5.2: Segmentation results using a 5 fold cross-validation on HVSMR16 dataset

Hybrid loss	**DC/%**	
	Myocardium	**Blood pool**
	75.15(±4.99)	82.33(±12.38)
✔	**78.36(±4.44)**	**85.07(±10.49)**

Compared with the myocardium segmentation results of A^0Net in Table. 4.3 of chapter 4, the myocardium segmentation results decrease 4% by using the attention network framework. The main reason is that the large model is applied to the too small dataset (HVSMR16 only provides 10 patients), leading to severe overfitting. The trainable parameters of the attention network framework exceed those of A^0Net by more than ten times.

5.3 Conclusion

In this chapter, we propose a novel attention network architecture, and a new hybrid loss. By using the attention module, the proposed network framework is able to prevent the interferences between the surrounding similar tissues and to capture large-scale and thiner structures. We propose a hybrid loss function that fairly treats regions and boundaries of objects, optimizes the convergence to the boundaries, while maintaining the segmentation precision of the regions. Compared to the state-of-the-arts methods on the AtriaSeg18 challenge dataset, our segmentation results overcome the best one by an average of 2.179% in terms of DC and 1.3 mm on 95HD. After that, we continue to experiment on multi-class task based on HVSMR16 dataset, and then the performance of hybrid loss still remain good. therefore, our method with attention module and hybrid loss is more robust. The computation time of our pipeline is less than 4 seconds for an entire 3D volume of a heart.

Part IV

Evaluation Methods of Fibrosis

Chapter 6

Evaluation of Fibrosis

This chapter mainly describes two parts: 1) Based on the heart segmentation results of chapter 4 and chapter 5, the fibrosis results are obtained by using one morphology method. 2) One end-to-end deep learning approach is used in segmenting the fibrosis.

6.1 Combine the deep learning and morphology method

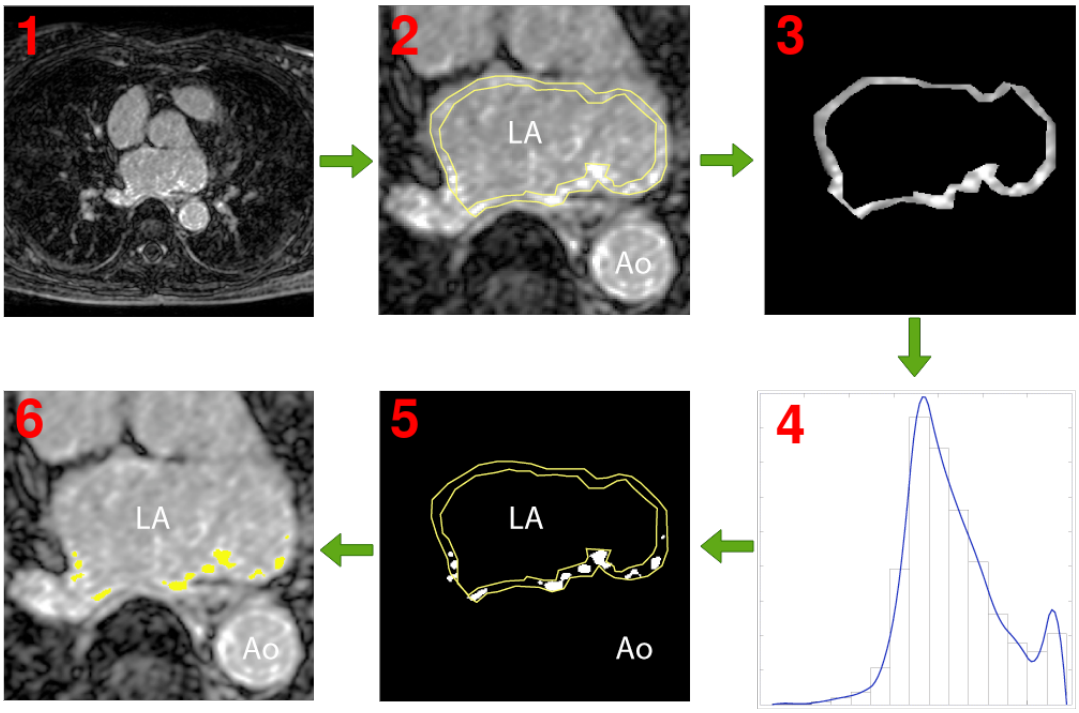


FIGURE 6.1: Scheme of the proposed process. 1. Input MRI. 2. Myocardial contours. 3. LA wall. 4. Histogram. 5. 3SD threshold. 6. Detection

Fig. 6.1 presents the expected workflow: segmentation of the heart volume leading to the identification of the left atrial wall, analysis of the radiometry within the wall, thresholding to quantify the fibrosis degree. The heart segmentation can be completed by the segmentation methods of chapter 4 and chapter 5. The analysis part can rely on a mathematical morphology approach.

Therefore, we continue to quantify the fibrosis degree based on the heart segmentation results of chapter 5. The thickness of atrial walls is 2 to 4 mm, which is 2- to 3-fold thinner than ventricular walls. The spatial resolution of AtriaSeg18 dataset[1] is $0.625\times\ 0.625\ \times\ 0.625\ mm^3$. The AtriaSeg18 dataset provides the label of left atrial (LA) cavity, so the heart segmentation results of chapter 5 is left atrial (LA) cavity and the endocardial border is obtained. Next, the endocardial border is morphologically dilated (by 4 pixel layers, 2.5 mm) and then manually adjusted to create the shell of the epicardial LA surface [28]. In a final step, the endocardial segmentation is subtracted from the epicardial layer to define the wall segmentation as shown in Fig 6.2.

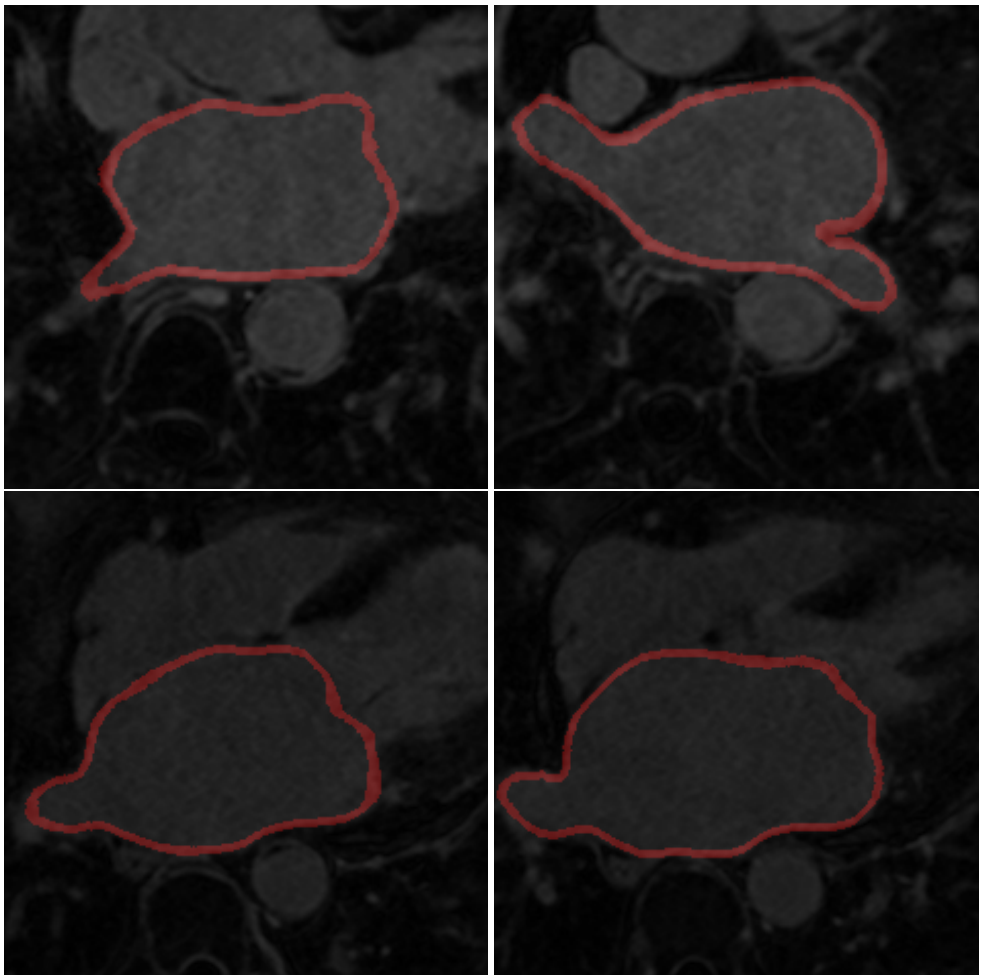


FIGURE 6.2: Left atrial wall segmentation

After obtaining the wall segmentation, we assume that the image only includes the left atrial A is defined as $A = ES \times I$, where ES denotes the endocardial segmentation result (binary image) and I denotes the gray image of heart. Then we calculate the mean value M and the standard deviation SD of $A > 0$, and let threshold is set to $M + 3SD$. Finally, the fibrosis is obtained by $W > (M + 3SD)$ (W denotes that the the image only includes the left atrial wall) as shown in Fig 6.3.

[1] https://atriaseg2018.cardiacatlas.org/

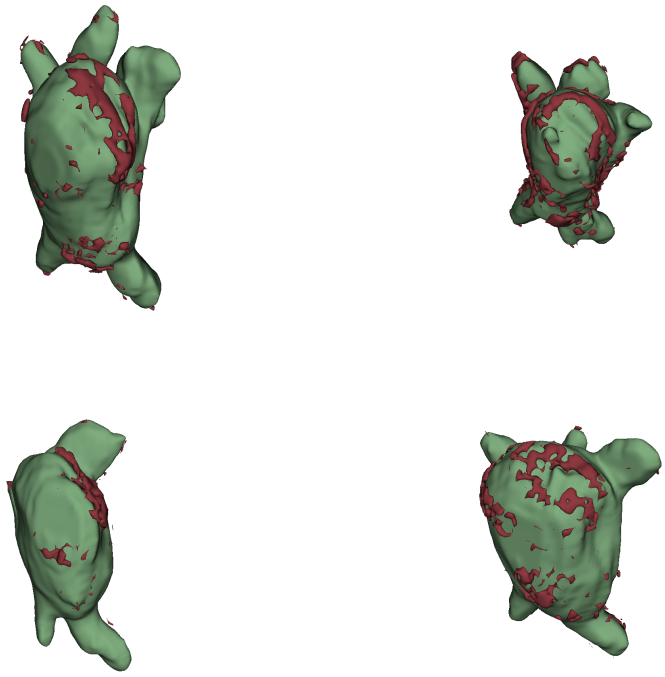


FIGURE 6.3: 3D view of fibrosis and left atrial wall; red color denotes fibrosis and green color denotes left atrial wall

Because the AtriaSeg18 dataset does not provide the label of scar, we continue to test our method on Left Atrial and Scar Quantification & Segmentation Challenge [2] (LAScarQS2022) [29–31], and the LAScarQS2022 aims to segment the left atrium and evaluates the scar. It includes two tasks (Task 1 and Task 2) and Task 1 contains the scar data, so we only use the dataset of Task 1. Task 1 contains 60 annotated 3D MRIs from patients with atrial fibrillation for training and validating. The voxel size of the MR images is different: 1.25 × 1.25 × 2.5 mm, 1.4 × 1.4 × 1.4 mm, and 1.3 × 1.3 × 4.0 mm. The dataset includes two different image sizes: $44\times576\times576$ pixels and $44\times640\times640$ pixels.

TABLE 6.1: Ablation study of SD on LAScarQS2022 dataset using a 5 fold cross-validation.

Different SD	DC of scar
SD	0.328±0.035
2SD	0.305±0.067
3SD	0.062±0.038

The use of the threshold value for all patients is no feasible because the contrast between normal and fibrotic myocardium in LGE-MRI of the left atrium depends on multiple factors: patient heart rate and rhythm during MRI study, type and dosage

[2] https://zmiclab.github.io/projects/lascarqs22/

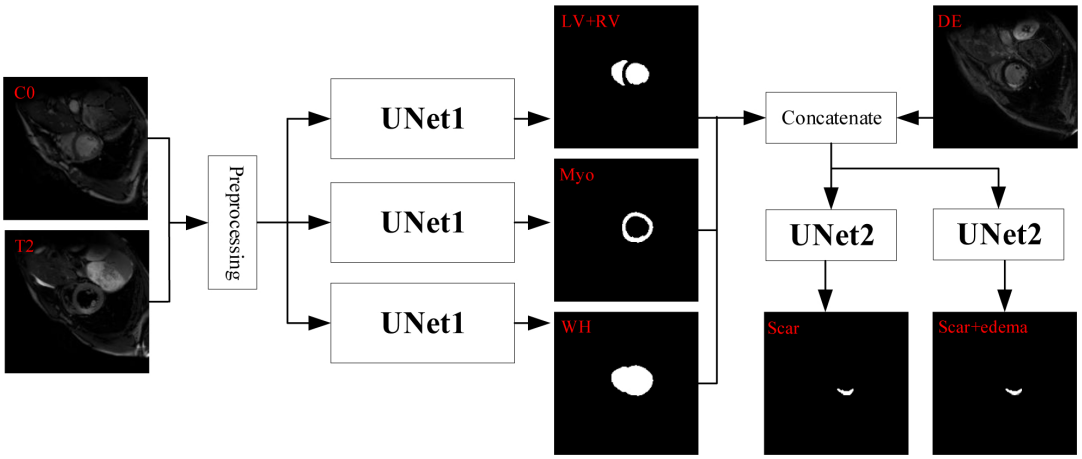


FIGURE 6.4: Global overview of the proposed method.

of contrast agent, time between contrast administration and LGE-MRI scan, patient-specific contrast clearance rate, choice of TI value for LGE scan, strength of the main field of MRI scanner, patient body mass index (BMI), blood hematocrit, and oxygenation level [28]. Therefore, we continue to consider other methods. Next, we will try deep learning method.

6.2 Deep learning method

6.2.1 Methodology

6.2.1.1 Overview of Network Architecture

We propose a hybrid network (see Fig. 6.4) using UNet [53] to the myocardial pathology segmentation, which is consisted by five UNet frameworks. The main difference between **UNet1** and **UNet2** is the filter number as shown in Table. 6.2: the filter number of **UNet1** is [64 128 256 512 256 128 64] and the filter number of **UNet2** is [8 16 32 64 32 16 8], but their framework is same, which consists of two parts as shown in Fig. 6.5: a down-sampling part and an up-sampling part and fuses high-level features and low-level features by a shortcut connection between the two parts. **UNet1** is used to segment the normal tissue around myocardial pathology and obtain three segmentation results on LV+RV, Myo, and WH, respectively. **UNet2** is used to segment myocardial pathology by learning the relationship between the surrounding normal tissue and myocardial pathology. Since the number of myocardial pathology samples is much smaller than the number of normal tissues around it, compared with **UNet1**, we reduce the filter number **UNet2** in order to reduce the impact of overfitting.

TABLE 6.2: The structural configuration of UNet.

Layers	Input size		Operation	kernel	Stride	Regularization	Output size	
	UNet1	UNet2					UNet1	UNet2
Input image	(240,240,2)	(240,240,4)	-	-	-	-	(240,240,2)	(240,240,4)
C1	(240,240,2)	(240,240,4)	[Conv2d+relu]*2	3	1	L2	(240,240,64)	(240,240,8)
C2	(240,240,64)	(240,240,8)	Maxpooling2d	2	-	-	(120,120,64)	(120,120,8)
C3	(120,120,64)	(120,120,8)	[Conv2d+relu]*2	3	1	L2	(120,120,128)	(120,120,16)
C4	(120,120,128)	(120,120,16)	Maxpooling2d	2	-	-	(60,60,128)	(60,60,16)
C5	(60,60,128)	(60,60,16)	[Conv2d+relu]*2	3	1	L2	(60,60,256)	(60,60,32)
C6	(60,60,256)	(60,60,32)	Maxpooling2d	2	-	-	(30,30,256)	(30,30,32)
C7	(30,30,256)	(30,30,32)	[Conv2d+relu]*2+Dropout	3	1	L2	(30,30,512)	(30,30,64)
C8	(30,30,512)	(30,30,64)	Maxpooling2d	2	-	-	(15,15,512)	(15,15,64)
C9	(15,15,512)	(15,15,64)	[Conv2d+relu]*2+Dropout	3	1	L2	(15,15,1024)	(15,15,128)
O1	(240,240,2)	(240,240,2)	Sigmoid	-	-	-	(240,240,1)	(240,240,1)

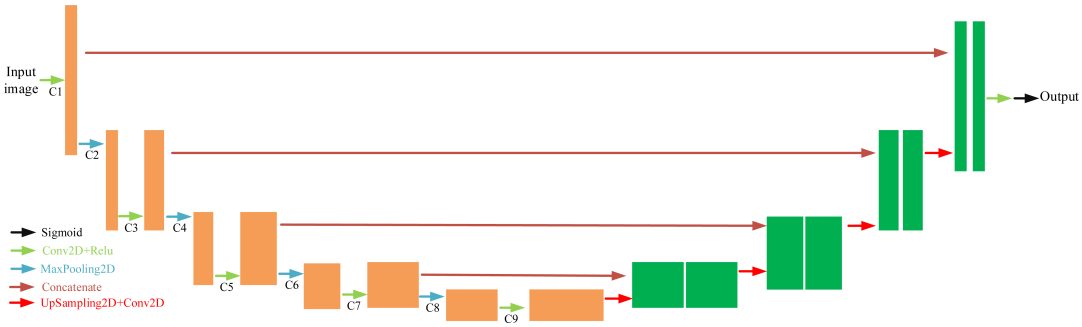


FIGURE 6.5: Architecture of networks.

6.2.2 Experimental Results

Dataset Description. We evaluate our method on the myocardial pathology segmentation combining multi-sequence CMR [3] (MyoPS 2020). Its aim is to myocardial pathology segmentation. It contains 45 cases of multi-sequence CMR (25 cases for training and 20 cases for testing). Each case refers to a patient with three sequence CMR, i.e., LGE, T2 and balanced-Steady State Free Precession (bSSFP) CMR. The LGE CMR sequence can visualize myocardial infarction. The T2-weighted CMR shows the acute injury and ischemic regions. The bSSFP cine sequence captures cardiac motions and presents clear boundaries. The slice spacings of multi-sequence CMR volume range from 11.999 mm/pixel to 23.000 mm/pixel, while in-plane resolution ranged from 0.729 mm/pixel to 0.762 mm/pixel. The average sizes: $482 \times 479 \times 4$ pixels.

Preprocessing and Postprocessing. We cropped each slice to 240×240 pixels and we do not use data augmentation. The pre-processing begins with a Gaussian normalization. For post-processing, we pad with zeros to get back a initial width and height of a slice.

Implementation and Experimental Setup. We implemented our experiments on Keras/TensorFlow using a NVidia Quadro P6000 GPU. We used five different loss functions for training the network and used sigmoid to get a probability distribution of the left and right ventricle, myocardium, whole heart, scar and edema, and scar, respectively (as shown in Fig. 4.1). Adam optimizer (batchsize = 1, $\beta_1 = 0.9$, $\beta_2 =$

[3] http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/MyoPS20/index.html

TABLE 6.3: Evaluation results on 5-fold-cross-validation.

Patient	101-105	106-110	111-115	116-120	121-125	Average	Test datasets
Edema	0.284	0.153	0.189	0.122	0.280	0.206	–
Scar	0.473	0.496	0.515	0.464	0.602	0.510	0.586
Myo	0.844	0.852	0.811	0.859	0.869	0.847	–
LV+RV	0.818	0.854	0.812	0.897	0.864	0.849	–
WH	0.925	0.937	0.876	0.918	0.959	0.923	–

0.999, $\varepsilon = 0.001$, $\text{lr} = 0.0001$) and did not use learning rate decay. We trained the network during 300 epochs.

Training Step. First, we kept weight of **UNet2** unchanged, which means **UNet2** was not trained at the beginning, then we trained **UNet1**. After finished the train of **UNet1**, we kept weight of **UNet1** unchanged, then trained **UNet2**.

Evaluation Methods. One metric is used to evaluate our method: dice coefficient (DC) to evaluate the regions of myocardial pathology.

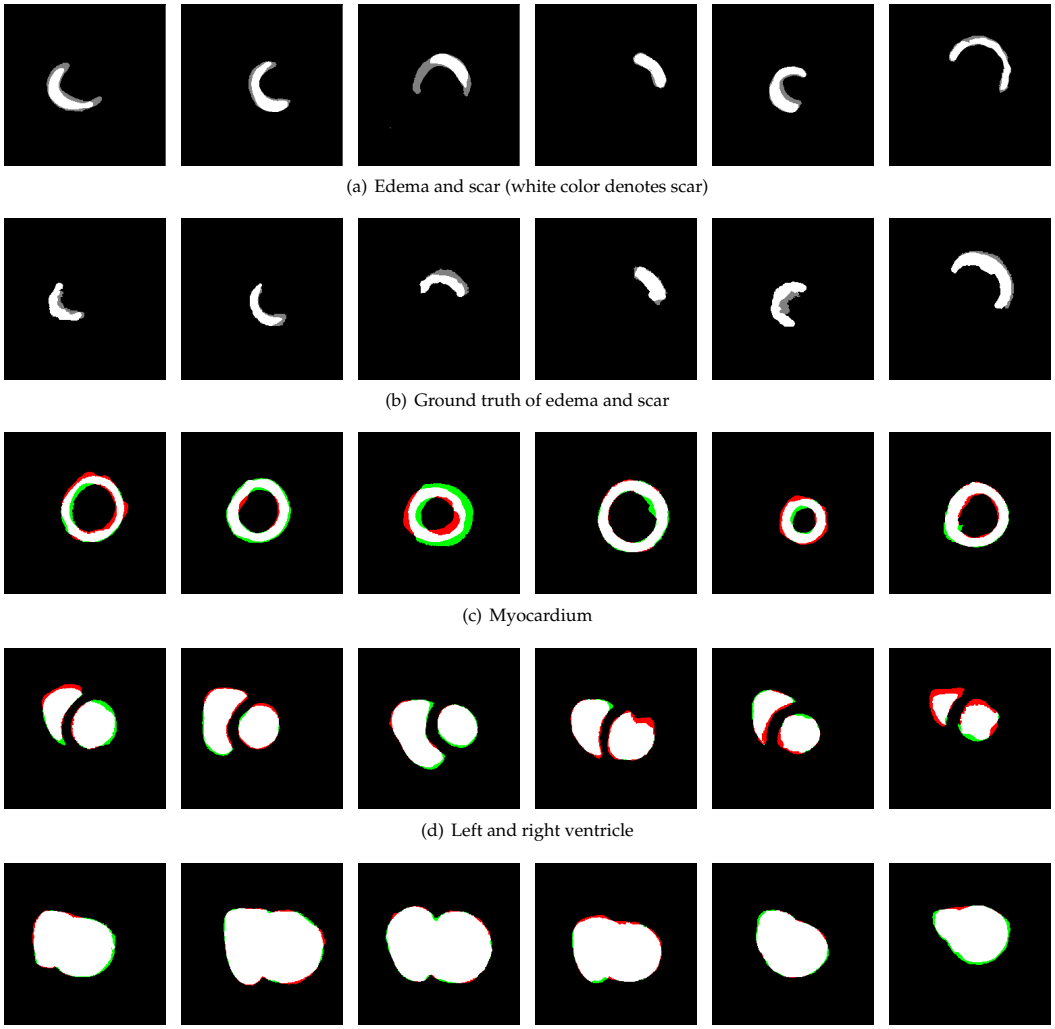
6.2.2.1 Segmentation Results

As shown in Table. 6.3, we evaluate the proposed method with 5-fold-cross-validation. We obtain a mean DC of 92.3% on WH, 84.9% on LV+RV, and 84.7% on Myo by **UNet1**. Without using data augmentation, based on the original dataset, we obtain a higher segmentation accuracy, which lays the foundation for the subsequent segmentation of myocardial pathology. Finally, we obtain a mean DC of 20.6% on edema, 51% on scar by **UNet2**. We used the trained network to predict the testset (20 cases) and received the evaluation of our prediction results from the MyoPS2020 organizer: the mean DC of 58.6% on scar and the mean DC of 63.9% on scar and edema.

As shown in Fig. 6.6, for the segmentation results of whole heart, left and right ventricle, and myocardium, as the number of positive samples continues to decrease, the segmentation accuracy is also decreasing, and false segmentation is mainly concentrated at the boundary, which is mainly because ambiguities often appear near the boundaries of the target domains due to tissue similarities. For the segmentation results of edema and scar, the poorly segmentation result is not only on the boundary, but also in regions. In the original dataset, edem does not exist in many slices, which further leads to a reduction in the effective dataset for edema, therefore, the segmentation network is very difficult to segment edema.

6.2.2.2 Conclusion

In this chapter, we propose a way of reverse thinking, not to segment the myocardial pathology directly, but to learn a relationship between the surrounding normal tissue and it by designing one stacked and parallel UNets with multi-output framework. We evaluate the proposed method with 5-fold-cross-validation on the



(a) Edema and scar (white color denotes scar)

(b) Ground truth of edema and scar

(c) Myocardium

(d) Left and right ventricle

(e) Whole heart

FIGURE 6.6: Segmentation results. Red color denotes false positive and green color denotes false negative.

MICCAI 2020 myocardial pathology segmentation combining multi-sequence CMR Challenge dataset (MyoPS 2020) and achieve a mean DC of 20.6%, 51% on edema and scar, respectively. The computation time of the entire pipeline is less than 3 seconds on Quadro P6000 GPU for an entire 3D volume, making it usable for clinical practice. However, the segmentation accuracy of myocardial pathology is affected by the segmentation accuracy of surrounding normal tissues.

Part V

Conclusion

Chapter 7

Conclusion and Perspectives

According to the World Health Organization (WHO), cardiovascular diseases (CVDs) are the leading cause of death globally. Medical imaging becomes increasingly important for the diagnosis and treatments of CVDs. Medical imaging contains many modalities such as computed tomography (CT), positron emission tomography (PET), ultrasound (US) and magnetic resonance imaging (MRI) and so on. Comparing with the others modalities, MRI has one great contrast between soft tissues and relatively high spatial resolutions (this is why we choose the MRI dataset.). But there are some difficulties for using MRI images to segment:

- there is a poor contrast between myocardium and surrounding structures;
- brightness due to blood flow;
- non-homogeneous partial volume due to limited MRI resolution;
- noise due to motion artifacts and heart dynamics;
- shape and intensity variability due to different patients and pathologies.

Therefore, based on the above difficulties, our research significance is derived. In this thesis, we mainly use deep learning methods to solve related problems in cardiac segmentation and evaluation of fibrosis.

7.1 Main results

Firstly, we explore the sensitivity of networks to noise for different preprocessing methods in chapter 3, Through comparative experiments, it is concluded that the standardization preprocessing method is the best for the output of the network. Therefore, we choose this preprocessing method for the subsequent processing of the dataset.

Secondly, we design novel network frameworks to segment heart in chapter 4 and chapter 5. The first proposed framework is one two-stage architecture, which includes two parts that are one localization network and one segmentation network. The localization network is used in localizing roughly the object position, which can reduce the useless information (negative sample). The segmentation network is devoted to accurately segment the object. Due to the fact that many methods mainly

focus on the region accuracy of the heart, more than to the quality of the boundaries, we prensent one novel hybrid loss that combines Categorical Cross Entropy (CCE), Structural Similarity (SSIM) and Dice Coefficient (DC) to study the transformation relationship between the input image and the corresponding label in a three level hierarchy (pixel-, patch- and map-level), which is helpful to improve segmentation and recovery of the boundaries. We demonstrate the efficiency of our approach on three public datasets in terms of regional and boundary segmentations. The second proposed framework is one end-to-end architecture, which is an attention full convolutional network framework based on the ResNet-101 architecture and focuses on boundaries as much as on regions. The additional attention module is added to have the network pay more attention on regions and then to reduce the impact of the misleading similarity of neighboring tissues. We also use a hybrid loss composed of a region loss and a boundary loss to treat boundaries and regions at the same time. We demonstrate the efficiency of the proposed approach on three public datasets.

Finally, in chapter 6, two different methods are used in evaluation of fibrosis. The first method is that we combine the deep learning method with morphology. The left atrial wall is obtained based on the segmentation results of chapter 5 by morphologically dilating, and then thresholds to quantify the fibrosis degree. The second method is that we provide one cascaded UNet framework and uses three different modalities (the late gadolinium enhancement (LGE) CMR sequence,the balanced-Steady State Free Precession (bSSFP) cine sequence and the T2-weighted CMR) to complete the segmentation of the myocardium, scar and edema in the context of the MICCAI 2020 myocardial pathology segmentation combining multi-sequence CMR Challenge dataset (MyoPS 2020). We evaluate the proposed method with 5-fold-cross-validation on the MyoPS 2020 dataset.

7.2 Future work

The work described in this PhD thesis provides many ways for further research.

7.2.1 Multi-modality and multi-task

Multi-modality:. The problem of overfitting is common in medical image segmentation, because the dataset of medical images is small, maybe only contains a few patients. However, it contains many modalities, and each modality contains different information. Therefore, trying to utilize multi-modality information to segment the heart is necessary as showed in fig.7.1.

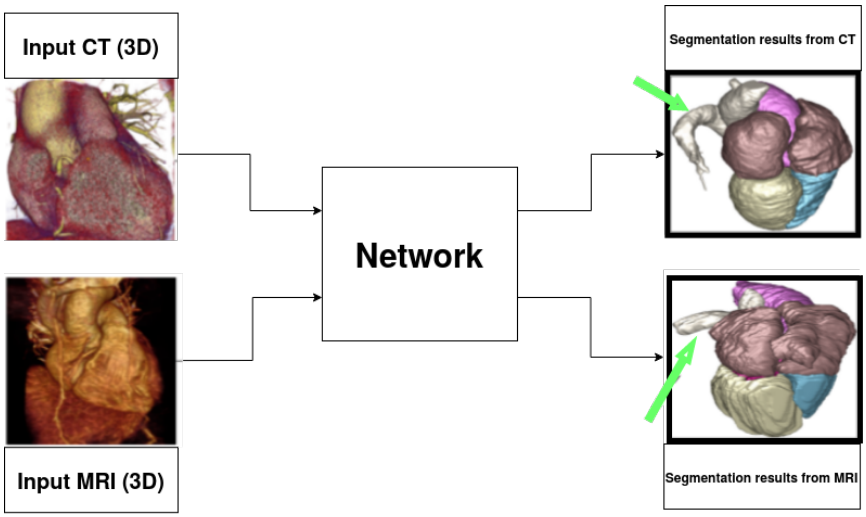


FIGURE 7.1: Multi-modality information. A single modality contains only limited information, but the information of several modalities can complement each other.

Multi-task:. For the same network framework, one multi-task such as classification and segmentation task is implemented. These tasks will affect each other during the training process, either positively or negatively. Therefore, trying to use multi-task method to segment the heart is worth exploring. For example, adding the ground truth of spatial constraint guides the segmentation network to study the transformation relationship between the input image and the corresponding label as showed in fig.7.2.

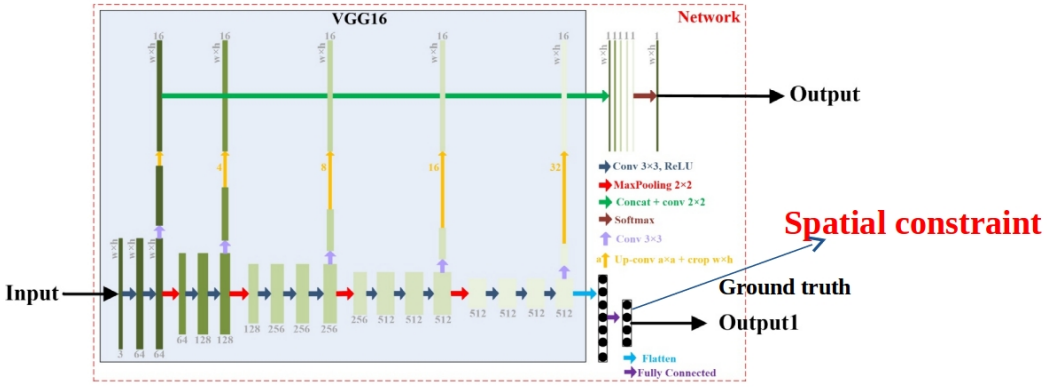


FIGURE 7.2: Multi-task.

7.2.2 Hybrid loss

In the training phase, the loss function is an essential part, which guides the network to learn the transformation relationship between the input image and the corresponding label. Therefore, it is very important to design a loss function that meets the requirements of the task. However, only using one loss function in the network

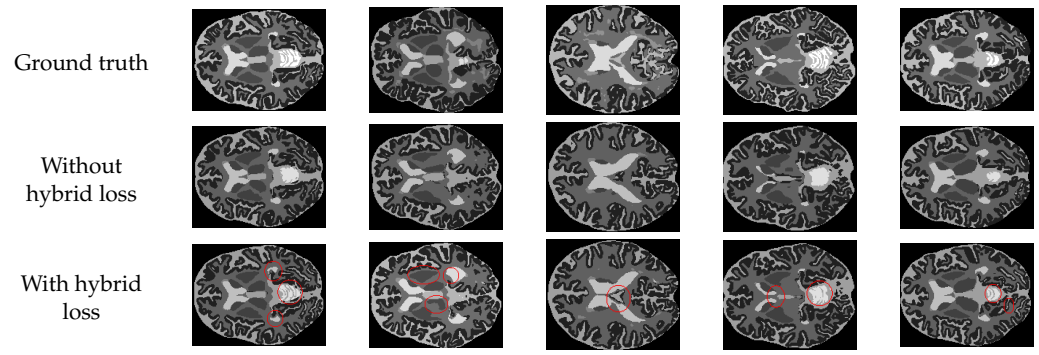


FIGURE 7.3: Partial segmentation results based on the MRBrains18 dataset: in this dataset, there exists a serious imbalance between the eight brain structures (different proportions). For the brain structures with a small proportion (see the red circle), the hybrid loss strongly helps to produce detailed segmentations; we say that the networks learn to see more clearly in the input images.

is not enough, so designing one hybrid loss that combines different loss function is required.

We have tested the proposed hybrid loss on other non-cardiac datasets, for example, Grand Challenge on MR Brain Segmentation at MICCAI 2018[1] (MRBrainS2018). The aim of **MRBrainS2018** is to segment the 8 brain structure such as cortical gray matter, basal ganglia, white matter, white matter lesions, cerebrospinal fluid in the extracerebral space, ventricles, cerebellum and brain stem. It contains 30 MRI scans, which provides contains three modalities such as T1-weighted, T1-weighted inversion recovery and T2-FLAIR. Seven of them are released as the training dataset. Another 23 scans are kept unreleased for test dataset. The dataset includes same image size: $48 \times 240 \times 240$.

For the MRBrainS2018 dataset, we use z-score normalization as preprocessing. Table 7.1 shows the segmentation results using a 7 fold cross-validation on MRBrainS2018 dataset. If adding the hybrid loss into the network, the segmentation results of each brain structure are both improved. Corresponding to Fig. 7.3, The partial segmentation results of brain are shown for with or without hybrid loss. Using hybrid loss in the network, more details are segmented by the network.

[1] https://mrbrains18.isi.uu.nl/

TABLE 7.1: Segmentation results using a 7 fold cross-validation on MRBrainS2018 dataset

Brain structure	Dice/%	
	Without hybrid loss	**With hybrid loss**
Cortical gray matter	84.89	85.44
Basal ganglia	81.37	82.64
White matter	85.36	86.07
White matter lesions	31.08	40.20
CSF	81.98	82.35
Ventricles	91.89	92.86
Cerebellum	89.74	90.65
Brain stem	71.83	74.44

CSF denotes cerebrospinal fluid in the extracerebral space.

7.2.3 Attention method

In the training phase, too much redundant information is reused. Therefore, it is necessary to add attention modules to the network to reduce the utilization of redundant information. Designing dedicated attention modules for different tasks is worth exploring.

Bibliography

[1] P. Kirchhof, S. Benussi, D. Kotecha, A. Ahlsson, D. Atar, B. Casadei, M. Castella, H.-C. Diener, H. Heidbuchel, J. Hendriks *et al.*, "2016 esc guidelines for the management of atrial fibrillation developed in collaboration with eacts," *European journal of cardio-thoracic surgery*, vol. 50, no. 5, pp. e1–e88, 2016.

[2] N. F. Marrouche, D. Wilber, G. Hindricks, P. Jais, N. Akoum, F. Marchlinski, E. Kholmovski, N. Burgon, N. Hu, L. Mont *et al.*, "Association of atrial tissue fibrosis identified by delayed enhancement mri and atrial fibrillation catheter ablation: the decaaf study," *Jama*, vol. 311, no. 5, pp. 498–506, 2014.

[3] R. S. Oakes, T. J. Badger, E. G. Kholmovski, N. Akoum, N. S. Burgon, E. N. Fish, J. J. Blauer, S. N. Rao, E. V. DiBella, N. M. Segerson *et al.*, "Detection and quantification of left atrial structural remodeling using delayed enhancement mri in patients with atrial fibrillation," *Circulation*, vol. 119, no. 13, p. 1758, 2009.

[4] J. Seitz, J. Horvilleur, J. Lacotte, D. OH-ICI, Y. Mouhoub, A. Maltret, F. Salerno, D. Mylotte, M. Monchi, and J. Garot, "Correlation between af substrate ablation difficulty and left atrial fibrosis quantified by delayed-enhancement cardiac magnetic resonance," *Pacing and clinical electrophysiology*, vol. 34, no. 10, pp. 1267–1277, 2011.

[5] E. M. Benito, A. Carlosena-Remirez, E. Guasch, S. Prat-González, R. J. Perea, R. Figueras, R. Borràs, D. Andreu, E. Arbelo, J. M. Tolosana *et al.*, "Left atrial fibrosis quantification by late gadolinium-enhanced magnetic resonance: a new method to standardize the thresholds for reproducibility," *Ep Europace*, vol. 19, no. 8, pp. 1272–1279, 2017.

[6] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi *et al.*, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 65, 2018.

[7] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "ω-net (omega-net): fully automatic, multi-view cardiac mr detection, orientation, and segmentation with deep neural networks," *Medical image analysis*, vol. 48, pp. 95–106, 2018.

[8] Z. Xiong, V. V. Fedorov, X. Fu, E. Cheng, R. Macleod, and J. Zhao, "Fully automatic left atrium segmentation from late gadolinium enhanced magnetic resonance imaging using a dual fully convolutional neural network," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 515–524, 2018.

[9] C. J. Preetha, S. Haridasan, V. Abdi, and S. Engelhardt, "Segmentation of the left atrium from 3d gadolinium-enhanced mr images with convolutional neural networks," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 265–272.

[10] C. Bian, X. Yang, J. Ma, S. Zheng, Y.-A. Liu, R. Nezafat, P.-A. Heng, and Y. Zheng, "Pyramid network with online hard example mining for accurate left atrium segmentation," in *international workshop on statistical atlases and computational models of the heart*. Springer, 2018, pp. 237–245.

[11] C. Chen, W. Bai, and D. Rueckert, "Multi-task learning for left atrial segmentation on ge-mri," in *International workshop on statistical atlases and computational models of the heart*. Springer, 2018, pp. 292–301.

[12] G. Yang, X. Zhuang, H. Khan, S. Haldar, E. Nyktari, X. Ye, G. Slabaugh, T. Wong, R. Mohiaddin, J. Keegan *et al.*, "Segmenting atrial fibrosis from late gadolinium-enhanced cardiac mri by deep-learned features with stacked sparse auto-encoders," in *Annual Conference on Medical Image Understanding and Analysis*. Springer, 2017, pp. 195–206.

[13] G. Yang, X. Zhuang, H. Khan, S. Haldar, E. Nyktari, L. Li, R. Wage, X. Ye, G. Slabaugh, R. Mohiaddin *et al.*, "Fully automatic segmentation and objective assessment of atrial scars for long-standing persistent atrial fibrillation patients using late gadolinium-enhanced mri," *Medical physics*, vol. 45, no. 4, pp. 1562–1576, 2018.

[14] J. Chen, G. Yang, Z. Gao, H. Ni, E. Angelini, R. Mohiaddin, T. Wong, Y. Zhang, X. Du, H. Zhang *et al.*, "Multiview two-task recursive attention model for left atrium and atrial scars segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 455–463.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR abs/1409.1556, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[18] K. K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Deep retinal image understanding," ser. Lecture Notes in Computer Science, vol. 9901, 2016, pp. 140–148.

[19] É. Puybareau, Z. Zhao, Y. Khoudli, E. Carlinet, Y. Xu, J. Lacotte, and T. Géraud, "Left atrial segmentation in a few seconds using fully convolutional network and transfer learning," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 339–347.

[20] Z. Zhao, N. Boutry, É. Puybareau, and T. Géraud, "A two-stage temporal-like fully convolutional network framework for left ventricle segmentation and quantification on MR images," ser. Lecture Notes in Computer Science, vol. 12009. Springer, 2019, pp. 405–413.

[21] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1409.1556, 2014.

[22] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. of the Intl. Conf. on Neural Information Processing Systems (NIPS)*, 2018, pp. 8792–8802.

[23] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. of the 37th Asilomar Conference on Signals, Systems and Computers*, vol. 2, 2003, pp. 1398–1402.

[24] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[25] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. of CVPR*, 2009, pp. 248–255.

[26] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. of CVPR*, 2019, pp. 3146–3154.

[27] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*. PMLR, 2019, pp. 285–296.

[28] J. Siebermair, E. G. Kholmovski, and N. Marrouche, "Assessment of left atrial fibrosis by late gadolinium enhancement magnetic resonance imaging: methodology and clinical implications," *JACC: Clinical Electrophysiology*, vol. 3, no. 8, pp. 791–802, 2017.

[29] L. Li, V. A. Zimmer, J. A. Schnabel, and X. Zhuang, "Atrialgeneral: Domain generalization for left atrial segmentation of multi-center lge mris," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 557–566.

[30] ——, "Medical image analysis on left atrial lge mri for atrial fibrillation studies: A review," *Medical Image Analysis*, p. 102360, 2022.

[31] ——, "Atrialjsqnet: A new framework for joint segmentation and quantification of left atrium and scars incorporating spatial and shape information," *Medical Image Analysis*, vol. 76, p. 102303, 2022.

[32] P. Kellman and A. E. Arai, "Cardiac imaging techniques for physicians: late enhancement," *Journal of magnetic resonance imaging*, vol. 36, no. 3, pp. 529–542, 2012.

[33] R. Karim, L.-E. Blake, J. Inoue, Q. Tao, S. Jia, R. J. Housden, P. Bhagirath, J.-L. Duval, M. Varela, J. M. Behar *et al.*, "Algorithms for left atrial wall segmentation and thickness–evaluation on an open-source ct and mri image database," *Medical image analysis*, vol. 50, pp. 36–53, 2018.

[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[35] R. Karim, R. J. Housden, M. Balasubramaniam, Z. Chen, D. Perry, A. Uddin, Y. Al-Beyatti, E. Palkhi, P. Acheampong, S. Obom *et al.*, "Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, pp. 1–17, 2013.

[36] B. Ambale-Venkatesh and J. A. Lima, "Cardiac mri: a central prognostic tool in myocardial fibrosis," *Nature Reviews Cardiology*, vol. 12, no. 1, p. 18, 2015.

[37] D. C. Preston, "Magnetic resonance imaging (mri) of the brain and spine: Basics," *MRI Basics, Case Med*, vol. 30, 2006.

[38] M. Y. Wang, X. Wang, and D. Guo, "A level set method for structural topology optimization," *Computer methods in applied mechanics and engineering*, vol. 192, no. 1-2, pp. 227–246, 2003.

[39] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[40] J. B. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundamenta informaticae*, vol. 41, no. 1, 2, pp. 187–228, 2000.

[41] R. J. Kim, D. S. Fieno, T. B. Parrish, K. Harris, E.-L. Chen, O. Simonetti, J. Bundy, J. P. Finn, F. J. Klocke, and R. M. Judd, "Relationship of mri delayed

contrast enhancement to irreversible injury, infarct age, and contractile function," *Circulation*, vol. 100, no. 19, pp. 1992–2002, 1999.

[42] A. Kolipaka, G. P. Chatzimavroudis, R. D. White, T. P. O'Donnell, and R. M. Setser, "Segmentation of non-viable myocardium in delayed enhancement magnetic resonance images," *The international journal of cardiovascular imaging*, vol. 21, no. 2, pp. 303–311, 2005.

[43] A. Schmidt, C. F. Azevedo, A. Cheng, D. A. Bluemke, G. Gerstenblith, Weiss *et al.*, "Infarct tissue heterogeneity by magnetic resonance imaging identifies enhanced cardiac arrhythmia susceptibility in patients with left ventricular dysfunction," *Circulation*, vol. 115, no. 15, p. 2006–2014, 2007.

[44] L. C. Amado, B. L. Gerber, S. N. Gupta, D. W. Rettmann, G. Szarf, R. Schock, K. Nasir, D. L. Kraitchman, and J. A. Lima, "Accurate and objective infarct sizing by contrast-enhanced magnetic resonance imaging in a canine myocardial infarction model," *Journal of the American College of Cardiology*, vol. 44, no. 12, pp. 2383–2389, 2004.

[45] D. M. Leistner, S. Palm, I. Ziegler, I. Diehl, C. Meyer-Woelden, I. Burck, B. Assmus, F. Seeger, T. J. Vogl, and A. M. Zeiher, "Characterization of the peri-infarct zone by contrast-enhanced magnetic resonance imaging and 18f-fdg positron emission tomography and its clinical impact in patients with coronary artery disease," 2010.

[46] V. Positano, A. Pingitore, A. Giorgetti, B. Favilli, M. F. Santarelli, L. Landini, P. Marzullo, and M. Lombardi, "A fast and effective method to assess myocardial necrosis by means of contrast magnetic resonance imaging," *Journal of Cardiovascular Magnetic Resonance*, vol. 7, no. 2, pp. 487–494, 2005.

[47] J. S. Detsky, G. Paul, A. J. Dick, and G. A. Wright, "Reproducible classification of infarct heterogeneity using fuzzy clustering on multicontrast delayed enhancement magnetic resonance images," *IEEE transactions on medical imaging*, vol. 28, no. 10, pp. 1606–1614, 2009.

[48] Y. Lu, Y. Yang, K. A. Connelly, G. A. Wright, and P. E. Radau, "Automated quantification of myocardial infarction using graph cuts on contrast delayed enhanced magnetic resonance images," *Quantitative imaging in medicine and surgery*, vol. 2, no. 2, p. 81, 2012.

[49] B. R. Knowles, D. Caulfield, M. Cooklin, C. A. Rinaldi, J. Gill, J. Bostock, R. Razavi, T. Schaeffter, and K. S. Rhode, "3-d visualization of acute rf ablation lesions using mri for the simultaneous determination of the patterns of necrosis and edema," *IEEE transactions on biomedical engineering*, vol. 57, no. 6, pp. 1467–1475, 2010.

[50] A. Hennemuth, A. Seeger, O. Friman, S. Miller, B. Klumpp, S. Oeltze, and H.-O. Peitgen, "A comprehensive approach to the analysis of contrast enhanced cardiac mr images," *IEEE Transactions on Medical Imaging*, vol. 27, no. 11, pp. 1592–1610, 2008.

[51] Q. Tao, J. Milles, K. Zeppenfeld, H. J. Lamb, J. J. Bax, J. H. Reiber, and R. J. van der Geest, "Automated segmentation of myocardial scar in late enhancement mri using combined intensity and spatial information," *Magnetic Resonance in Medicine*, vol. 64, no. 2, pp. 586–594, 2010.

[52] 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*, 2015, pp. 3431–3440.

[53] 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*. Springer, 2015, pp. 234–241.

[54] Q. Xia, Y. Yao, Z. Hu, and A. Hao, "Automatic 3d atrial segmentation from gemris using volumetric fully convolutional networks," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 211–220.

[55] N. Savioli, G. Montana, and P. Lamata, "V-fcnn: volumetric fully convolution neural network for automatic atrial segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 273–281.

[56] S. Jia, A. Despinasse, Z. Wang, H. Delingette, X. Pennec, P. Jaïs, H. Cochet, and M. Sermesant, "Automatically segmenting the left atrium from cardiac images using successive 3d u-nets and a contour loss," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 221–229.

[57] S. Vesal, N. Ravikumar, and A. Maier, "Dilated convolutions in neural networks for left atrial segmentation in 3d gadolinium enhanced-mri," in *International workshop on statistical atlases and computational models of the heart*. Springer, 2018, pp. 319–328.

[58] C. Li, Q. Tong, X. Liao, W. Si, Y. Sun, Q. Wang, and P.-A. Heng, "Attention based hierarchical aggregation network for 3d left atrial segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 255–264.

[59] A. Mortazi, R. Karim, K. Rhode, J. Burt, and U. Bagci, "Cardiacnet: Segmentation of left atrium and proximal pulmonary veins from mri using multi-view

cnn," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 377–385.

[60] G. Yang, J. Chen, Z. Gao, H. Zhang, H. Ni, E. Angelini, R. Mohiaddin, T. Wong, J. Keegan, and D. Firmin, "Multiview sequential learning and dilated residual learning for a fully automatic delineation of the left atrium and pulmonary veins from late gadolinium-enhanced cardiac mri images," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 1123–1127.

[61] Z. Zhou *et al.*, "Cross-modal attention-guided convolutional network for multi-modal cardiac segmentation," in *Proc. of the Intl. Workshop on Mach. Learning in Med. Imaging*, ser. LNCS, vol. 11861. Springer, 2019, pp. 601–610.

[62] Ö. Çiçek *et al.*, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," ser. LNCS, vol. 9901. Springer, 2016, pp. 424–432.

[63] T. Zhang *et al.*, "Multiple attention fully convolutional network for automated ventricle segmentation in cardiac magnetic resonance imaging," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 5, pp. 1037–1045, 2019.

[64] Q. Tong *et al.*, "RIANet: Recurrent interleaved attention network for cardiac MRI segmentation," *Comp. in Bio. and Med.*, vol. 109, pp. 290–302, 2019.

[65] H. Wei *et al.*, "Left ventricle segmentation and quantification with attention-enhanced segmentation and shape correction," in *Proc. of the Intl. Symp. on Image Computing and Digital Medicine*, 2019, pp. 226–230.

[66] F. Zabihollahy, J. A. White, and E. Ukwatta, "Myocardial scar segmentation from magnetic resonance images using convolutional neural network," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575. International Society for Optics and Photonics, 2018, p. 105752Z.

[67] M. C. Carminati, C. Boniotti, L. Fusini, D. Andreini, G. Pontone, M. Pepi, and E. G. Caiani, "Comparison of image processing techniques for nonviable tissue quantification in late gadolinium enhancement cardiac magnetic resonance images," *Journal of thoracic imaging*, vol. 31, no. 3, pp. 168–176, 2016.

[68] A. S. Fahmy, J. Rausch, U. Neisius, R. H. Chan, M. S. Maron, E. Appelbaum, B. Menze, and R. Nezafat, "Automated cardiac mr scar quantification in hypertrophic cardiomyopathy using deep convolutional neural networks," *JACC: Cardiovascular Imaging*, vol. 11, no. 12, pp. 1917–1918, 2018.

[69] S. Moccia, R. Banali, C. Martini, G. Muscogiuri, G. Pontone, M. Pepi, and E. G. Caiani, "Development and testing of a deep learning-based strategy for scar segmentation on cmr-lge images," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 32, no. 2, pp. 187–195, 2019.

[70] C. Xu, L. Xu, Z. Gao, S. Zhao, H. Zhang, Y. Zhang, X. Du, S. Zhao, D. Ghista, H. Liu *et al.*, "Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture," *Medical image analysis*, vol. 50, pp. 82–94, 2018.

[71] D. F. Pace, A. V. Dalca, T. Geva, A. J. Powell, M. H. Moghari, and P. Golland, "Interactive whole-heart segmentation in congenital heart disease," ser. Lecture Notes in Computer Science, vol. 9351. Springer, 2015, pp. 80–88.

[72] X. H. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Medical Image Analysis*, vol. 31, pp. 77–87, 2016.

[73] Z. Xiong, Q. Xia, Z. Hu, N. Huang, C. Bian, Y. Zheng, S. Vesal, N. Ravikumar, A. Maier, X. Yang *et al.*, "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical Image Analysis*, vol. 67, p. 101832, 2021.

[74] W. F. Xue, A. Lum, A. Mercado, M. Landis, J. Warringto, and S. Li, "Full quantification of left ventricle via deep multitask learning network respecting intra- and inter-task relatedness," ser. Lecture Notes in Computer Science, vol. 10435. Springer, 2017, pp. 276–284.

[75] W. F. Xue, G. Brahm, S. Pandey, S. Leung, and S. Li, "Full left ventricle quantification via deep multitask relationships learning," *Medical Image Analysis*, vol. 43, pp. 54–65, 2018.

[76] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[77] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[78] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[79] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[80] 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*, 2016, pp. 770–778.

[81] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, 2000.

[82] A. V. Fiacco and J. Liu, "Neural networks using a logistics sigmoid function: linear classifier bounds and global nonattainability," *Optimization*, vol. 32, no. 4, pp. 351–358, 1995.

[83] G. A. Anastassiou, "Univariate hyperbolic tangent neural network approximation," *Mathematical and Computer Modelling*, vol. 53, no. 5-6, pp. 1111–1132, 2011.

[84] 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.

[85] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1. Citeseer, 2013, p. 3.

[86] V. Avinash Sharma, "Understanding activation functions in neural networks," *Machine Learning Mastery. Available at https://medium. com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0*, 2017.

[87] S. Sharma and S. Sharma, "Activation functions in neural networks," *Towards Data Science*, vol. 6, no. 12, pp. 310–316, 2017.

[88] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[89] T. Tieleman and G. Hinton, "Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning," *Technical Report.*, 2017.

[90] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*. PMLR, 2013, pp. 1139–1147.

[91] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[92] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[93] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *arXiv preprint arXiv:1805.07836*, 2018.

[94] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[95] Y. S. Aurelio, G. M. de Almeida, C. L. de Castro, and A. P. Braga, "Learning from imbalanced data sets with weighted cross-entropy function," *Neural Processing Letters*, vol. 50, no. 2, pp. 1937–1949, 2019.

[96] 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," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

[97] N. Ketkar, "Stochastic gradient descent," in *Deep learning with Python*. Springer, 2017, pp. 113–132.

[98] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research*, vol. 12, no. 7, 2011.

[99] N. D. Marom, L. Rokach, and A. Shmilovici, "Using the confusion matrix for improving ensemble classifiers," in *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*. IEEE, 2010, pp. 000 555–000 559.

[100] T. Calders and S. Jaroszewicz, "Efficient auc optimization for classification," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2007, pp. 42–53.

[101] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[102] N. N. Prasad and J. N. Rao, "The estimation of the mean squared error of small-area estimators," *Journal of the American statistical association*, vol. 85, no. 409, pp. 163–171, 1990.

[103] Y. LeCun, D. Touresky, G. Hinton, and T. Sejnowski, "A theoretical framework for back-propagation," in *Proceedings of the 1988 connectionist models summer school*, vol. 1, 1988, pp. 21–28.

[104] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential data augmentation techniques for medical imaging classification tasks," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 979.

[105] R. Moore and J. DeNero, "L1 and l2 regularization for multiclass hinge loss models," in *Symposium on machine learning in speech and language processing*, 2011.

[106] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[107] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[108] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, no. 6, 2013, pp. 380–384.

[109] J. Serra, "Hausdorff distances and interpolations," *Computational Imaging and Vision*, vol. 12, pp. 107–114, 1998.

[110] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," *Advances in neural information processing systems*, vol. 23, pp. 1243–1251, 2010.

[111] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[112] P. Zhang, W. Liu, H. Wang, Y. Lei, and H. Lu, "Deep gated attention networks for large-scale street-level scene segmentation," *Pattern Recognition*, vol. 88, pp. 702–714, 2019.

[113] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[114] Y. Wang, Z. Deng, X. Hu, L. Zhu, X. Yang, X. Xu, P.-A. Heng, and D. Ni, "Deep attentional features for prostate segmentation in ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 523–530.

[115] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[116] J. Tian, K. Wu, K. Ma, H. Cheng, and C. Gu, "Exploration of different attention mechanisms on medical image segmentation," in *International Conference on Neural Information Processing*. Springer, 2019, pp. 598–606.

[117] A. Papadopoulos, P. Korus, and N. Memon, "Hard-attention for scalable image classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 694–14 707, 2021.

[118] G. Elsayed, S. Kornblith, and Q. V. Le, "Saccader: improving accuracy of hard attention models for vision," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[119] B. Uzkent and S. Ermon, "Learning when and where to zoom with deep reinforcement learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 345–12 354.

[120] A. Katharopoulos and F. Fleuret, "Processing megapixel images with deep attention-sampling models," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3282–3291.

[121] A. Torralba, "Contextual priming for object detection," *International Journal on Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[122] L. Wang, D. Nie, G. Li, Élodie Puybareau *et al.*, "Benchmark on automatic 6-month-old infant brain segmentation algorithms: The iSeg-2017 challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2219–2230, 2019.

[123] H. J. Kuijf *et al.*, "Standardized assessment of automatic segmentation of white matter hyperintensities: Results of the WMH segmentation challenge," *IEEE Transactions on Medical Imaging*, pp. 1–13, 2019, available as 'Early access'.

[124] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[125] Y. Xu, T. Géraud, and I. Bloch, "From neonatal to adult brain MR image segmentation in a few seconds using 3D-like fully convolutional network and transfer learning," in *Proc. of IEEE Intl. Conf. on Image Processing (ICIP)*, 2017, pp. 4417–4421.

[126] E. Puybareau, Z. Zhao, Y. Khoudli, E. Carlinet, Y. Xu, J. Lacotte, and T. Géraud, "Left atrial segmentation in a few seconds using fully convolutional network and transfer learning," ser. Lecture Notes in Computer Science, vol. 11395. Springer, 2018, pp. 339–347.

[127] L. Yu, X. Yang, J. Qin, and P.-A. Heng, "3d fractalnet: dense volumetric segmentation for cardiovascular mri volumes," in *Reconstruction, segmentation, and analysis of medical images*, ser. Lecture Notes in Computer Science. Springer, 2016, vol. 10129, pp. 103–110.

[128] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease," in *Reconstruction, Segmentation, and Analysis of Medical Images*, ser. Lecture Notes in Computer Science, vol. 10129. Springer, 2017, pp. 95–102.

[129] J. Li, R. Z. Zhang, L. Shi, and D. Wang, "Automatic whole-heart segmentation in congenital heart disease using deeply-supervised 3D FCN," in *Reconstruction, Segmentation, and Analysis of Medical Images*, ser. Lecture Notes in Computer Science, vol. 10129. Springer, 2017, pp. 111–118.

[130] R. Shahzad, S. Gao, Q. Tao, O. Dzyubachyk, and R. van der Geest, "Automated cardiovascular segmentation in patients with congenital heart disease from 3D CMR scans: Combining multi-atlases and level-sets," in *Reconstruction, Segmentation, and Analysis of Medical Images*, ser. Lecture Notes in Computer Science, vol. 10129. Springer, 2017, pp. 147–155.

[131] G. Tziritas, "Fully-automatic segmentation of cardiac images using 3D MRF model optimization and substructures tracking," in *Reconstruction, Segmentation, and Analysis of Medical Images*, ser. Lecture Notes in Computer Science, vol. 10129. Springer, 2017, pp. 129–136.

[132] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, ser. Lecture Notes in Computer Science, vol. 9901. Springer, 2016, pp. 424–432.

[133] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, vol. 170, pp. 446–455, 2018.

[134] M. P. Heinrich and J. Oster, "MRI whole heart segmentation using discrete nonlinear registration and fast non-local fusion," ser. Lecture Notes in Computer Science, vol. 10663. Springer, 2017, pp. 233–241.

[135] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Multi-label whole heart segmentation using CNNs and anatomical label configurations," ser. Lecture Notes in Computer Science, vol. 10663. Springer, 2017, pp. 190–198.

[136] Z. Shi, G. Zeng, L. Zhang, X. Zhuang, L. Li, G. Yang, and G. Zheng, "Bayesian voxdrn: A probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3d mr images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 569–577.

[137] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "3d convolutional networks for fully automatic fine-grained whole heart partition," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, ser. Lecture Notes in Computer Science, vol. 10663. Springer, 2017, pp. 181–189.

[138] X. Zhuang, L. Li, C. Payer, D. Štern, M. Urschler, M. P. Heinrich, J. Oster, C. Wang, Ö. Smedby, C. Bian *et al.*, "Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge," *Medical image analysis*, vol. 58, p. 101537, 2019.

[139] L.-C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of ECCV*, 2018, pp. 801–818.

Part VI

Appendices

Appendix A

Publication List

Some of the materials presented in this manuscript have been published in peer-reviewed conferences.

- **Zhao Z.**, et al. *Multi-purpose Tactile Perception Based on Deep Learning in a New Tendon-driven Optical Tactile Sensor*[C]// 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). DOI: 10.1109/IROS47612.2022.9981477
- **Zhao Z.**, et al. *FOANet: A Focus of Attention Network with Application to Myocardium Segmentation*[C]// 2020 25th International Conference on Pattern Recognition (ICPR). DOI: 10.1109/ICPR48806.2021.9412016
- **Zhao Z.**, et al. *Do not Treat Boundaries and Regions Differently: An Example on Heart Left Atrial Segmentation*[C]// 2020 25th International Conference on Pattern Recognition (ICPR). (**Oral paper**) DOI: 10.1109/ICPR48806.2021.9412755
- **Zhao Z.**, et al.. *Stacked and Parallel U-Nets with Multi-output for Myocardial Pathology Segmentation*. Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images. MyoPS 2020. DOI: 10.1007/978-3-030-65651-5_13
- **Zhao Z.**, et al.. *A Two-Stage Temporal-Like Fully Convolutional Network Framework for Left Ventricle Segmentation and Quantification on MR Images*. Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges. STACOM 2019. (**Oral paper**) DOI: 10.1007/978-3-030-39074-7_42
- Puybareau É., **Zhao Z.**. et al. *Left Atrial Segmentation in a Few Seconds Using Fully Convolutional Network and Transfer Learning*. Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. STACOM 2018. DOI: 10.1007/978-3-030-12029-0_37
- Li, L., Wu, F. P.. et al. *MyoPS: A Benchmark of Myocardial Pathology Segmentation Combining Three-Sequence Cardiac Magnetic Resonance Images*. arXiv preprint arXiv:2201.03186.

Some honors about deep learning competitions have been awarded by the proposed method in this manuscript

- MICCAI 2019 Left Ventricle Full Quantification Challenge (**The third prize**)
- MICCAI 2018 Atrial Segmentation Challenge (**The third prize**)

Some work has been already submitted to journals.

- **Zhou Zhao**, Nicolas Boutry, Élodie Puybareau, Thierry Géraud. *A Multi-class Hybrid Loss Function to Handle Ambiguities in Biomedical MRI Images* (submitted to international journal)
- **Zhou Zhao**, Élodie Puybareau, Nicolas Boutry, Thierry Géraud. *A^0Net: Single-minded Attention Network with Application to Myocardium Segmentation* (submitted to international journal)

Some works presented in this thesis are in preparation for journal papers.

FOANet: A Focus of Attention Network with Application to Myocardium Segmentation

Zhou Zhao, Élodie Puybareau, Nicolas Boutry, Thierry Géraud
EPITA Research and Development Laboratory (LRDE), Le Kremlin-Bicêtre, France
Email: elodie.puybareau@lrde.epita.fr

***Abstract*—In myocardium segmentation of cardiac magnetic resonance images, ambiguities often appear near the boundaries of the target domains due to tissue similarities. To address this issue, we propose a new architecture, called FOANet, which can be decomposed in three main steps: a localization step, a Gaussian-based contrast enhancement step, and a segmentation step. This architecture is supplied with a hybrid loss function that guides the FOANet to study the transformation relationship between the input image and the corresponding label in a three-level hierarchy (pixel-, patch- and map-level), which is helpful to improve segmentation and recovery of the boundaries. We demonstrate the efficiency of our approach on two public datasets in terms of regional and boundary segmentations.**

I. Introduction

In order to accurately segment the myocardium in cardiac magnetic resonance (MR) images, numerous methods have been developed by world-wide researchers. Among these methods, the most common method is atlas-based, which offers good accuracy for myocardium segmentation, but often looses efficiency due to heavy calculations with the registration algorithm. Recently, methods based on deep learning are replacing the conventional methods in the field of myocardium segmentation. For example, Zabihollahy et al. [1] proposed a novel method to segment myocardium using a U-Net convolutional neural network (CNN)-based model, and the algorithm-generated results demonstrated its usefulness for myocardium segmentation. Do et al. [2] proposed a network architecture of Monte Carlo dropout (MCD) UNet for myocardium segmentation, and the MCD was mainly applied to measure a global score of model uncertainty without using the reference segmentation, which was valuable for automatic quality control at production. Dangi et al. [3] proposed a multi-task learning (MTL)-based regularization of a CNN, and used the rich information available in the distance map of the segmentation mask as an auxiliary task for the myocardium segmentation network. Since each pixel in the distance map represented its distance from the closest object boundary, which was more redundant and robust than the per-pixel image label directly used for segmentation. Furthermore, the distance map contained the shape and boundary information of the object. Therefore, predicting the distance map, as an additional task, was beneficial to enforce shape and boundary constraints during the process of training.

However, there are many difficulties to segment myocardium from cardiac MR images, for example, the presence of poor contrast between the segmented tissue and surrounding

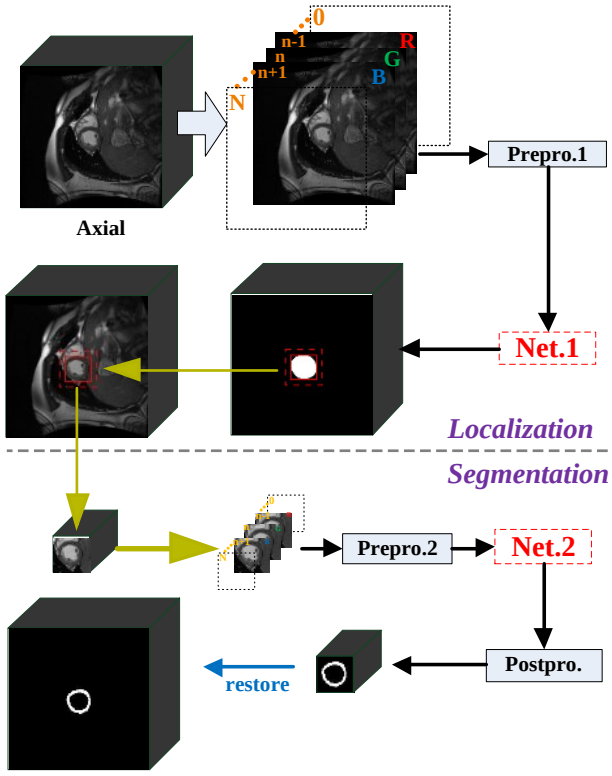


Fig. 1: Global overview of the proposed method (FOANet).

structures, the brightness heterogeneities due to blood flow, the shape and intensity variabilities of the structures across patients and pathologies, and so on [4]. To decrease the effect of blood flow and accurately segment the blood pool and myocardium from cardiac MR, Qi et al. [5] proposed a multi-scale feature fusion (MSFF) CNN with a new weighted dice index loss function to segment myocardium, using MSFF modules to obtain feature maps of different scale, and then concatenating them through short and long skip connections in the encoder and decoder path to capture more complete context information and geometry structure for better segmentation. To capture the valuable dynamics of heart motion, Zhang et al. [6] proposed a method based on recurrent neural network (RNN), in order to take the motion of the heart into consideration, and extract myocardium-related image features at both the low- and high resolution levels in consecutive frames of a cardiac cycle. Faced with variability in contrast, appearance, orien-

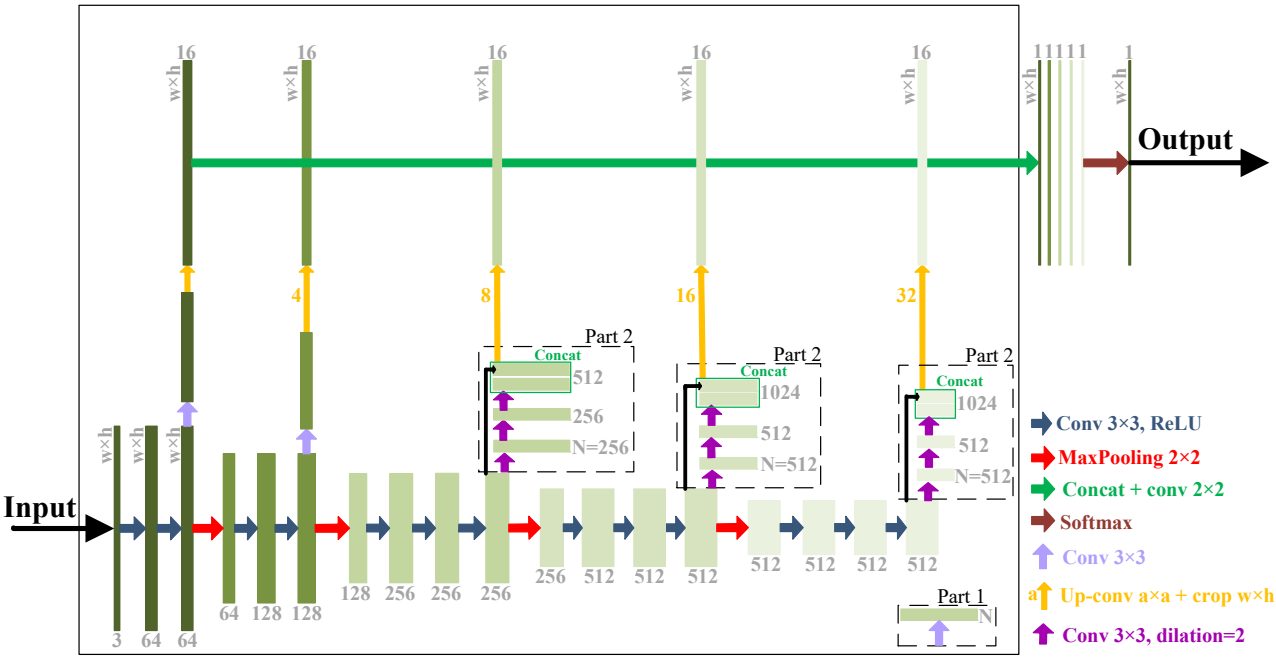


Fig. 2: Architecture of our networks. **Part 1** and **Part 2** correspond to the components of **Net.1** and **Net.2** of Fig. 1, respectively. Because the role of **Net.1** is only to roughly locate the target, using **Part 1** instead of **Part 2** can both reduce model parameters and improve the speed of model prediction. N denotes the number of feature map

tation, and placement of the heart between patients, clinical views, scanners, and protocols, Davis et al. [7] proposed a fully automatic semantic segmentation method: Omega-Net that included three steps to segment, first, roughly located the object on the input image; second, learned the features based on the obtained object during the first step, which is used to predict the parameters needed to transform the input image into a canonical orientation; and third, the transformed image from the second step is used to finally segment. Despite the fact that these methods continue to improve segmentation accuracy, a large number of mis-segmentations still exist, which is due to the fact that they mainly pay attention to region accuracy, more than to the quality of the boundaries. However, issues often occur at indistinguishable boundaries. To maintain region accuracy without losing the boundary quality, we propose a focus of attention architecture that we call *FOANet*, and a new hybrid loss for region- and boundary-aware segmentation. The main contributions of our work are:

— A novel region- and boundary-aware segmentation network, FOANet, which consists of a localization and a segmentation parts.
— A novel hybrid loss that combines Categorical Cross Entropy (CCE), Structural Similarity (SSIM) and Dice Coefficient (DC) to guide the training process at three levels: pixel-level, patch-level, and map-level.
— A novel Focus of Attention (FOA) that decreases the impact of surrounding similar tissues.
— A temporal-like method that lets the FOANet take advantage of the temporal information by stacking 3 successive 2D frames.

II. Methodology

A. Overview of Network Architecture

The global overview of our FOANet consists of two parts (localization and segmentation) as depicted in Fig. 1, and the architecture of our networks in Fig. 2. The first part (the "localization network") is used to localize roughly the object position. The second part is devoted to segment the object (the "segmentation network").

B. Localization Network

The localization network (**Net.1**) is depicted in Fig. 2. The black dotted box **Part 1** is dedicated to the localization network, it can be replaced by **Part 2** to become the segmentation network (**Net.2**). For **Net.1** and **Net.2**, the difference concerns only **Part 1** and **Part 2** as shown in Fig. 2, while the other components of the architecture are the same. **Part 1** consists of one convolutional layers with 256 or 512. First, we rely on the original VGG16 [8] network architecture, pre-trained on millions of natural images of ImageNet for image classification [9]. We then discard its fully connected layers to keep only the sub-network made of five convolution-based "stages" (the base network). Each stage is made of two convolutional layers, a ReLU activation function, and a max-pooling layer. Since the max-pooling layers decrease the resolution of the input image, we obtain a set of fine to coarse feature maps (with 5 levels of features). Inspired by the works in [10, 11, 12, 13], we added *specialized* convolutional layers

(with a 3×3 kernel size) with K (*e.g.* $K = 16$) feature maps after the up-convolutional layers placed at the end of each stage. The outputs of the specialized layers show the same resolution than the input image, and are concatenated together. We add a 1×1 convolutional layer at the output of the concatenation layer to linearly combine the fine to coarse feature maps [1].

C. Segmentation Network

As mentioned above, we replace **Part 1** of **Net.1** with **Part 2**, which becomes the segmentation network (**Net.2**). Because the role of **Net.2** is mainly to obtain accurate segmentation results, we use **Part 2** that is more complicated than **Part 1** in Fig. 2. It can capture the global information and decrease the effect of surrounding similar tissues. **Part 2** consists of three convolutional layers with 256 or 512 dilated (dilation = 2) [14] 3×3 filters, and one layer of concatenation.

D. Hybrid Loss

To obtain high quality regional segmentation and nice boundaries, we define ℓ as a hybrid loss: $\ell = \ell_{\text{CCE}} + \ell_{\text{SSIM}} + \ell_{\text{DC}}$, where ℓ_{CCE}, ℓ_{SSIM} and ℓ_{DC} respectively denote CCE loss [15], SSIM loss [16] and DC loss [17] respectively.

CCE [15] loss is commonly used for multi-class classification and segmentation. It is defined as:

$$\ell_{\text{CCE}} = -\sum_{i=1}^{C} \sum_{a=1}^{H} \sum_{b=1}^{W} y^{i}_{(a,b)} \ln {y_*}^{i}_{(a,b)}, \quad (1)$$

where C is the number of classes of each image, H and W are the height and width of image, $y^{i}_{(a,b)} \in \{0, 1\}$ is the ground truth one-hot label of class i in the position (a, b) and ${y_*}^{i}_{(a,b)}$ is the predicted probability of class i.

SSIM loss can assess image quality [16], and can be used to capture the structural information, which will decrease the mis-segmentation rate of surrounding similar tissues. Therefore, we integrated it into our training loss to learn the differences between the segmented domain and similar tissues around the segmented domain. Let **S** and **G** be the predicted probability map and the ground truth mask respectively, the SSIM of **S** and **G** is defined as:

$$\ell_{\text{SSIM}} = 1 - \frac{(2\mu_{\text{S}}\mu_{\text{G}} + C_1)(2\sigma_{\text{SG}} + C_2)}{(\mu_{\text{S}}^2 + \mu_{\text{G}}^2 + C_1)(\sigma_{\text{S}}^2 + \sigma_{\text{G}}^2 + C_2)}, \quad (2)$$

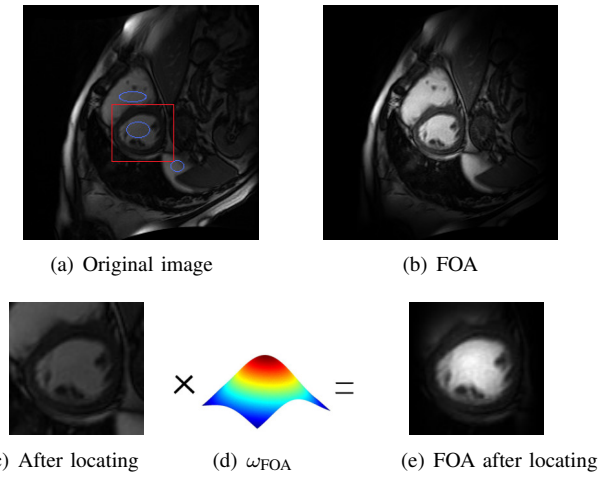
where μ_{S}, μ_{G} and σ_{S}, σ_{G} are the mean and standard deviations of **S** and **G** respectively, σ_{SG} is their covariance, $C_1 = 0.01^2$ and $C_2 = 0.03^2$ are used to avoid a division by zero.

DC [17] loss is used to measure the similarity between two sets as defined in Eq. 3. But for the multi-class segmentation task, Eq. 3 is not suitable due to the class imbalance problem in such cases. Therefore, we extend the definition of the DC loss to multiclass segmentation in the following manner:

$$dice_i = (\epsilon + 2\sum_{n=1}^{N_i} y^{i}_{n}\, {y_*}^{i}_{n}) \,/\, (\epsilon + \sum_{n=1}^{N_i} (y^{i}_{n} + {y_*}^{i}_{n})) \quad (3)$$

$$\ell_{\text{DC}} = 1 - \sum_{i=1}^{C} dice_i / (N_i + \epsilon), \quad (4)$$

[1]Note that we designed our network's architecture to work with any input shape.



(a) Original image (b) FOA (c) After locating (d) ω_{FOA} (e) FOA after locating

Fig. 3: Focus of attention (FOA).

where N_i denotes the numbers of class i and ϵ is a smooth factor.

E. Focus of Attention

The image of Fig. 3a is from the MICCAI 2019 left ventricle (LV) Full Quantification Challenge dataset[2] (LVQuan19) [18, 19]. The red box denotes the object domain, here the LV. There are a large number of similar tissues around it, highlighted by the blue ellipses. Even after a localization procedure, these tissues are still present. To decrease the impact of similar tissues on segmentation results, we built on the biological visual system, which concentrates on certain image regions requiring detailed analysis [20]. We define the FOA as: $I_{\text{FOA}}(a, b) = I(a, b)\omega_{\text{FOA}}(a, b)$, where $I(a, b)$ denotes the image intensity at location (a, b) and $\omega_{\text{FOA}}(a, b)$ is a Gaussian weighted function defined by

$$\omega_{\text{FOA}}(a, b) = \alpha \exp(-|(a, b) - (a^*, b^*)|^2 / \delta^2), \quad (5)$$

where (a^*, b^*) denotes the object center, α is a normalization constant, δ is a scale parameter.

If we used $I_{\text{FOA}}(a, b)$ on each original image, we would probably miss the object of interest. Therefore, we must first localize the domain of interest; then we use $I_{\text{FOA}}(a, b)$ to focus on the object. This methodology is depicted in Fig. 3e, where similar tissues are less visible when compared to Fig. 3c.

III. Experimental Results

A. Dataset Description

We evaluated our method on two datasets: LVQuan19 and Multi-Modality Whole Heart Segmentation [3] (MM-WHS2017). The aim of **LVQuan19** is to segment the myocardium of the left ventricle and estimate a set of clinical significant LV indices such as regional wall thicknesses, cavity dimensions, and cardiac phase and so on. It contains the processed SAX MR sequences of 56 patients. For each patient, 20 temporal frames are given and cover a whole cardiac cycle.

[2]https://lvquan19.github.io

[3]http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs17/index.html

All ground truth (GT) values of the LV indices are provided for every single frame. The pixel spacings of the MR images range from 0.6836 mm/pixel to 1.5625 mm/pixel, with mean values of 1.1809 mm/pixel. The LV dataset includes two different image sizes: 256×256 or 512×512 pixels. **MM-WHS2017** [21] aims to segment 7 substructures of the whole heart. Although it contains 20 cardiac MRI and 20 CT images, we only use the MRI modality. The slice spacings of MRI volume range from 0.899 mm/pixel to 1.60 mm/pixel, while in-plane resolution ranged from 0.78 mm/pixel to 1.2 mm/pixel. The average sizes: $324 \times 325 \times 171$ pixels.

B. Preprocessings

Since the VGG-16 network's input is an RGB image, we propose to take advantage of the temporal information by stacking 3 successive 2D frames: to segment the n^{th} slice, we use the n^{th} slice of the MR volume, and its neighboring $(n-1)^{th}$ and $(n+1)^{th}$ slices, as green, red and blue channels, respectively. This new image, named "temporal-like" image, enhances the area of motions, here the heart, as shown in Fig. 4.

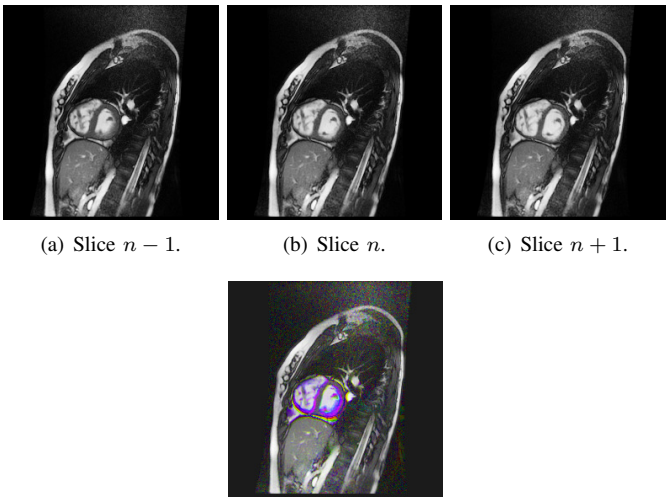
Let us remind what we call *Gauss normalization*: for each $(2D+t)$-image I corresponding to a given patient, we compute $I := (I - \mu)/\sigma$ where μ is the mean of I and σ its standard deviation (σ is assumed not to be equal to zero). There are then two different pre-processing steps as depicted in Fig. 1.

1) The first pre-processing (see **Prepro.1** in Fig. 1) begins with a Gauss normalization. Then, for each n, we created the $width \times height \times 3$ pseudo-color ("temporal-like") image where R, G, B correspond respectively to the $n-1, n, n+1$ frames and we concatenate them (we do not detail the cases $n = 1$ and $n = n_{end}$, the first and last slice of the volume, because of lack of space).

2) The second pre-processing (**Prepro.2** in Fig. 1) follows five steps: (1) data augmentation using rotations and flips for the LVQuan19 dataset (only for the training phase), but it is not used on the MM-WHS2017 dataset, (2) resizing with a fixed pixel-spacing ($0.65mm$), (3) FOA, (4) Gauss normalization, and (5) pseudo-color concatenated image like above. Such a use of a pseudo-color image in the context of 3D medical imaging has been proven effective in [22] to segment brain structures and in [23] to extract white matter hyperintensities in brain volumes.

C. Postprocessing

Let us assume that we crop an initial volume of T frames of size $T \times W \times H$ into an image of size $T \times w \times h$ (where the crop is due to the localization procedure, and W and H are the initial width and height of a slice). After **Prepro.2** we obtain a $T \times w \times h \times 3$ image. Then we filter the ouput of the segmentation network, of size $T \times w \times h$, by keeping only the greatest connected component, in order to get back the initial pixel-spacing. Finally, we add a padding of zeros to get back a $T \times W \times H$ image.



(a) Slice $n-1$. (b) Slice n. (c) Slice $n+1$.

(d) RGB concatenation at n.

Fig. 4: Illustration of our "temporal-like" procedure.

D. Implementation and Experimental Setup

We implemented our experiments on Keras/TensorFlow using a NVidia Quadro P6000 GPU. For the localization network, we used the multinomial logistic loss function for a one-of-many classification task, passing real-valued predictions through a softmax to get a probability distribution over classes. We used an Adam optimizer (batchsize = 1, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 0.001$, $\mathrm{lr} = 0.002$) and we did not use learning rate decay. We trained the network during 10 epochs. For this step, we merged all the classes into the object class to obtain a binary segmentation. For the segmentation network, we used the same optimizer and parameters detailed previously. We used the hybrid loss as loss function. For this task, we considered three different classes (background, myocardium, cavity) for LVQuan19 and eight different classes (background, myocardium, left atrium, left ventricle, right atrium, right ventricle, ascending aorta and pulmonary artery) for MM-WHS2017.

E. Evaluation Methods

Three measures are used to evaluate our method: DC given in Eq. 3, 95% in the Hausdorff distance (95HD) [24] and Boundary of Dice Coefficient (BDC) to quantitatively evaluate the boundaries. As many diseases appear in the myocardium wall, we chose to quantitatively evaluate the precision of the segmentation on boundaries.

For the BDC evaluation method, given a segmentation map M, we first convert the class i to a binary mask, M^i_{bm}. Then, we obtain the mask of class i of its one pixel wide boundary by conducting an XOR(M^i_{bm}, M^i_{erd}) operation where M^i_{erd} is the eroded binary mask of M^i_{bm}. The same method is used to get the GT mask boundaries, M^i_{g}. Then the DC is calculated on the boundaries of the GT and segmentation masks to obtain the BDC.

TABLE I: Ablation study; Dice values are for the myocardium.

Ablation	Configurations	DC	95HD	BDC
Architecture	a: B. + ℓ_{CCE}	0.842	3.186	0.269
	b: B. + L. + ℓ_{CCE} [13]	0.867	2.209	0.281
	c: BLP + ℓ_{CCE}	0.877	2.019	0.303
Loss	d: BLP + ℓ_{SSIM}	0.873	2.094	0.297
	e: BLP + ℓ_{DC}	0.871	2.193	0.295
FOA (our)	i: BLP + FOA + ℓ_{CSD}	**0.879**	**1.826**	**0.306**
UNet [25]	-	0.862	3.976	0.291

"B." means "baseline" (**Net.1**) [26, 27]; "L." means "localization"; "P2." means "Part 2"(**Net.2**); "BLP" means "baseline + localization + Part2".
Note: $\ell_{CSD} = \ell_{CCE} + \ell_{SSIM} + \ell_{DC}$

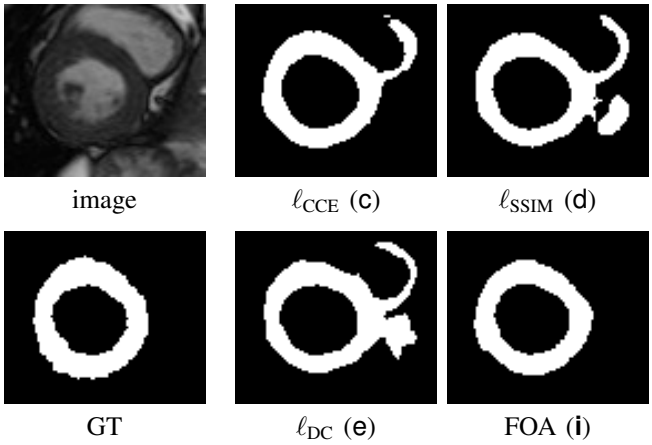


Fig. 5: The comparative results trained with our FOANet on different losses.

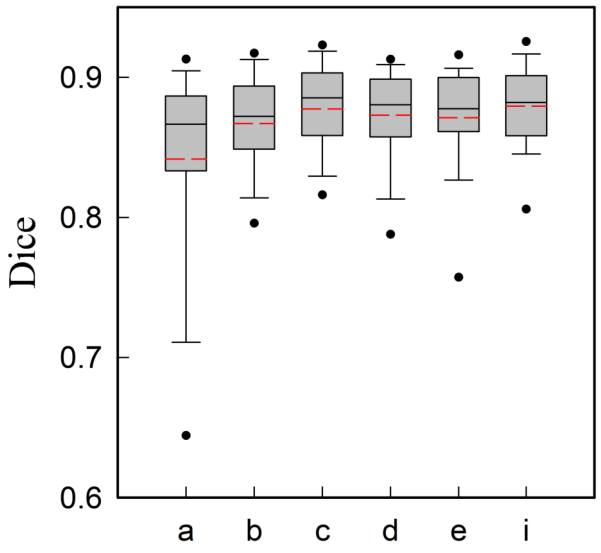


Fig. 6: Box plots of dice scores for the 56 patients. The red dotted line represents the average value, and a, b, c, etc. on the abscissa correspond to the methods of Tbl. I

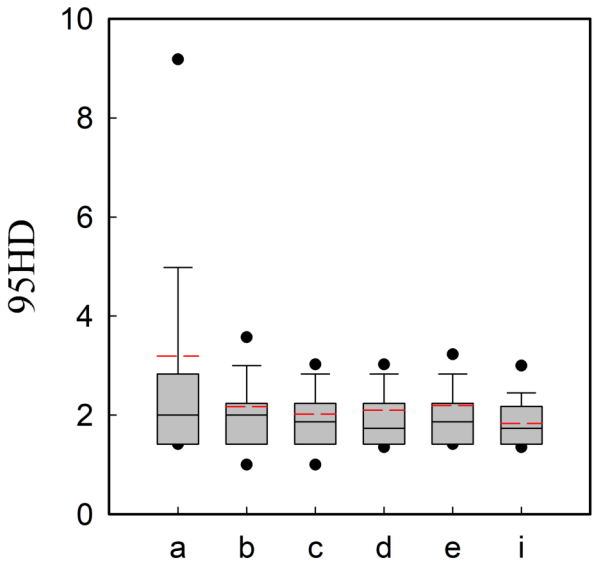


Fig. 7: Box plots of 95HD for the 56 patients. The red dotted line represents the average value, and a, b, c, etc. on the abscissa correspond to Tbl. I

F. Ablation Study

To validate the influence of each component used in our method, we conducted the ablation study that includes three parts (architecture, loss and FOA) on the LVQuan19 dataset with 5-fold cross-validation. Results are shown in Tbl. I. **Architecture ablation**: To demonstrate the effects of our FOANet, we compared the results of our method with other related frameworks. We took a network used in our previous works [26, 27] as baseline network (**Net.1**). First, we added a localization module (as shown in Fig. 1) based on the baseline; with this module, we obtained a mean improvement of 1.89% in terms of DC, 0.9772 on 95HD, which meant that reducing the proportion of the background in the image is beneficial to improve segmentation accuracy. This architecture was the one we presented for the Challenge LVQUAN19 [13]. Further, we added the **Part 2** module, so **Net.1** was changed to **Net.2** (Baseline+Part2) as shown in Fig. 2. We learned from our comparison results that, when using dilated convolution and capturing the global information in the feature maps of high level, we could refine the segmentation results, which meant further improvement of 1.70% in terms of DC, 0.1893 on 95HD. **Loss ablation**: To prove the effects of our hybrid loss, we conducted comparative experiments over different losses based on our method. The results in Tbl. I illustrate that the proposed hybrid loss helps to improve the performance, and, compared with other combinations, that loss function based on three-level hierarchy (pixel-, patch- and map-level) can fully guide the network to study the transformation relationship between the input image and the corresponding label. **FOA ablation**: As shown in Fig. 5, without FOA, the surrounding similar tissues are mis-segmented, meaning that the segmentation results are disturbed by these similar tissues, and mis-

TABLE II: Comparison of our method and other challengers on the MM-WHS2017 MRI training dataset for segmenting the myocardium.

Method	DC (train)	DC (test)	Computation time	Data augmentation
Our (best)	0.851	?	$<$ 2s	No
Best [28]	0.796	0.781	$<$ 2min	No
Second-best [29]	0.752	0.778	-	Yes
UB2 [30]	?	0.811	?	Yes

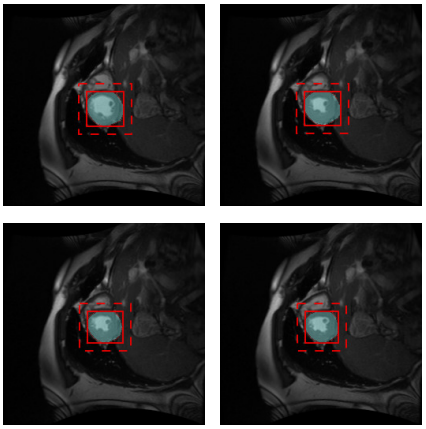
segmented parts are connected to the ground truth, which is very difficult to remove. Therefore, by using our FOA module, we decrease the impact of the surrounding similar tissues, and the segmentation results are better.

Statistical analysis Fig. 6 shows the box plots of the evaluation on different framework configurations for dice scores. Compared with others configurations, the segmentation results obtained by our method (configuration:**i**) have a small standard deviation, which shows that our method is more stable on region segmentation. Fig. 7 shows the box plots of the evaluation for 95HD. Compared with others configurations, based on the median quantile of box plots and the average of 56 patients, most of the values of our method are low, which shows that our method optimizes the boundary quality.

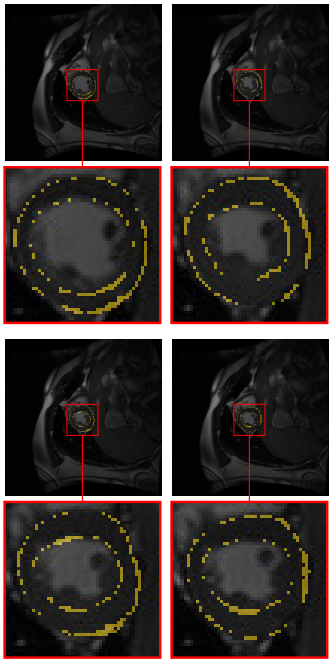
Fig. 8 shows several localization and segmentation results of our FOANet on LVQuan19. Fig. 8a indicates that we started with finding the smallest rectangular box for each slice of the patient's heart, ensuring that each box contained the segmentation object. Then we found the biggest rectangular box on the basis of these smallest rectangular boxes; and based on its shape, we cropped a new 3D volume from the original 3D volume as shown in the segmentation module of Fig. 1. Thanks to the localization results of Fig. 8a, we knew that the object was contained in/by the box, which greatly increased the proportion of objects in the image and reduced class imbalance. Fig. 8b compares ground truth and prediction, and we can see that the differences mainly are near the boundaries.

G. Comparison with State-of-the-Art Methods

We continued to test our method on the MM-WHS2017 challenge with 5-fold cross-validation and we obtained segmentation results for each class. As we focus in this article on the myocardium segmentation, we will only present our results for this structure. For the comparison with state-of-the-art methods, we choose to compare our results with the results of the first and second prizes of the challenge, who respectively get dices of 0.87 and 0.863 in average for all classes. We reported their results on the training and on the testing sets. We also add a comparison with a late submission on the testing set only (scores on the training set are not available), having the best actual score of the challenge [30, 31]. As shown in Tbl. II, compared with the first and second prizes of the MM-WHS2017 challenge, without using data augmentation, our method outperformed them for the segmentation of the myocardium of the left ventricle. Furthermore, our method needs less time to compute the prediction, which further



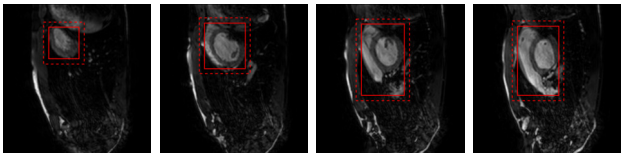
(a) Some localizations of the LV (in blue) of the 9^{th} patient. The red dotted box denotes that we extend next to the box by a size equal to 10 pixels to ensure that the whole LV is included into the bounding box.



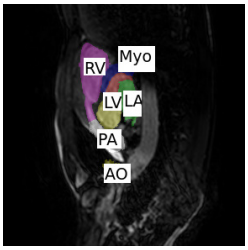
(b) Different comparisions between ground truth and prediction corresponding to (a); yellow denotes the difference.

Fig. 8: Localization and segmentation of our FOANet on LVQuan19.

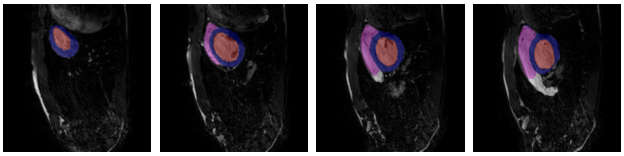
validates the results in LVQuan19. We are still waiting for the quantitative results on the testing dataset to be able to compare our method fairly with [30]. Fig. 9 shows some localization and segmentation results. Concerning the whole heart segmentation, the class imbalance causes a lot of damage without the localization module, because the seven structures of the heart do not always appear at the same time in a slice of the same 3D volume of a same patient. Without the FOA module and **Part 2**, the network can confuse one class with another: the RA can be confused with the RV, the LV can be confused with the LA, and so on. Accordingly, a good segmentation requires to capture the global information by



(a) Some localization results in one patient.



(b) Seven structures of the whole heart. Myo: myocardium, LA: left atrium, LV: left ventricle, RA: right atrium, RV: right ventricle, AO: ascending aorta, PA: pulmonary artery.



(c) Some segmentation results in one patient corresponding to (a).

Fig. 9: Localization and segmentation of our FOANet on MM-WHS2017.

dilated convolutions and to enhance contrast using the FOA module.

IV. Conclusion

In this paper, we propose a new focus of attention network framework, FOANet, and present a new hybrid loss for boundary-aware segmentation. FOANet is able to prevent the interferences of surrounding similar tissues, while the hybrid loss guides it at several levels. Both generate a better capture not only of large-scale information but also of fine structures to produce segmentations with nice boundaries. The computation time of the entire pipeline is less than 2 seconds for an entire 3D volume, making it usable for clinical practice. In our future work, we will continue to study the impact of the hybrid loss by weighting differently the segmentation loss and the boundary loss. Furthermore, we will add constraints on shapes in the network.

References

[1] F. Zabihollahy, J. A. White, and E. Ukwatta, "Fully automated segmentation of left ventricular myocardium from 3d late gadolinium enhancement magnetic resonance images using a u-net convolutional neural network-based model," in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950, 2019, p. 109503C.

[2] H. P. Do, Y. Guo, A. J. Yoon, and K. S. Nayak, "Accuracy, uncertainty, and adaptability of automatic myocardial asl segmentation using deep cnn," *Magnetic Resonance in Medicine*, vol. 83, no. 5, pp. 1863–1874, 2020.

[3] S. Dangi, C. A. Linte, and Z. Yaniv, "A distance map regularized cnn for cardiac cine mr image segmentation," *Medical physics*, vol. 46, no. 12, pp. 5637–5651, 2019.

[4] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structure segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[5] L. Qi, H. Zhang, X. Cao, X. Lyu, L. Xu, B. Yang, and Y. Ou, "Multi-scale feature fusion convolutional neural network for concurrent segmentation of left ventricle and myocardium in cardiac mr images," *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 5, pp. 1023–1032, 2020.

[6] D. Zhang, I. Icke, B. Dogdas, S. Parimal, S. Sampath, J. Forbes, A. Bagchi, C.-L. Chin, and A. Chen, "A multi-level convolutional lstm model for the segmentation of left ventricle myocardium in infarcted porcine cine mr images," in *Proc. of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2018, pp. 470–473.

[7] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "ω-net (omega-net): fully automatic, multi-view cardiac mr detection, orientation, and segmentation with deep neural networks," *Medical Image Analysis*, vol. 48, pp. 95–106, 2018.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR abs/1409.1556, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. of the Intl. Conf. on Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[11] K. K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Deep retinal image understanding," in *Proc. of MICCAI, Part II*, ser. Lecture Notes in Computer Science, vol. 9901, 2016, pp. 140–148.

[12] É. Puybareau, Z. Zhao, Y. Khoudli, E. Carlinet, Y. Xu, J. Lacotte, and T. Géraud, "Left atrial segmentation in a few seconds using fully convolutional network and transfer learning," in *Proc. of the Intl. Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*, ser. Lecture Notes in Computer Science, vol. 11395. Springer, 2018, pp. 339–347.

[13] Z. Zhao, N. Boutry, É. Puybareau, and T. Géraud, "A two-stage temporal-like fully convolutional network framework for left ventricle segmentation and quantification on MR images," in *Proc. of the Intl. Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*, ser. Lecture Notes in Computer Sci-

ence, vol. 12009. Springer, 2019, pp. 405–413.

[14] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1409.1556, 2014.

[15] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. of the Intl. Conf. on Neural Information Processing Systems (NIPS)*, 2018, pp. 8792–8802.

[16] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. of the 37th Asilomar Conference on Signals, Systems and Computers*, vol. 2, 2003, pp. 1398–1402.

[17] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[18] W. F. Xue, A. Lum, A. Mercado, M. Landis, J. Warringto, and S. Li, "Full quantification of left ventricle via deep multitask learning network respecting intra- and inter-task relatedness," in *Proc. of IEEE Intl. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 10435. Springer, 2017, pp. 276–284.

[19] W. F. Xue, G. Brahm, S. Pandey, S. Leung, and S. Li, "Full left ventricle quantification via deep multitask relationships learning," *Medical Image Analysis*, vol. 43, pp. 54–65, 2018.

[20] A. Torralba, "Contextual priming for object detection," *International Journal on Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[21] X. H. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Medical Image Analysis*, vol. 31, pp. 77–87, 2016.

[22] L. Wang, D. Nie, G. Li, Élodie Puybareau *et al.*, "Benchmark on automatic 6-month-old infant brain segmentation algorithms: The iSeg-2017 challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2219–2230, 2019.

[23] H. J. Kuijf *et al.*, "Standardized assessment of automatic segmentation of white matter hyperintensities: Results of the WMH segmentation challenge," *IEEE Transactions on Medical Imaging*, pp. 1–13, 2019, available as 'Early access'.

[24] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.

[26] Y. Xu, T. Géraud, and I. Bloch, "From neonatal to adult brain MR image segmentation in a few seconds using 3D-like fully convolutional network and transfer learning," in *Proc. of IEEE Intl. Conf. on Image Processing (ICIP)*, 2017, pp. 4417–4421.

[27] E. Puybareau, Z. Zhao, Y. Khoudli, E. Carlinet, Y. Xu, J. Lacotte, and T. Géraud, "Left atrial segmentation in a few seconds using fully convolutional network and transfer learning," in *Proc. of the Intl. Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*, ser. Lecture Notes in Computer Science, vol. 11395. Springer, 2018, pp. 339–347.

[28] M. P. Heinrich and J. Oster, "MRI whole heart segmentation using discrete nonlinear registration and fast non-local fusion," in *Proc. of the Intl. Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*, ser. Lecture Notes in Computer Science, vol. 10663. Springer, 2017, pp. 233–241.

[29] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Multi-label whole heart segmentation using CNNs and anatomical label configurations," in *Proc. of the Intl. Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*, ser. Lecture Notes in Computer Science, vol. 10663. Springer, 2017, pp. 190–198.

[30] Z. Shi, G. Zeng, L. Zhang, X. Zhuang, L. Li, G. Yang, and G. Zheng, "Bayesian voxdrn: A probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3d mr images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 569–577.

[31] X. Zhuang, L. Li, C. Payer, D. Štern, M. Urschler, M. P. Heinrich, J. Oster, C. Wang, Ö. Smedby, C. Bian *et al.*, "Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge," *Medical image analysis*, vol. 58, p. 101537, 2019.

Do not Treat Boundaries and Regions Differently: An Example on Heart Left Atrial Segmentation

Zhou Zhao, Élodie Puybareau, Nicolas Boutry, Thierry Géraud
EPITA Research and Development Laboratory (LRDE), Le Kremlin-Bicêtre, France
Email: thierry.geraud@lrde.epita.fr

***Abstract*—Atrial fibrillation is the most common heart rhythm disease. Due to a lack of understanding in matter of underlying atrial structures, current treatments are still not satisfying. Recently, with the popularity of deep learning, many segmentation methods based on fully convolutional networks have been proposed to analyze atrial structures, especially from late gadolinium-enhanced magnetic resonance imaging. However, two problems still occur: 1) segmentation results include the atrial-like background; 2) boundaries are very hard to segment. Most segmentation approaches design a specific network that mainly focuses on the regions, to the detriment of the boundaries. Therefore, this paper proposes an attention full convolutional network framework based on the ResNet-101 architecture, which focuses on boundaries as much as on regions. The additional attention module is added to have the network pay more attention on regions and then to reduce the impact of the misleading similarity of neighboring tissues. We also use a hybrid loss composed of a region loss and a boundary loss to treat boundaries and regions at the same time. We demonstrate the efficiency of the proposed approach on the MICCAI 2018 Atrial Segmentation Challenge public dataset.**

I. Introduction

Segmentation of left atrium in 3D late gadolinium-enhanced magnetic resonance (LGE-MR) images with high precision is a key step for atrial fibrillation (AF) ablation. Although a lot of research has been made on the automation of this task, manual annotations are still commonly used in the medical community, which is highly time-consuming and is subject to inter- and intra-observer variabilities [1]. With the recent development of convolutional neural networks (CNNs), remarkable progress has been made in matter of automatic segmentation [2]. However, the heterogeneity of the features corresponding to a same label may introduce intra-class inconsistencies and affect the accuracy of the segmentation [3]. Although the full convolutional network (FCN) [4] or U-Net [5] architectures can make up for the spatial resolution loss to a certain extent, it performs poorly on small parts of objects. The main issues are then the lack of precision regarding the boundaries of the segmented objects and the loss of small objects and small parts of objects. Therefore, in this paper, we consider two challenging problems applyied on cardiac imaging: 1) how to enlarge the receptive field of a CNN and improve the segmentation accuracy on small parts of objects; 2) how to balance the importance of the regions and the boundaries of objects. Many challenging problems are linked with cardiac imaging: poor contrast between the segmented domain and surrounding structures, heterogeneities in matter of brightness due to the blood flow, non-homogeneous partial volume effects due to limited cardiac magnetic resonance (CMR) resolution (1.5T, 3.0T, *etc.*), and so on [6]. Most of the proposed network frameworks are based on FCN or on U-Net. They use upsampling layers and combine the feature maps from lower to higher resolutions. Many extensions to these networks have been proposed already: Chen [7] proposes a shape-aware multi-view autoencoder (thanks to some modifications to the original U-Net) to achieve high segmentation performance on cardiac magnetic resonance (MR) image segmentation; Khened [8] proposes DenseNet, based on FCNs, for cardiac segmentation and tries to overcome the feature map explosion, but still fails at the boundaries. In fact, the most used loss functions for segmentation network such as dice or cross-entropy (CE) are based on regional integrals, which are convenient for training deep neural networks [9]. However, the CE has well-known drawbacks in the context of highly unbalanced problems, and dice losses may undergo diffculties when dealing with very small structures, and are both region-based. Some methods incorporated boundary information into the loss function. Shen [10] proposes a multi-task FCN architecture where the boundary information is directly incorporated into the loss function, improving its results of segmentation. Kervadec [9] designs one novel boundary loss, and combines it with the standard regional losses, improving the boundary accuracy without losing the region one. Su [11] and Qin [12] propose a novel boundary-aware network, using the hybrid loss to help the network focus on region segmentation without neglecting boundaries. These kind of losses improve the boundary quality but not the differenciation between similar objects or small objects segmentation.

To enlarge the receptive field to segment small objects, Yu [13] proposes what he calls *dilated convolutions*. By combining them with deep residual networks [14], he introduces dilated residual networks [15]. Wang [16] proposes a multi-path dilated residual network based on Mask-RCNN model [17], and solves the problem of information loss of small objects in deep neural networks. Liu [18] proposes a context embedding object detection network capturing both details and context information to boost the performance on small object detection. However, dilated convolutions often lead to gridding artifacts [13]. Attention plays an important role in human perception [19, 20, 21]. An important property of the human visual system is to not process a whole scene at once. Instead, humans exploit a sequence of partial glimpses

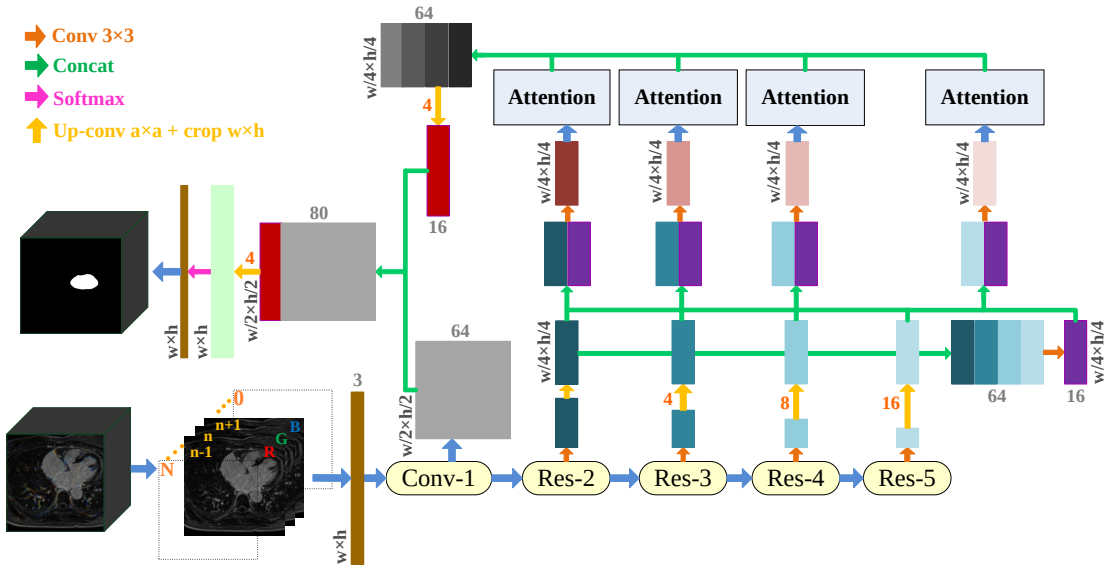


Fig. 1: Architecture of our network.

and selectively focus on salient parts in order to capture the visual structure in a better way [22, 23]. For this reason, attention modules have been developed: they focus on important regions, filter irrelevant information, and make up the limited receptive field of CNNs. They get good performance on segmentation tasks [24, 25, 26, 27]. For example, Zhang [24] proposes an efficient multi-scale feature interaction mechanism with attention, paying more attention to the important regions of objects, capturing more detail information, and so improving segmentation accuracy on small objects. Attention modules are also used for cardiac segmentation. Zhou [28] designed a cross-modal attention module between the encoder and the decoder, which leverages the correlated information between modalities to benefit the cross-modal cardiac segmentation. Based on 3D U-Net [29], Li [30] designed an attention module based on hierarchical aggregation to force the network to focus on the left atrium. Zhang [31] designed three types of attention modules (spatial, channel, and region) achieving good segmentation results on ventricles. Tong [32] presents an interleaved attention mechanism, improving the performance of cardiac MRI segmentation when applied to recurrent FCNs. Wei [33] proposes a spatial constrained channel attention module to pay more attention to the left ventricle and to decrease the impact of surrounding similar tissues. This approach leads to an effective segmentation of multiply connected domains but do not take the boundaries into account.

Facing these difficulties, we propose a novel attention FCN framework that focuses on the region of interest and is region- and boundary-aware. The main contributions of our work are: 1) a novel attention network framework based on the pretrained Resnet-101 with attention module, which can improve the segmentation accuracy on small parts of objects; 2) a novel hybrid loss that considers regions and boundaries of objects equally by combining region loss with boundary loss.

II. Methodology

A. Overview of Network Architecture

We propose a new attention network (see Fig. 1) using ResNet-101 pretrained on ImageNet [34] to compute feature maps. We discard its average pooling and fully connected layers, and keep only the sub-network made of one convolution-based and four residual-based "stages". Since the resolution decreases at each stage, we obtain a set of fine to coarse feature maps (with five levels of features).We add *specialized* convolutional layers (with a 3×3 kernel size) with K (*e.g.* $K = 16$) feature maps placed at the end of four residual-based "stages". They are concatenated together after up-convolutional layers. These last feature maps are combined with each of the outputs of the specialized layers, and then fed into the attention module to generate the attention features. Finally, we concatenate the attention features with the outputs of *Conv1* and we fed them into the softmax layer.

Attention Module. As mentioned before, in a traditional segmentation model, the usual issue is that receptive fields are too small, which leads to poor contextual representations. Furthermore, the relationship between the different channels should be explored since each channel map represents one feature-specific response. Therefore, improving the dependencies among channel maps can lead to richer features. To solve these issues, we use an attention module inspired by [3]. As shown in Fig. 2, $\mathbf{F} \in \mathbb{R}^{C \times W \times H}$ acts as an input feature map for the attention module, where C, W, H are the channel, the width and the height of the feature map respectively. The upper branch $\mathbf{F}$ is fed into a convolutional, a Reshape and then a Transpose layers, resulting in a feature map $\mathbf{F}_0^u \in \mathbb{R}^{(W \times H) \times C}$. In the second branch (consider the order from top to bottom), the input feature map $\mathbf{F}$ follows the same operations minus the Transpose layer, resulting in

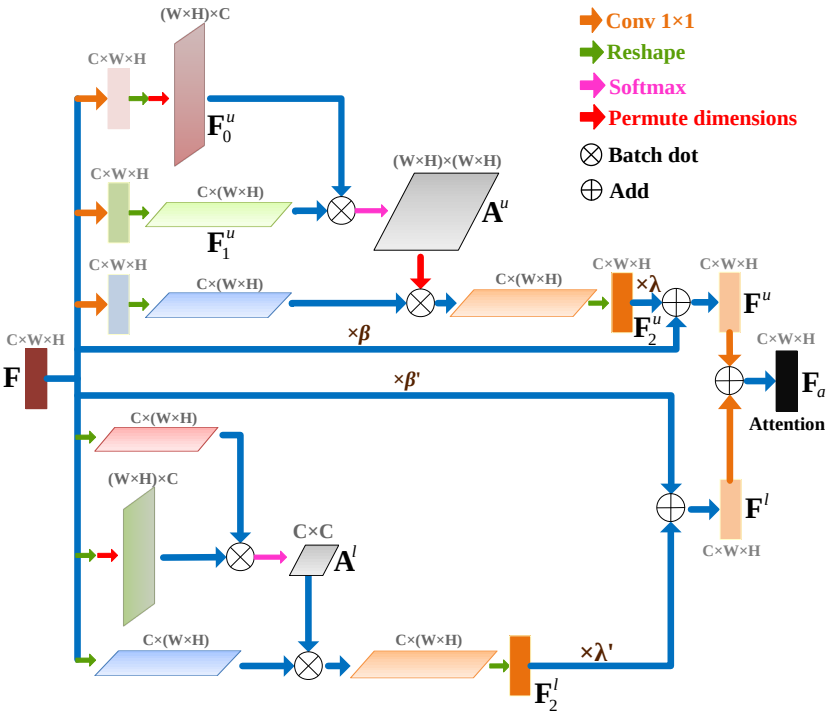


Fig. 2: Attention Module. λ, λ', β and β' as hyperparameters, which is trained like the convolutional kernel. They decrease the weight of the unimportant feature maps.

$\mathbf{F}_1^u \in \mathbb{R}^{C\times(W\times H)}$. Then, the Multiply and the Softmax layers follow; they are applied on $\mathbf{F}_0^u$ and $\mathbf{F}_1^u$ to obtain the spatial attention map $\mathbf{A}^u \in \mathbb{R}^{(W\times H)\times(W\times H)}$. The input $\mathbf{F}$ is fed into a different convolutional layer in the third branch, and is then multiplied by $\mathbf{A}^u$ fed into the Transpose layer, resulting in $\mathbf{F}_2^u$. Therefore the output $\mathbf{F}^u$ of the upper branch can be formulated as follows:

$$\mathbf{F}^u = \lambda \times \mathbf{F}_2^u + \beta \times \mathbf{F}, \tag{1}$$

where $\lambda \in \mathbb{R}^C$ is initialized to [0,..,0], and $\beta \in \mathbb{R}^C$ is initialized to [1,..,1]. The values λ and β are used to gradually learn the importance of the spatial attention map.

In the lower branch, the attention module mainly focuses on the most important channels. The channel attention map $\mathbf{A}^l$ can be obtained by different combinations of convolutional, Reshape and Transpose layers as shown at the bottom of Fig. 2. Finally, the output $\mathbf{F}^l$ of the lowest branch can be defined as follows: $\mathbf{F}^l = \lambda' \times \mathbf{F}_2^l + \beta' \times \mathbf{F}$, where $\lambda' \in \mathbb{R}^C$ is initialized to [0,..,0], and $\beta' \in \mathbb{R}^C$ is initialized to [1,..,1]. The feature map $\mathbf{F}_2^l$ denotes the results of the product of the input $\mathbf{F}$ with $\mathbf{A}^l$ fed into a convolutional passing through the transpose block. Therefore, the attention feature map $\mathbf{F}_a$ is defined as:

$$\mathbf{F}_a = Conv\left(\mathbf{F}^u\right) + Conv\left(\mathbf{F}^l\right). \tag{2}$$

Compared to [3], we make learnable the coefficient beta multiplying F in the channel and position attention modules (Eq. 1) so that the improved attention modules focus more on important features. Furthermore, we do not use a convolution layer before the channel attention module like in [3], so we do not destroy the relationships between channel maps. Finally, we apply one attention module for each scale explaining that we have four attention modules, contrary to [3] where the attention modules are only used at the output of the network.

B. Hybrid Loss

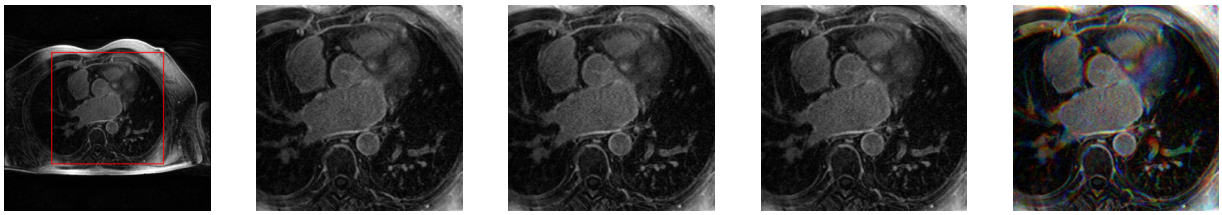
Most of medical segmentation methods directly use Categorical Cross Entropy[35] (CCE) or Dice Coefficient [36] (DC) losses. Models trained with CCE loss usually have low confidence in differentiating boundary pixels, leading to blurry boundaries. DC were proposed for biased training sets but are not specifically designed for capturing fine structures.

In our framework, we combine four losses: the dice loss, the cross-entropy (CE) loss, the structure similarity (SSIM) loss [37], and our self-made boundary loss. When used alone, the dice and CE losses have respectively shown issues in capturing fine structures and in segmenting correctly boundary pixels. Combined together with in addition the SSIM loss (used to reduce the impact of the misleading similarities of neighboring tissues), we obtain an efficient region loss. By adding to it our own boundary loss, we are then able to refine the segmentation which converges to the boundaries.

Our hybrid loss consists of two parts: region loss and boundary one. It is defined as: $\ell_{\mathrm{H}} = \ell_{\mathrm{R}} + \ell_{\mathrm{B}}$, where ℓ_{R} denotes the region loss and ℓ_{B} denotes the boundary loss. They are explained hereafter.

Region Loss.

To obtain high quality regional segmentation, we define ℓ_{R} as a region loss: $\ell_{\mathrm{R}} = \ell_{\mathrm{CCE}} + \ell_{\mathrm{SSIM}} + \ell_{\mathrm{DC}}$, where ℓ_{CCE},



(a) Initial image (b) Slice $n-1$ (c) Slice n (d) Slice $n+1$ (e) "3D-Like" slice

Fig. 3: Illustration of our "3D-Like" procedure. The red box depicts the boundary of the cropped input image. Three successive cropped slices (b-d) are used to build a "3D-Like" image (e).

ℓ_{SSIM} and ℓ_{DC} denote Categorical Cross Entropy (CCE) loss , Structural Similarity (SSIM) loss and Dice Coefficient (DC) loss respectively.

CCE [35] loss is commonly used for multi-class classification and segmentation. It is defined as $\ell_{\text{CCE}} = -\sum_{i=1}^{C}\sum_{a=1}^{H}\sum_{b=1}^{W} y^i_{(a,b)} \ln {y_*}^i_{(a,b)}$, where C is the number of classes of each image, H and W are the height and width of image, $y^i_{(a,b)} \in \{0,1\}$ is the ground truth one-hot label of class i at position (a,b) and ${y_*}^i_{(a,b)}$ is the predicted probability that (a,b) belongs to class i.

SSIM [37] loss can assess image quality [37], and can be used to capture the structural information, which will decrease the mis-segmentation rate of surrounding similar tissues. Therefore, we integrated it into our training loss to learn the differences between the segmented domain and similar tissues around the segmented domain. Let $\mathbf{S}$ and $\mathbf{G}$ be the predicted probability map and the ground truth mask respectively, the SSIM loss function of $\mathbf{S}$ and $\mathbf{G}$ is defined as $\ell_{\text{SSIM}} = 1 - ((2\mu_{\text{S}}\mu_{\text{G}}+\varepsilon_1)(2\sigma_{\text{SG}}+\varepsilon_2)) / ((\mu_{\text{S}}^2+\mu_{\text{G}}^2+\varepsilon_1)(\sigma_{\text{S}}^2+\sigma_{\text{G}}^2+\varepsilon_2))$, where μ_{S}, μ_{G} and σ_{S}, σ_{G} are the means and standard deviations of $\mathbf{S}$ and $\mathbf{G}$ respectively, σ_{SG} is their covariance, $\varepsilon_1 = 0.01^2$ and $\varepsilon_2 = 0.03^2$ are used to avoid a division by zero.

DC [36] loss is used to measure the similarity between two sets as defined in Eq. 2. But for the multi-class segmentation task, Eq. 2 is not suitable due to the class imbalance problem in such cases. Therefore, we extend the definition of the DC loss to multiclass segmentation in the following manner:

$$dice_i = (\epsilon + 2\sum_{n=1}^{N_i} y^i_n \, {y_*}^i_n) \,/\, (\epsilon + \sum_{n=1}^{N_i} (y^i_n + {y_*}^i_n)) \quad (3)$$

$$\ell_{\text{DC}} = 1 - \sum_{i=1}^{C} dice_i / (N_i + \epsilon), \quad (4)$$

where N_i denotes the numbers of class i and $\epsilon > 0$ is a smooth factor.

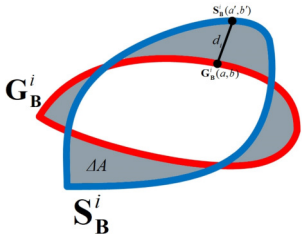


Fig. 4: Illustration of calculating boundary loss.

Boundary Loss.

The loss functions mentioned before are mainly for region segmentation, so we propose a boundary loss function to optimize the segmentation result. As shown in Fig. 4, ΔA denotes the difference between the boundary $\mathbf{G}^i_{\mathbf{B}}$ of the ground truth of class i and the boundary $\mathbf{S}^i_{\mathbf{B}}$ of the prediction of class i. When ΔA tends to zero, it means that the segmentation results are becoming better around the boundaries. Therefore the boundary loss is defined as

$$\ell_{\text{B}} = \sum_{i}^{C} \int_{\mathbf{G}^i_{\mathbf{B}}} \left\| \mathbf{S}^i_{\mathbf{B}}(a',b') - \mathbf{G}^i_{\mathbf{B}}(a,b) \right\|^2 \mathrm{d}(a,b), \quad (5)$$

where $\mathbf{G}^i_{\mathbf{B}}(a,b)$ is a point on boundary $\mathbf{G}^i_{\mathbf{B}}$ and $\mathbf{S}^i_{\mathbf{B}}(a',b')$ denotes the corresponding point on boundary $\mathbf{S}^i_{\mathbf{B}}$, along the direction normal to $\mathbf{G}^i_{\mathbf{B}}$, i.e., $\mathbf{S}^i_{\mathbf{B}}(a',b')$ is the intersection of $\mathbf{S}^i_{\mathbf{B}}$ and the line that is normal to $\mathbf{G}^i_{\mathbf{B}}$ at position (a',b') (see Fig. 4 for an illustration), $\|\cdot\|$denotes the L2 norm.

III. Experimental Results

Dataset Description. We evaluate our method on the MICCAI 2018 Atrial Segmentation Challenge [1] (AtriaSeg18). Its aim is to segment the left atrium. It contains 100 annotated 3D MRIs from patients with atrial fibrillation. The pixel spacing of the MR images is 0.625 x 0.625 x 0.625 mm/pixel. The dataset includes two different image sizes: $88\times576\times576$ and $88\times640\times640$.

Preprocessing. We cropped each slice to 346×346 pixels as shown in Fig. 3a. The pre-processing begins with a Gaussian normalization. Because ResNet-101 network's input is an RGB image, we propose to take advantage of the 3D information by stacking 3 successive 2D frames, as presented in our previous works [38, 39]: to segment the n^{th} slice, we use the n^{th} slice of the MR volume, and its neighboring $(n-1)^{th}$ and $(n+1)^{th}$ slices, as green, red and blue channels, respectively. This new image, named "3D-Like" image, enhances the boundaries of objects, as shown in Fig. 3.

[1] http://atriaseg2018.cardiacatlas.org/

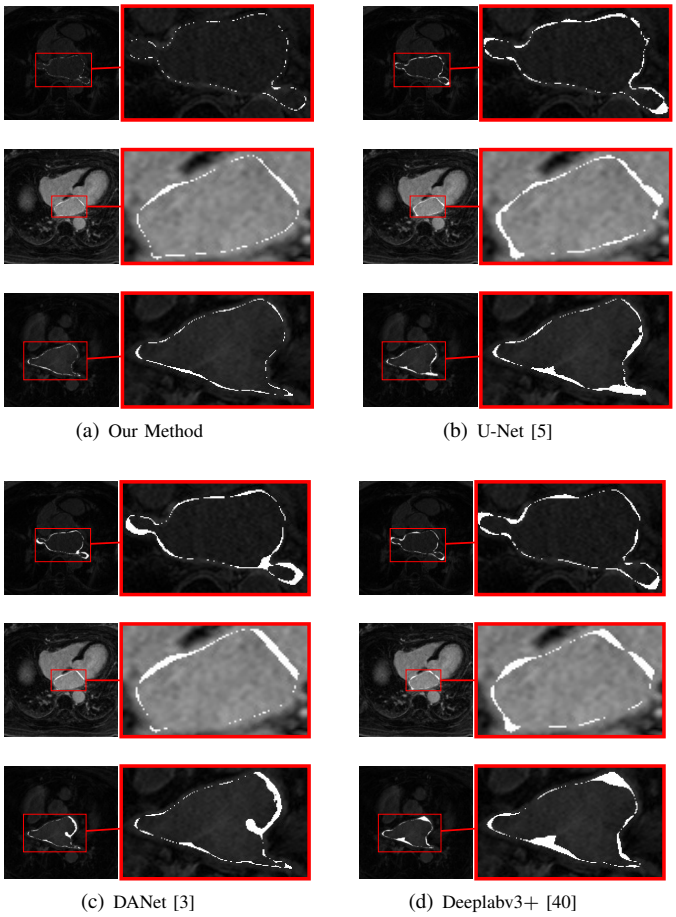
Postprocessing. We crop the initial volume of size $88 \times W \times H$ into an image of size $88 \times w \times h$ (where W and H are the initial width and height of a slice). We keep only the greatest connected component of the output segmentation and pad with zeros to get back a $T \times W \times H$ image.

Implementation and Experimental Setup. We implemented our experiments on Keras/TensorFlow using a NVidia Quadro P6000 GPU. We used the hybrid loss function, softmax to get a probability distribution over classes, Adam optimizer (batchsize = 3, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 0.001$, $\mathrm{lr} = 0.01$) and did not use learning rate decay. We trained the network during 30 epochs.

Evaluation Methods. Three metrics are used to evaluate our method: dice to evaluate the regions, and 95% Hausdorff distance (95HD) and Average Hausdorff distance (AHD) to quantitatively evaluate the boundaries.

Comparison with State-of-the-arts Methods. The experimental results obtained by several state-of-the-art segmentation networks are reported in Table I. Compared to other networks proposed in the context of medical image segmentation ,i.e., U-Net [5], DANet [3] and Deeplabv3+ [40], our network achieves a mean improvement of 3.236%, 7.563% and 6.348% (in terms of DC), 1.579 mm, 3.277 mm and 3.004 mm (on 95HD) and 0.082 mm, 0.384 mm and 0.374 mm (on AHD), respectively. The attention module increases segmentation performance by 0.552% (DC), 0.215 mm (95HD), and 0.015 mm (AHD), respectively as shown in Table I.

Ablation Study. To explain the advantages of the proposed hybrid loss, we conduct an ablation study. We compare the segmentation results with and without hybrid loss (see Table I). Segmentation performance increases for DC, 95HD and AHD for the 4 architectures, proving the benefits of the proposed hybrid loss.



(a) Our Method (b) U-Net [5] (c) DANet [3] (d) Deeplabv3+ [40]

Fig. 5: Comparison of the proposed method and other state-of-the-art architectures. The white pixels are the differences between the prediction and the GT.

IV. Conclusion

In this paper, we propose a novel attention network architecture, and a new hybrid loss. Unlike a traditional FCN, we first add multi-layer features to keep as much details as possible, then we concatenate them with level features, and input them in the attention modules to obtain the attentional features. By using the attention module, the proposed network framework is able to prevent the interferences between the surrounding similar tissues and to capture large-scale and thiner structures. We propose a hybrid loss function that fairly treats regions and boundaries of objects, optimizes the convergence to the boundaries, while maintaining the segmentation precision of the regions. Compared to the state-of-the-arts methods on the AtriaSeg18 challenge dataset, our segmentation results overcome the best one by an average of 2.179% in terms of DC and 1.3 mm on 95HD. Taking into account regions as well as boundaries in our loss permits to have a segmentation more precise, especially at the boundaries. Moreover, our method with attention module and hybrid loss is more robust. The computation time of our pipeline is less than 4 seconds for an entire 3D volume of a heart. As future works, we plan to continue to study the impact of the hybrid loss when the region of interest and the background are imbalanced. We plan also to add shape constraints to the predicted boundary of the LA in the attention module. The final aim is to be able to accurately segment LA wall to diagnose fibrosis.

References

[1] A. Sinha and J. Dolz, "Multi-scale guided attention for medical image segmentation," *arXiv preprint arXiv:1906.02849*, 2019.

[2] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[3] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the*

TABLE I: Comparison of our method and other state-of-the-art architectures using a 5 fold cross-validation.

Method	Att. Module	Hyb. Loss	DC/%	95HD/mm	AHD/mm
U-Net [5]			88.556 (±2.586)	4.447 (±0.996)	0.212 (±0.077)
		✔	89.613 (±2.257)	4.169 (±0.960)	0.210 (±0.118)
DANet [3]			84.229 (±3.774)	6.145 (±2.341)	0.514 (±0.477)
		✔	87.584 (±2.765)	4.903 (±1.448)	0.280 (±0.179)
Deeplabv3+ [40]			85.444 (±3.079)	5.872 (±2.345)	0.504 (±0.614)
		✔	87.556 (±1.155)	5.210 (±1.087)	0.273 (±0.074)
Our Method			90.774 (±1.568)	3.312 (±1.277)	0.158 (±0.092)
	✔		91.326 (±1.174)	3.097 (±0.810)	0.143 (±0.055)
	✔	✔	**91.792 (±1.065)**	**2.868 (±0.667)**	**0.130 (±0.042)**

IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. of IEEE Intl. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.

[6] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. 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 Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[7] C. Chen, C. Biffi, G. Tarroni, S. Petersen, W. Bai, and D. Rueckert, "Learning shape priors for robust cardiac MR segmentation from multi-view images," in *Proc. of IEEE Intl. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. LNCS, vol. 11765. Springer, 2019, pp. 523–531.

[8] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *Medical Image Analysis*, vol. 51, pp. 21–45, 2019.

[9] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," *arXiv preprint arXiv:1812.07032*, 2018.

[10] H. Shen, R. Wang, J. Zhang, and S. J. McKenna, "Boundary-aware fully convolutional network for brain tumor segmentation," in *Proc. of IEEE Intl. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. LNCS, vol. 10434. Springer, 2017, pp. 433–441.

[11] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," 2019, pp. 3799–3808.

[12] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[13] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[14] 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*, 2016, pp. 770–778.

[15] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 472–480.

[16] E. K. Wang, X. Zhang, L. Pan, C. Cheng, A. Dimitrakopoulou-Strauss, Y. Li, and N. Zhe, "Multi-path dilated residual network for nuclei segmentation and detection," *Cells*, vol. 8, no. 5, p. 499, 2019.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," 2017, pp. 2961–2969.

[18] T. Liu, Y. Zhao, Y. Wei, Y. Zhao, and S. Wei, "Concealed object detection for activate millimeter wave image," *IEEE Trans. on Industrial Electronics*, vol. 66, no. 12, pp. 9909–9917, 2019.

[19] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," vol. 20, no. 11, pp. 1254–1259, 1998.

[20] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[21] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Reviews Neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.

[22] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," in *Advances in Neural Information Processing Systems*, 2010, pp. 1243–1251.

[23] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," 2018, pp. 3–19.

[24] P. Zhang, W. Liu, H. Wang, Y. Lei, and H. Lu, "Deep gated attention networks for large-scale street-level scene segmentation," *Pattern Recognition*, vol. 88, pp. 702–714, 2019.

[25] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3640–3649.

[26] Y. Wang, Z. Deng, X. Hu, L. Zhu, X. Yang, X. Xu, P.-A. Heng, and D. Ni, "Deep attentional features for prostate segmentation in ultrasound," in *Proc. of IEEE Intl. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. LNCS, vol. 11073. Springer, 2018, pp. 523–530.

[27] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[28] Z. Zhou, X. Guo, W. Yang, Y. Shi, L. Zhou, L. Wang,

and M. Yang, "Cross-modal attention-guided convolutional network for multi-modal cardiac segmentation," in *Proc. of the Intl. Workshop on Mach. Learning in Med. Imaging*, ser. LNCS, vol. 11861. Springer, 2019, pp. 601–610.

[29] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. of IEEE Intl. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9901. Springer, 2016, pp. 424–432.

[30] C. Li, Q. Tong, X. Liao, W. Si, Y. Sun, Q. Wang, and P.-A. Heng, "Attention based hierarchical aggregation network for 3D left atrial segmentation," ser. LNCS, vol. 11395. Springer, 2018, pp. 255–264.

[31] T. Zhang, A. Li, M. Wang, X. Wu, and B. Qiu, "Multiple attention fully convolutional network for automated ventricle segmentation in cardiac magnetic resonance imaging," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 5, pp. 1037–1045, 2019.

[32] Q. Tong, C. Li, W. Si, X. Liao, Y. Tong, Z. Yuan, and P. A. Heng, "RIANet: Recurrent interleaved attention network for cardiac MRI segmentation," *Comp. in Bio. and Med.*, vol. 109, pp. 290–302, 2019.

[33] H. Wei, W. Xue, and D. Ni, "Left ventricle segmentation and quantification with attention-enhanced segmentation and shape correction," in *Proc. of the Intl. Symp. on Image Computing and Digital Medicine*, 2019, pp. 226–230.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[35] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. of the Intl. Conf. on Neural Information Processing Systems (NIPS)*, 2018, pp. 8792–8802.

[36] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[37] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. of the 37th Asilomar Conference on Signals, Systems and Computers*, vol. 2, 2003, pp. 1398–1402.

[38] Y. Xu, T. Géraud, and I. Bloch, "From neonatal to adult brain MR image segmentation in a few seconds using 3D-like fully convolutional network and transfer learning," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4417–4421.

[39] É. Puybareau *et al.*, "Left atrial segmentation in a few seconds using fully convolutional network and transfer learning," ser. LNCS, vol. 11395. Springer, 2018, pp. 339–347.

[40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, pp. 801–818.

Stacked and Parallel U-Nets with Multi-Output for Myocardial Pathology Segmentation

Zhou Zhao*, Nicolas Boutry[0000−0001−6278−4638], and Élodie Puybareau[0000−0002−2748−6624]

EPITA Research and Development Laboratory (LRDE), Le Kremlin-Bicêtre, France
*Corresponding author
zz@lrde.epita.fr, nicolas.boutry@lrde.epita.fr,
elodie.puybareau@lrde.epita.fr

Abstract. In the field of medical imaging, many different image modalities contain different information, helping practitionners to make diagnostic, follow-up, etc. To better analyze images, mixing multi-modalities information has become a trend. This paper provides one cascaded UNet framework and uses three different modalities (the late gadolinium enhancement (LGE) CMR sequence,the balanced- Steady State Free Precession (bSSFP) cine sequence and the T2-weighted CMR) to complete the segmentation of the myocardium, scar and edema in the context of the MICCAI 2020 myocardial pathology segmentation combining multi-sequence CMR Challenge dataset (MyoPS 2020). We evaluate the proposed method with 5-fold-cross-validation on the MyoPS 2020 dataset.

Keywords: Deep Learning · Myocardial Pathology · Segmentation · UNet.

1 Introduction

The assessment of myocardial viability is essential for diagnosis and follow-up of patients suffering from myocardial infarction (MI) [17, 16]. However, many different images modalities in the field of medical imaging are available and are complementary. Late gadolinium enhancement (LGE) cardiac magnetic resonance (CMR) sequence which visualizes MI, T2-weighted CMR (imaging the acute injury and ischemic regions) and balanced-Steady State Free Precession (bSSFP) cine sequence (which captures cardiac motions and presents clear boundaries) are examples of such imaging modalities. Therefore, making a better use of the information in these different modalities has become a research focus. In recent years, many semi-automated and automated methods have been proposed for multi-modal medical image segmentation using deep learning-based methods, such as convolutional neural networks (CNNs) [8] and fully convolutional networks (FCNs) [9] especially the U-Net architecture [11]. For example, Guo [3, 4] proposed a conceptual image fusion architecture for supervised biomedical image analysis. They designed and implemented an image segmentation system based on deep CNNs to contour the lesions of soft tissue sarcomas using multimodal images by fusing the information derived from different modalities.

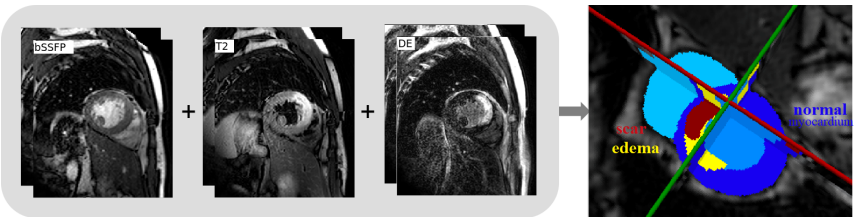


Fig. 1: Myocardial pathology, the picture is from MyoPS2020 challenge [1].

Although we can use multi-modal information to improve the myocardial pathology segmentation, class imbalance remains a problem to tackle. Network overfitting is common in the field of medical imagingbecause of the relatively small size of handled datasets. Data augmentation is classically used in the pre-processing stage to overcome this limitation, and weighted loss functions are designed. For example, Zhao et al. [15, 10] used data augmentation by rotating and flipping the heart segmentations to reduce the impact of overfitting. Zhao et al. [14] proposed an automated data augmentation method for synthesizing labeled medical images, which provided significant improvements over state-of-the-art methods for one-shot biomedical image segmentation. Sudre et al. [13] proposed the generalized dice to solve the problem of highly unbalanced segmentations. Abraham et al. [1] proposed a generalized focal loss function based on the Tversky index to address the issue of data imbalance in medical image segmentation. Examples of data augmentation methods to overcome this issue can be found in [2, 12, 6, 5, 7]. However, datasets obtained through data augmentation are strongly correlated with the original datasets, Therefore, the proportion of negative samples remains significantly larger than the proportion of positive samples after data augmentation. Thus, data augmentation does not reduces the risk of overfitting. For the proposed improved loss function can effectively reduce the issues of class imbalance, it does not fundamentally address the problems caused by the lack of datasets.

Therefore, in this paper, in order to segment myocardial pathology (see Fig. 1), we begin with a segmentation of the anatomical tissue (left ventricle (LV), right ventricle (RV), whole heart (WH), myocardium (myo)) around it, and then let the network learn a relationship between these segmentation results to obtain the myocardial pathology. Compared with direct segmentation of myocardial pathology, the effect of class imbalance can be reduced by the segmentation of surrounding anatomical tissues, because it helps the network to focus on the small lesions regarding to the surrounding tissues.

2 Methodology

2.1 Overview of Network Architecture

We propose a hybrid network (see Fig. 2) using 5 UNet [11] to segment myocardial pathology. Our network is composed of three UNet named **UNet1** and two named **UNet2**. The main difference between **UNet1** and **UNet2** is number of

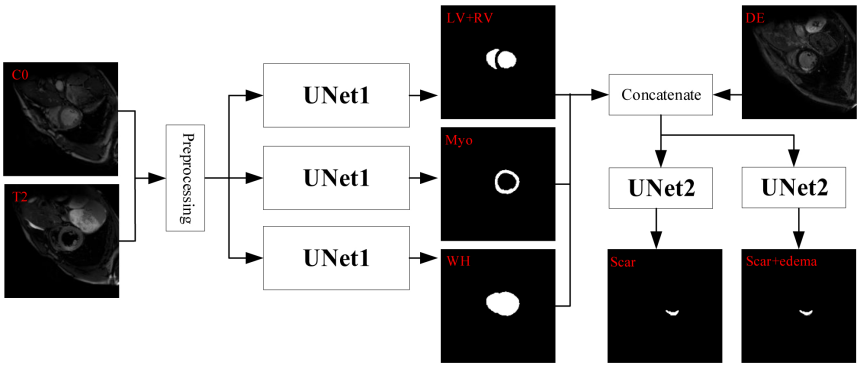


Fig. 2: Global overview of the proposed method.

Table 1: The structural configuration of UNet.

Layers	Input size		Operation	Kernel	Stride	Regul.	Output size	
	UNet1	UNet2					UNet1	UNet2
Input image	(240,240,2)	(240,240,4)	-	-	-	-	(240,240,2)	(240,240,4)
C1	(240,240,2)	(240,240,4)	[Conv2d+relu]*2	3	1	L2	(240,240,64)	(240,240,8)
C2	(240,240,64)	(240,240,8)	Maxpooling2d	2	-	-	(120,120,64)	(120,120,8)
C3	(120,120,64)	(120,120,8)	[Conv2d+relu]*2	3	1	L2	(120,120,128)	(120,120,16)
C4	(120,120,128)	(120,120,16)	Maxpooling2d	2	-	-	(60,60,128)	(60,60,16)
C5	(60,60,128)	(60,60,16)	[Conv2d+relu]*2	3	1	L2	(60,60,256)	(60,60,32)
C6	(60,60,256)	(60,60,32)	Maxpooling2d	2	-	-	(30,30,256)	(30,30,32)
C7	(30,30,256)	(30,30,32)	[Conv2d+relu]*2+Dropout	3	1	L2	(30,30,512)	(30,30,64)
C8	(30,30,512)	(30,30,64)	Maxpooling2d	2	-	-	(15,15,512)	(15,15,64)
C9	(15,15,512)	(15,15,64)	[Conv2d+relu]*2+Dropout	3	1	L2	(15,15,1024)	(15,15,128)
O1	(240,240,2)	(240,240,2)	Sigmoid	-	-	-	(240,240,1)	(240,240,1)

filters as shown in Table. 1: the number of filters of **UNet1** is [64 128 256 512 256 128 64] and the number of filters of **UNet2** is [8 16 32 64 32 16 8]. Their framework is same. It consists of the classical two parts of the UNet network as shown in Fig. 3: a down-sampling part and an up-sampling part, and shortcut connections between the two parts to fuse high-level features and low-level features. **UNet1** is used to segment the anatomical tissue around myocardial pathology and obtain three segmentation results: LV+RV, Myo, and WH. **UNet2** is used to segment myocardial pathology by learning the relationships between the surrounding anatomical tissue and the pathological ones. Since the lesions are very small and unbalanced, we reduce the number of filters of **UNet2** in order to reduce the impact of overfitting.

3 Experimental Results

Dataset Description. We evaluate our method on the myocardial pathology segmentation combining multi-sequence CMR [2] dataset (MyoPS 2020). Its aim is to segment myocardial pathology, especially scar (infarcted) and edema regions.

[2] http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/MyoPS20/index.html

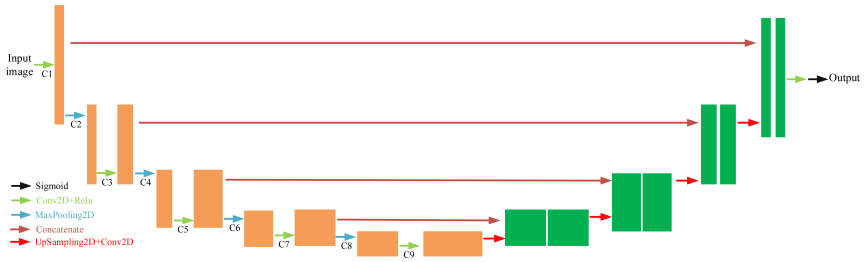


Fig. 3: Architecture of networks.

It contains 45 cases of multi-sequence CMR (25 cases for training and 20 cases for testing). Each case refers to a patient with three sequence CMR, i.e., LGE, T2 and bSSFP CMR. The slice spacings of multi-sequence CMR volume range from 11.999 mm/pixel to 23.000 mm/pixel, while in-plane resolution ranged from 0.729 mm/pixel to 0.762 mm/pixel. The average sizes: $482 \times 479 \times 4$ pixels.

Preprocessing and Postprocessing. We cropped each slice to 240×240 pixels and we do not use data augmentation. The pre-processing begins with a Gaussian normalization. For post-processing, we pad with zeros to get back a initial width and height of a slice.

Implementation and Experimental Setup. We implemented our experiments on Keras/TensorFlow using a NVidia Quadro P6000 GPU. We used five different loss functions for training the network and used sigmoid to get a probability distribution of the left and right ventricle, myocardium, whole heart, scar and edema, and scar, respectively (as shown in Fig. 2). Adam optimizer (batch-size $= 1$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 0.001$, lr $= 0.0001$) and did not use learning rate decay. We trained the network during 300 epochs.

Training Step. First, we kept weight of **UNet2** unchanged, which means **UNet2** was not trained at the beginning, then we trained **UNet1**. After finished the train of **UNet1**, we kept weight of **UNet1** unchanged, then trained **UNet2**.

Evaluation Methods. One metric is used to evaluate our method: dice coefficient (DC) to evaluate the regions of myocardial pathology.

3.1 Segmentation Results

As shown in Table. 2, we evaluate the proposed method with 5-fold-cross-validation. We obtain a mean DC of 92.3% on WH, 84.9% on LV+RV, and 84.7% on Myo by **UNet1**. Without using data augmentation, based on the original dataset, we obtain a higher segmentation accuracy, which lays the foundation for the subsequent segmentation of myocardial pathology. Finally, we obtain a mean DC of 20.6% on edema, 51% on scar by **UNet2**. We used the trained network to predict the testset (20 cases) and received the evaluation of our prediction results from the MyoPS2020 organizer: the mean DC of 58.6% on scar and the mean DC of 63.9% on scar and edema.

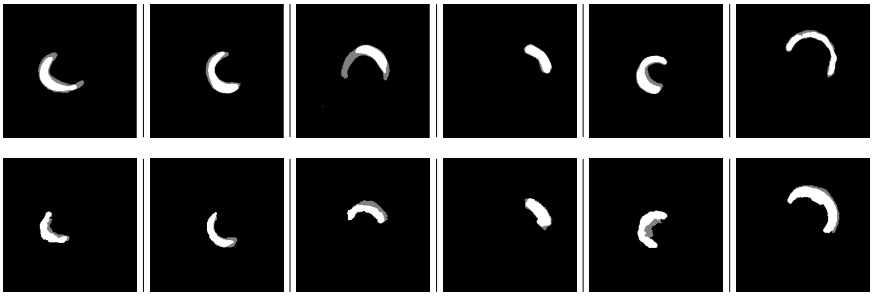
Table 2: Evaluation results on 5-fold-cross-validation.

Patient	**101-105**	**106-110**	**111-115**	**116-120**	**121-125**	**Average**	**Test datasets**
Edema	0.284	0.153	0.189	0.122	0.280	0.206	–
Scar	0.473	0.496	0.515	0.464	0.602	0.510	0.586
Myo	0.844	0.852	0.811	0.859	0.869	0.847	–
LV+RV	0.818	0.854	0.812	0.897	0.864	0.849	–
WH	0.925	0.937	0.876	0.918	0.959	0.923	–

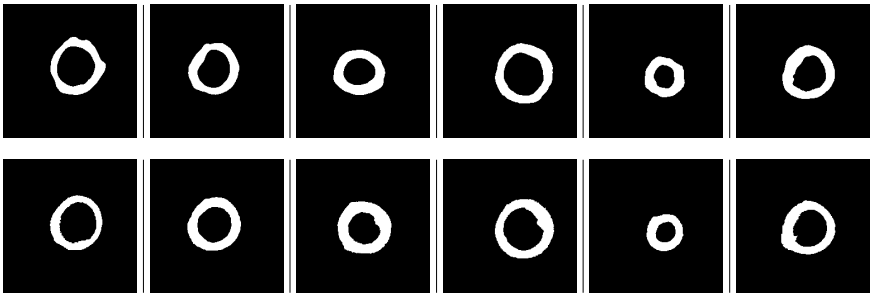
As shown in Fig. 4, for the segmentation results of whole heart, left and right ventricle, and myocardium, as the number of positive samples continues to decrease, the segmentation accuracy is also decreasing, and false segmentation is mainly concentrated at the boundary, which is mainly because ambiguities often appear near the boundaries of the target domains due to tissue similarities. For the segmentation results of edema and scar, the poorly segmentation result is not only on the boundary, but also in regions. In the original dataset, edema does not exist in many slices, which further leads to a reduction in the effective dataset for edema, therefore, the segmentation network is very difficult to segment edema.

4 Conclusion

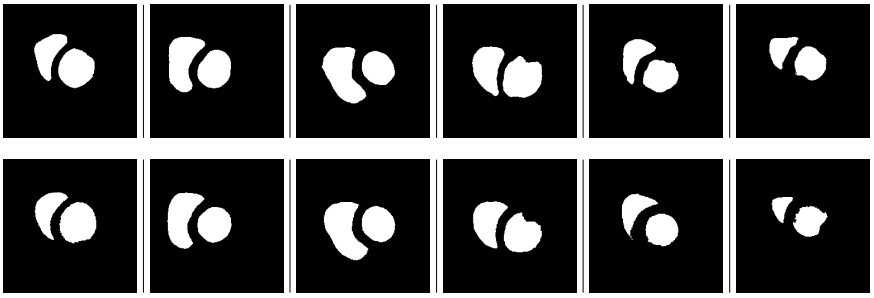
In this paper, we propose a way of reverse thinking, not to segment the myocardial pathology directly, but to learn a relationship between the surrounding normal tissue and it by designing one stacked and parallel UNets with multi-output framework. We evaluate the proposed method with 5-fold-cross-validation on the MICCAI 2020 myocardial pathology segmentation combining multi-sequence CMR Challenge dataset (MyoPS 2020) and achieve a mean DC of 20.6%, 51% on edema and scar,respectively. The computation time of the entire pipeline is less than 3 seconds for an entire 3D volume, making it usable for clinical practice. However, the segmentation accuracy of myocardial pathology is affected by the segmentation accuracy of surrounding normal tissues. Therefore, in our future work, we will continue to study the relationship between the surrounding normal tissue and myocardial pathology and improve the segmentation accuracy of surrounding normal tissues.



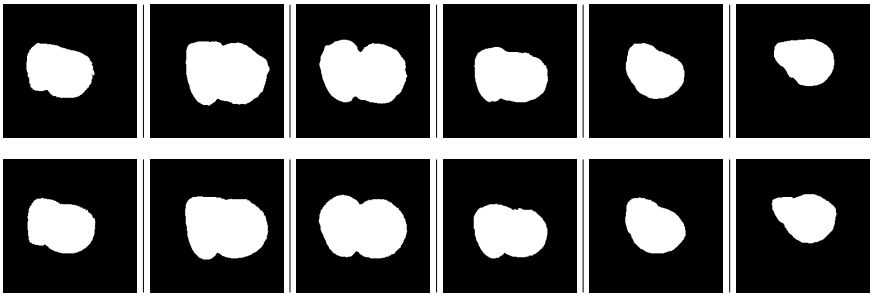
(b) Edema and scar. Scar is in white. Top = segmentation, bottom = Ground Truth



(d) Myocardium. Top = segmentation, bottom = Ground Truth



(f) Left and right ventricle. Top = segmentation, bottom = Ground Truth



(h) Whole heart. Top = segmentation, bottom = Ground Truth

Fig. 4: Qualitative segmentation results.

References

[1] Abraham, N., Khan, N.M.: A novel focal tversky loss function with improved attention u-net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 683–687. IEEE (2019)

[2] Eaton-Rosen, Z., Bragman, F., Ourselin, S., Cardoso, M.J.: Improving data augmentation for medical image segmentation (2018)

[3] Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.: Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 903–907. IEEE (2018)

[4] Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.: Deep learning-based image segmentation on multimodal medical imaging. IEEE Transactions on Radiation and Plasma Medical Sciences **3**(2), 162–169 (2019)

[5] Hashemi, S.R., Salehi, S.S.M., Erdogmus, D., Prabhu, S.P., Warfield, S.K., Gholipour, A.: Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access **7**, 1721–1735 (2018)

[6] Hussain, Z., Gimenez, F., Yi, D., Rubin, D.: Differential data augmentation techniques for medical imaging classification tasks. In: AMIA Annual Symposium Proceedings. vol. 2017, p. 979. American Medical Informatics Association (2017)

[7] Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B.: Boundary loss for highly unbalanced segmentation. In: International conference on medical imaging with deep learning. pp. 285–296 (2019)

[8] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

[9] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. of CVPR. pp. 3431–3440 (2015)

[10] Puybareau, É., et al.: Left atrial segmentation in a few seconds using fully convolutional network and transfer learning. In: Proc. of the Intl. Workshop on STACOM. LNCS, vol. 11395, pp. 339–347. Springer (2018)

[11] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of MICCAI. LNCS, vol. 9351, pp. 234–241. Springer (2015)

[12] Shin, H.C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P., Michalski, M.: Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: International workshop on simulation and synthesis in medical imaging. pp. 1–11. Springer (2018)

[13] Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 240–248. Springer (2017)

[14] Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8543–8553 (2019)

[15] Zhao, Z., et al.: A two-stage temporal-like fully convolutional network framework for left ventricle segmentation and quantification on MR images. In: Proc. of the Intl. Workshop on STACOM. LNCS, vol. 12009, pp. 405–413. Springer (2019)
[16] Zhuang, X.: Multivariate mixture model for cardiac segmentation from multi-sequence mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 581–588. Springer (2016)
[17] Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2019)

A Two-Stage Temporal-Like Fully Convolutional Network Framework for Left Ventricle Segmentation and Quantification on MR Images

Zhou Zhao, Nicolas Boutry, Élodie Puybareau, and Thierry Géraud

EPITA Research and Development Laboratory (LRDE), Le Kremlin-Bicêtre, France
elodie.puybareau@lrde.epita.fr

Abstract. Automatic segmentation of the left ventricle (LV) of a living human heart in a magnetic resonance (MR) image (2D+t) allows to measure some clinical significant indices like the regional wall thicknesses (RWT), cavity dimensions, cavity and myocardium areas, and cardiac phase. Here, we propose a novel framework made of a sequence of two fully convolutional networks (FCN). The first is a modified temporal-like VGG16 (the "localization network") and is used to localize roughly the LV (filled-in) epicardium position in each MR volume. The second FCN is a modified temporal-like VGG16 too, but devoted to segment the LV myocardium and cavity (the "segmentation network"). We evaluate the proposed method with 5-fold-cross-validation on the MICCAI 2019 LV Full Quantification Challenge dataset. For the network used to localize the epicardium, we obtain an average dice index of 0.8953 on validation set. For the segmentation network, we obtain an average dice index of 0.8664 on validation set (there, data augmentation is used). The mean absolute error (MAE) of average cavity and myocardium areas, dimensions, RWT are 114.77 mm^2; 0.9220 mm; 0.9185 mm respectively. The computation time of the pipeline is less than 2 seconds for an entire 3D volume. The error rate of phase classification is 7.6364%, which indicates that the proposed approach has a promising performance to estimate all these parameters.

Keywords: Deep learning · VGG · Left ventricle quantification · Segmentation · Fully convolutional network.

1 Introduction

Left ventricle (LV) full quantification is critical to evaluate cardiac functionality and diagnose cardiac diseases. Full quantification aims to simultaneously quantify all LV indices, including the two areas of the LV (the area of its cavity and the area of its myocardium), six RWT's (along different directions and at different positions), three LV dimensions (along different directions), and the cardiac phase (diastole or systole) [1, 2], as shown in Fig. 1. However, the LV full quantification is challenging: LV samples are variable, not only because the samples can be obtained from different hospital, but also because some of them are not concerned by cardiac diseases. It is also challenging because there are complex correlations between the LV indices. For example, the cavity area has a direct influence on the three LV dimensions and the cardiac phase.

The MICCAI 2019 Challenge on Left Ventricle Full Quantification[1] (LVQuan19) is an extension of the one of 2018 [2] with the difference that now the original data is given without preprocessing for training and testing phases, to be closer to clinical reality.

We propose then in this paper a two-stage temporal-like FCN framework that segments and estimates the parameters of interest in 2D+t sequences of the MR image of a LV. First, in each temporal frame, we localize the greatest connected component detected by the localization network, we dilate it using a size equal to 10 pixels, and we compute the corresponding bounding box. This results in a sequence of cropped LV's (that we will abusively call cropped volume). Second, we use these cropped volumes to train the LV segmentation network. The procedure is depicted in Fig. 2. Finally, the segmentation results are used for the LV full quantification.

The pipeline is based on our previous works [3, 4] but with a new step: we added one localization network before the segmentation network. Compared with [5], our localization precision is higher, because we localize the entire LV region (the filled-in epicardium) instead of the center of the bounding box containing the LV structure. Compared with [6], our method is quicker and do not have memory limit problems. To take advantages of time information, we use 3 successive 2D frames $(n-1, n, n+1)$ at time n as inputs in the localization and in the segmentation networks, yielding to better results than the traditional approach which used only the information at time n for the n^{th} slice.

We evaluated the proposed method using the dataset provided by LVQuan19 with 5-fold-cross-validation. Experiments with (very) limited training data have shown that our model has a stable performance. We added pre-processing and post-processing steps to enhance and refine our results.

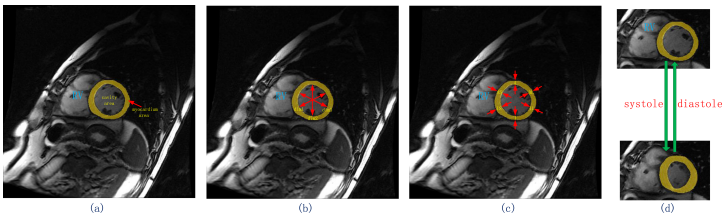


Fig. 1. Illustration of LV indices, including (a) the cavity area and the myocardium area, (b) three LV dimensions, (c) six regional wall thicknesses and (d) the cardiac phase (diastole or systole).

The plan is the following: we detail our methodology in Section 2, we detail our experiments in Section 3, and then Section 4 concludes.

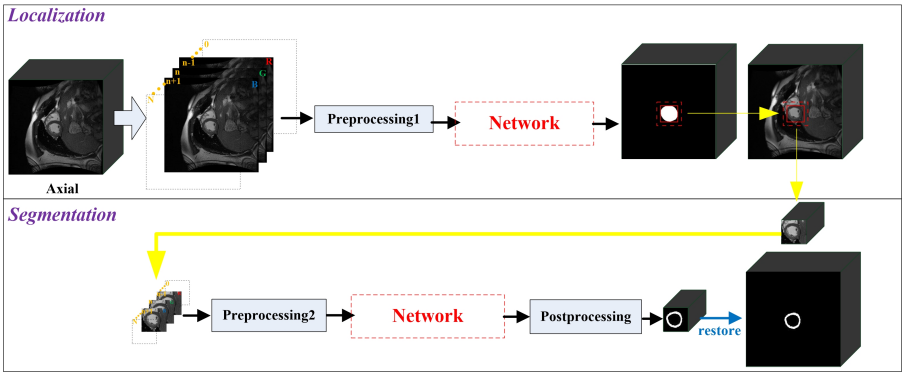


Fig. 2. Global overview of the proposed method.

2 Methodology

2.1 Dataset description

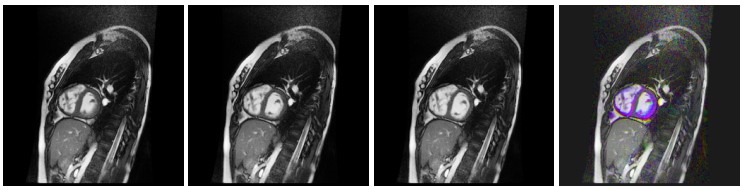
LV dataset used for this work was provided by the LVQuan19 challenge. It contains 56 patients processed SAX MR sequences. For each patient, 20 temporal frames are given and correspond to a whole cardiac cycle. All ground truth (GT) values of the LV indices are provided for every single frame. The pixel spacings of the MR images range from 0.6836 mm/pixel to 1.5625 mm/pixel, with mean values of 1.1809 mm/pixel. LV dataset includes two different image sizes: 256×256 or 512×512 pixels.

2.2 Preprocessings

Let us recall what we call *Gauss normalization*: for the $(2D+t)$-image I corresponding to a given patient, we compute $I := \frac{I-\mu}{\sigma}$ where μ is the mean of I and σ its standard deviation (σ is assumed not to be equal to zero). There are then two different pre-processing steps as depicted in Fig. 2.

- The first pre-processing (see preprocessing1 in Fig. 2) begins with a Gauss normalization. When we treat training data, we crop the initial slices into a 256×256 image to optimize the dice of the network (we do not do this for test datasets). Then we concatenate them for each n into a $256 \times 256 \times 3$ pseudo-color image where R, G, B correspond respectively to $n-1, n, n+1$ (we do not detail the cases $n = 1$ and $n = 20$ because of a lack of space).
- The second pre-processing (preprocessing2 in Fig. 2) is in four steps: (1) data augmentation using rotations and flips, (2) resizing with a fixed inter-pixel spacing $(0.65mm)$, (3) Gauss normalization, and (4) we concatenate into a pseudo-color image like above.

[1] https://lvquan19.github.io
[2] https://lvquan18.github.io



(a) $n-1$ (b) n (c) $n+1$ (d) concatenation

Fig. 3. Illustration of our "temporal-like" procedure.

Because the VGG-16 network's input is an RGB image, we propose to take advantage of the temporal information by stacking 3 successive 2D frames: to segment the n^{th} slice, we use the n^{th} slice of the MR volume, and its neighboring $(n-1)^{th}$ and $(n+1)^{th}$ slices, as green, red and blue channels, respectively. This new image, named "temporal-like" image, enhances the area of motions, here the heart, as shown in Fig. 3.

2.3 Network architecture

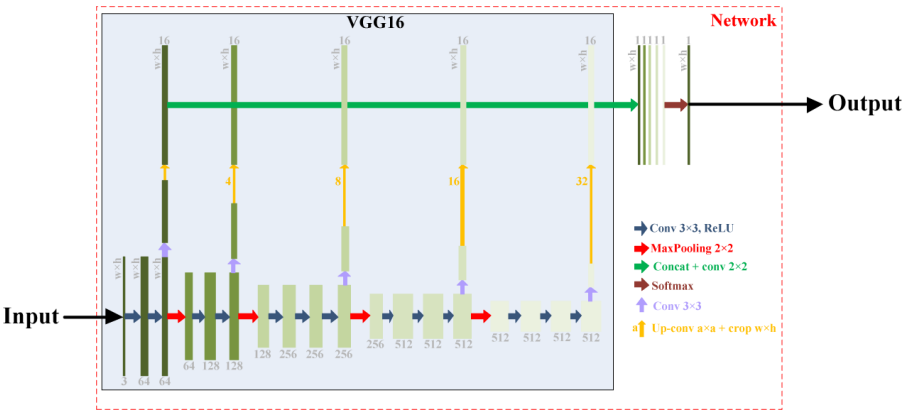


Fig. 4. Architecture of our networks.

The localization and the segmentation networks have the same architecture (see Fig. 4). First we downloaded the pre-trained original VGG16 [7] network architecture. We recall that this network has been pre-trained on millions of natural images of ImageNet for image classification [8]. Second, we discard its fully connected layers and this way we keep only the sub-network made of five convolution-based "stages" (the base network). Each stage is made of two convolutional layers, a ReLU activation function, and a max-pooling layer. Since the max-pooling layers decrease the resolution of the input image, we obtain a set of fine to coarse feature maps (with 5 levels of features). Inspired by the work in [9, 10], we added *specialized* convolutional layers (with

a 3×3 kernel size) with K (*e.g.* $K = 16$) feature maps after the up-convolutional layers placed at the end of each stage. The outputs of the specialized layers have then the same resolution as the input image, and are then concatenated together. We add a 1×1 convolutional layer at the output of the concatenation layer to linearly combine the fine to coarse feature maps. This complete network provides the final segmentation result.[3]

2.4 Postprocessing

Let us assume that we input the 20 cropped temporal slices of a patient into an image of size $20 \times width \times height$ (where the crop is due to the localization procedure) in preprocessing2 to obtain a $20 \times width \times height \times 3$ image. We filter then the ouput of size $20 \times width \times height$ by keeping only the greatest connected component in the segmented $(2D + t)$-image, and we compute the inverse interpolation on the x and y axes to get back the initial inter-pixel spacing. Finally, we add a zero-valued border to get back a $20 \times 256 \times 256$ or a $20 \times 512 \times 512$ image (depending on the shape of the input).

2.5 Evaluation Methods

The LV quantification as defined in LVquan19 relies on 11 parameters: the areas of the LV cavity and the myocardium, 3 dimensions of the cavity and 6 measurements of the wall thickness. We measure the areas (see Fig. 1 (a)) by computing the number of pixels in the segmented regions corresponding to the LV cavity and the myocardium. To measure the three cavity dimension values (dim1, dim2, dim3) (see Fig. 1 (b)), we proceed this way: because our final segmentation results is the LV myocardium, we first extracted the LV cavity from the segmentation results. We then compute the boundary of the LV cavity and calculate the distances between the points of the boundary and the centroid of the LV cavity along the integral angles $\theta \in [-30, 30[$ (in degrees). Finally, we average these distances. We do this for the six separated regions of the wall. Finally, we compute the mean dimensions for each pair of opposite regions and we obtain $(dim1, dim2, dim3)$. To measure the RWT's values, we first find the boundaries of epicardium and endocardium respectively, and we compute the distances between the points on the boundary of epicardium and the points on the boundary of endocardium along the same integral angles as before where zero corresponds to the normal. Finally, we compute the mean among 60 distance values for each region. To classify the phase as systolic or diastolic, we use a simple method: we detect the time $n_{\max}$ when the cavity is maximal, and $n_{\min}$ when the cavity is minimal. Assuming that we have the case $n_{\min} > n_{\max}$, then for each time $n \in [n_{\max}, n_{\min}]$, we label the image as systolic phase, and otherwise it is a diastolic phase. We do the converse when we have $n_{\max} < n_{\min}$.

3 Experiments

We implemented our experiments on Keras/TensorFlow using a NVidia Quadro P6000 GPU. We used the multinomial logistic loss function for a one-of-many classification

[3] Note that we designed our network's architecture to work with any input shape.

task, passing real-valued predictions through a softmax to get a probability distribution over classes. For the localization network, we used an Adam optimizer (batchsize=4, $\beta 1$=0.9, $\beta 2$=0.999, epsilon=0.001, lr = 0.002) and we did not use learning rate decay. We trained the network during 10 epochs. We recall that we used the filled-in epicardium connected component given in the GT as the "ones" of the output of our network. For the segmentation network, we used the same optimizer and the same parameters but we changed the batchsize to 1. Also, we considered three different classes[4] in the given GT: the background (0), the myocardum (1), the cavity (2) (we merge then 0 and 2 after the segmentation). This way, we obtained better results than using only the wall of the LV.

3.1 Results

We tested our method with 3- and 5-fold-cross-validations on the challenge dataset. An example of bounding box is depicted in red (we did not do any dilation here) in Fig. 5. We obtain an average dice index of 0.8953 on validation set. In practice, we extend next the box by a size equal to 10 pixels to ensure that the whole LV is included into the bounding box.

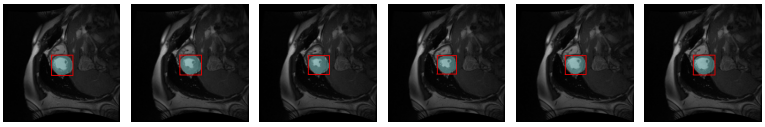


Fig. 5. Some localizations (in red) of the LV (in blue) of the 9^{th} patient.

For the segmentation, we compared ResNet50 with VGG16 as feature extraction on 3-fold-cross-validation (18, 19, 19) (see Fig. 6). VGG16 is then more efficient to detect boundaries than ResNet50 in our application.

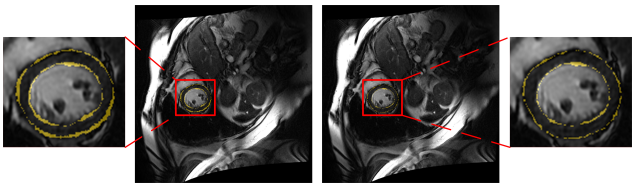


Fig. 6. Segmentation results (ResNet50-FCN on the left side vs. VGG16-FCN on the right side) for one same patient. The yellow color shows the false negatives.

Table 1 presents the average results for the two compared methods. The 11 indices of LV full quantification and dice using the VGG16-FCN are better than when we use

[4] From a technical point of view, we proceeded to a classification more than to a segmentation.

Table 1. Average results of compared methods on 3-fold-cross-validation. Values are shown as mean absolute error.

Dataset	Method	Cavity	Myocardium	Dims(mm)				RWT(mm)							Phase	Dice
		Areas(mm^2)	Areas(mm^2)	dim1	dim2	dim3	average	IS	I	IL	AL	A	AS	average	Error(%)	(%)
Validating data	ResNet50-FCN	279.32	284.84	1.8359	1.6320	1.7767	1.7482	1.2106	1.3059	1.7157	1.6225	1.3303	1.2437	1.4048	15.1267	79.20
	VGG16-FCN (our method)	88.84	157.01	0.9799	1.0691	0.9443	0.9978	0.8320	0.9173	1.1190	1.1124	0.8895	0.8408	0.9518	8.0311	86.04

the ResNet50-FCN. For these reasons, we used the VGG16-FCN for the segmentation of the LV.

To verify the stability of our algorithm, we evaluated the proposed method with 5-fold-cross-validation (11, 11, 11, 11, 12). In Table 2, the average results are showed. Compared with 3-fold-cross-validation, the average areas error is improved from 122.93 mm^2 to 114.77 mm^2, the average dims error is improved from 0.9978 mm to 0.9220 mm, the average RWT error is improved from 0.9518 mm to 0.9185 mm, the average phase error is improved from 8.0311% to 7.6364% and the dice is improved from 86.04% to 86.64%.

Table 2. Average results on 5-fold-cross-validation. Values are shown as mean absolute error.

Dataset	Cavity	Myocardium	Dims(mm)				RWT(mm)							Phase	Dice
	Areas(mm^2)	Areas(mm^2))	dim1	dim2	dim3	average	IS	I	IL	AL	A	AS	average	Error(%)	(%)
Validating data	94.31	135.23	0.9067	0.9792	0.8801	0.9220	0.8362	0.9147	1.0798	1.0560	0.8270	0.7973	0.9185	7.6364	86.64
Testing data	226.80	577.50	6.4934	3.8814	3.9835	4.7861	4.2693	1.8585	2.0570	1.9129	1.6441	3.6039	2.5576	9.83	-

In Table 2, we also reported the results on test dataset given by the organizers of LVQuan19. The test dataset was composed of processed SAX MR sequences of 30 patients. For each patient, only the SAX image sequences of 20 frames were provided (no GT).

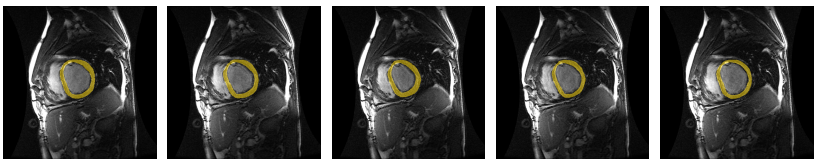


Fig. 7. Some segmentation results on the 5^{th} patient of test dataset.

In Fig. 7, the segmentation results on fifth patient of test dataset are showed, the yellow ring denotes the segmentation results.

4 Conclusion

In this paper, we propose to use a modified VGG16 to proceed to pixelwise image segmentation, in particular to segment the wall of the heart LV in temporal MR images. The

proposed method provides promising results at the same time in matter of localization and segmentation, and leads to realistic physical measures of clinical values relative to the human heart. Our perspective is to try to better segment the boundary of the wall of the LV, either by increasing the weights relative to the boundary regions in the loss function, or by separating the boundary and the interior of the wall into two classes during the classification procedure.

Acknowledgements We thank the organizers of the MICCAI 2019 LV Full Quantification Challenge for providing the LV dataset, NVidia for giving us a Quadro P6000 GPU for this research, and the financial support from China Scholarship Council (CSC, File No.201806290010)

References

1. Xue, W. F., Brahm, G., Pandey, S., Leung, S., Li, S.: Full left ventricle quantification via deep multitask relationships learning. Med. Image Anal. **43**, 54–65 (2018).
2. Xue, W. F., Lum, A., Mercado, A., Landis, M., Warringto, J., Li, S.: Full quantification of left ventricle via deep multitask learning network respecting intra-and inter-task relatedness. In: Descoteaux, M.Maier-Hein, L., Franz, A., Jannin, P., Collins, P.,Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 276–284. Springer, Cham (2017). https://doi.org/10.1007/978–3–319–66179–7_32
3. Xu, Y., Géraud, T., Bloch, I.: From neonatal to adult brain MR image segmentation in a few seconds using 3D-like fully convolutional network and transfer learning, Proc. of ICIP, pp.4417–4421. IEEE, Beijing (2017). https://doi.org/10.1109/$ICIP$.2017.8297117
4. Puybareau, E., Zhao, Z., Khoudli, Y., Carlinet, E., Xu Y. C., Lacotte J., Géraud T.: Left atrial segmentation in a few seconds using fully convolutional network and transfer learning, In: Pop, M., Sermesant M.,Zhao J. C., Li, S., McLeod, K., Young, A., Rhode, K., Mansi, T. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 339–347. Springer, Cham (2018). https://doi.org/10.1007/978–3–030–12029–0_37
5. Payer, C.,Stern, D., Bischof, H., Carlinet, E., Urschler, M.: Multi-label Whole Heart Segmentation Using CNNs and Anatomical Label Configurations, In: Pop M., Sermesant, M., Jodoin, P. M., Lalande, A., Zhuang, X. H., Yang, G., Young, A., Bernard, O.(eds.) STACOM 2017. LNCS, vol. 10663, pp. 190–198. Springer, Cham (2017). https://doi.org/10.1007/978–3–319–75541–0_20
6. Wang, C. J., MacGillivray, T., Macnaught, G., Yang, G., Newby, D.: A two-stage 3D Unet framework for multi-class segmentation on full resolution image. CoRR abs/1804.04341 (2018)
7. Simonyan, K., Zisserman A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
8. Krizhevsky, A., Sutskever, I., Hinton G. E.: ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems, pp. 1097–1105, 2012
9. Long J., Shelhamer E., Darrell T.: Fully convolutional networks for semantic segmentation. Proc. of CVPR, pp.3431–3440. IEEE, Boston (2015).
10. Maninis, K.K.,Pont-Tuset, J.,Arbeláez, P., Van Gool, L.: Deep Retinal Image Understanding. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9351, pp. 140–148. Springer, Cham (2016). https://doi.org/10.1007/978–3–319–46723–8_17