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

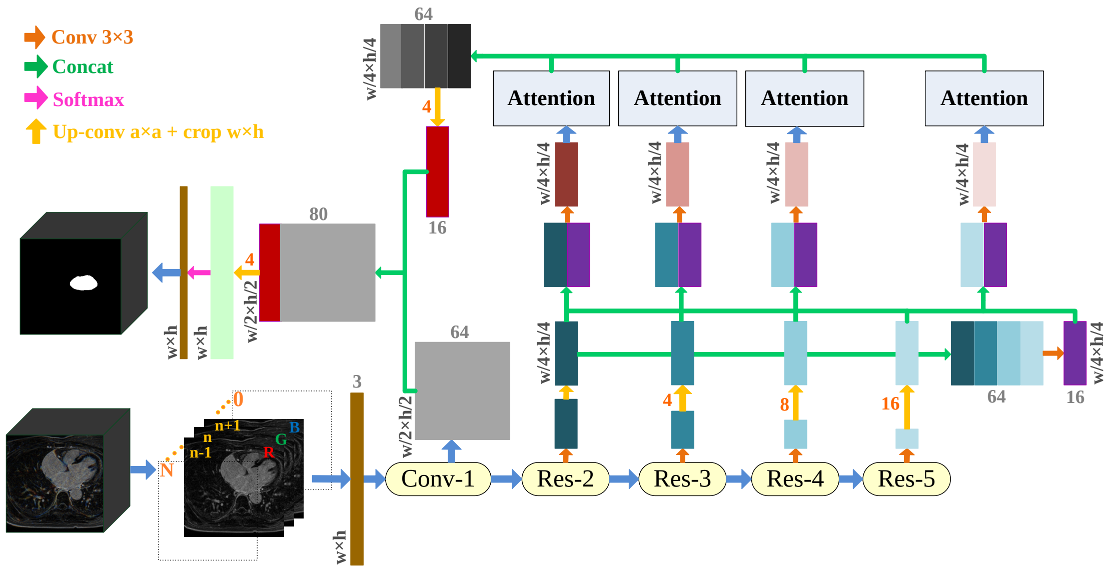


Fig. 1: Architecture of our network.

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 pre-trained 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

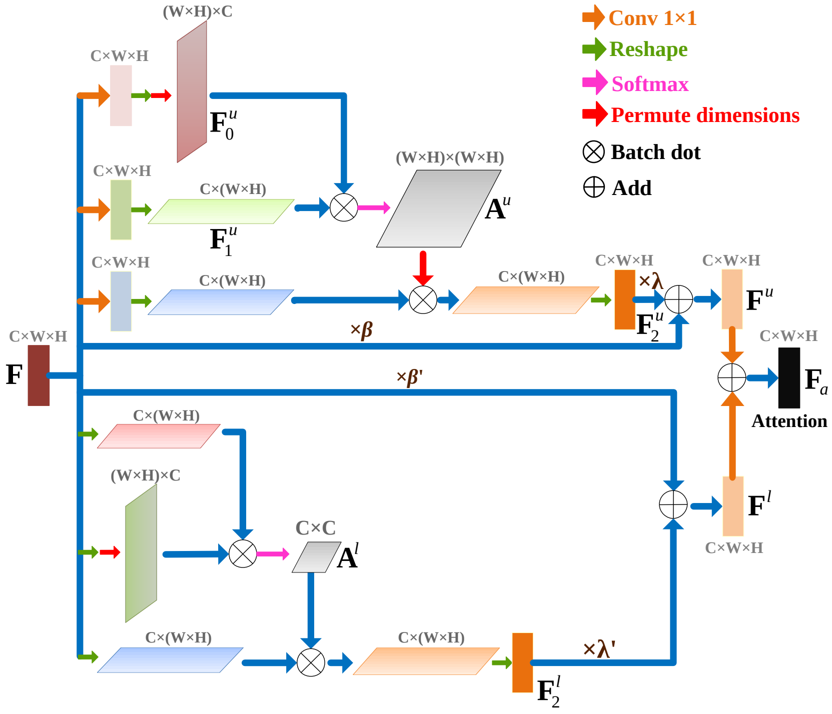


Fig. 2: Attention Module. λ, λ', β and β' as hyperparameters, which is trained like the convolutional kernel. They decrease the weight of the unimportant feature maps.

$\mathbf{F}_1^u \in \mathbb{R}^{C\times(W\times H)}$. 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:

$$\mathbf{F}^u \;=\; \lambda \times \mathbf{F}_2^u + \beta \times \mathbf{F}, \qquad (1)$$

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:

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

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

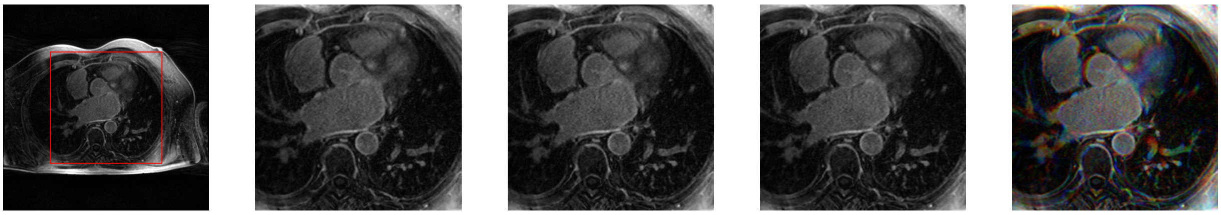
Most of medical segmentation methods directly use Categorical Cross Entropy[35] (CCE) or Dice Coefficient [36] (DC) losses. Models trained with CCE loss usually have low confidence in differentiating boundary pixels, leading to blurry boundaries. DC were proposed for biased training sets but are not 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: $\ell_\text{H} \;=\; \ell_\text{R} + \ell_\text{B}$, 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: $\ell_\text{R} \;=\; \ell_\text{CCE} + \ell_\text{SSIM} + \ell_\text{DC}$, 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 $\ell_{\text{CCE}} = -\sum_{i=1}^{C}\sum_{a=1}^{H}\sum_{b=1}^{W} y_{(a,b)}^{i} \ln {y_*}_{(a,b)}^{i}$, where C is the number of classes of each image, H and W are the height and width of image, $y_{(a,b)}^{i} \in \{0,1\}$ is the ground truth one-hot label of class i at position (a,b) and ${y_*}_{(a,b)}^{i}$ 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 **S** and **G** be the predicted probability map and the ground truth mask respectively, the SSIM loss function of **S** and **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 **S** and **G** respectively, σ_{SG} is their covariance, ε_1= 0.01^2 and ε_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:

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

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

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

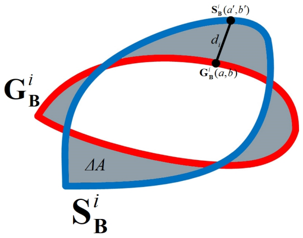

$\mathbf{S}_{\mathbf{B}}^{i}(a',b')$
d
$\mathbf{G}_{\mathbf{B}}^{i}(a,b)$
$\mathbf{G}_{\mathbf{B}}^{i}$
ΔA
$\mathbf{S}_{\mathbf{B}}^{i}$

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 $\mathbf{G}_{\mathbf{B}}^{i}$ of the ground truth of class i and the boundary $\mathbf{S}_{\mathbf{B}}^{i}$ 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

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

where $\mathbf{G}_{\mathbf{B}}^{i}(a,b)$ is a point on boundary $\mathbf{G}_{\mathbf{B}}^{i}$ and $\mathbf{S}_{\mathbf{B}}^{i}(a',b')$ denotes the corresponding point on boundary $\mathbf{S}_{\mathbf{B}}^{i}$, along the direction normal to $\mathbf{G}_{\mathbf{B}}^{i}$, i.e., $\mathbf{S}_{\mathbf{B}}^{i}(a',b')$ is the intersection of $\mathbf{S}_{\mathbf{B}}^{i}$ and the line that is normal to $\mathbf{G}_{\mathbf{B}}^{i}$ 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, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 0.001$, lr $= 0.01$) and did not use learning rate decay. We trained the network during 30 epochs.

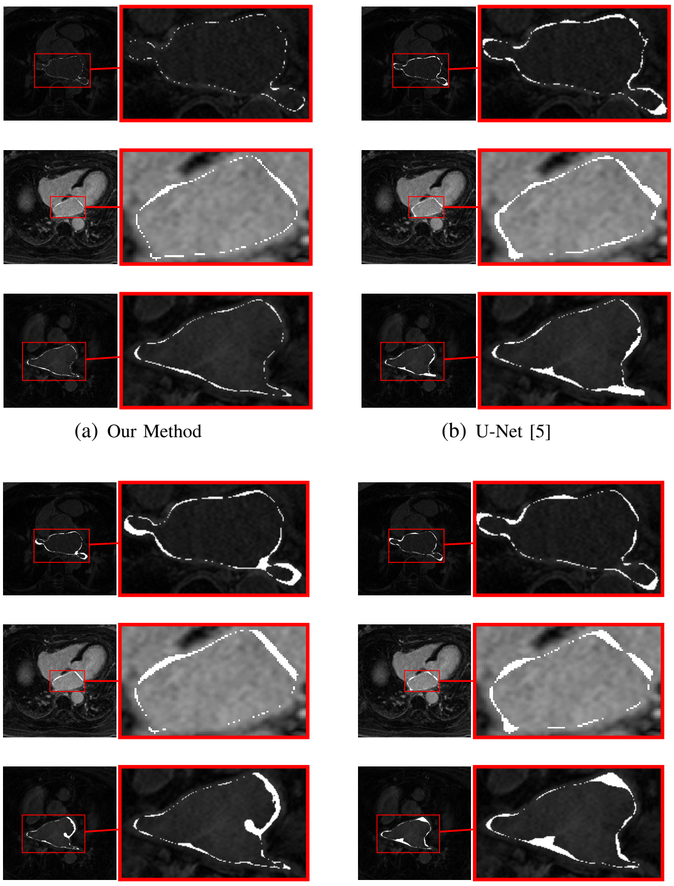
Evaluation Methods. Three metrics are used to evaluate our method: dice to evaluate the regions, and 95% Hausdorff distance (95HD) and Average Hausdorff distance (AHD) to quantitatively evaluate the boundaries.

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.

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.


(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.

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.