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.

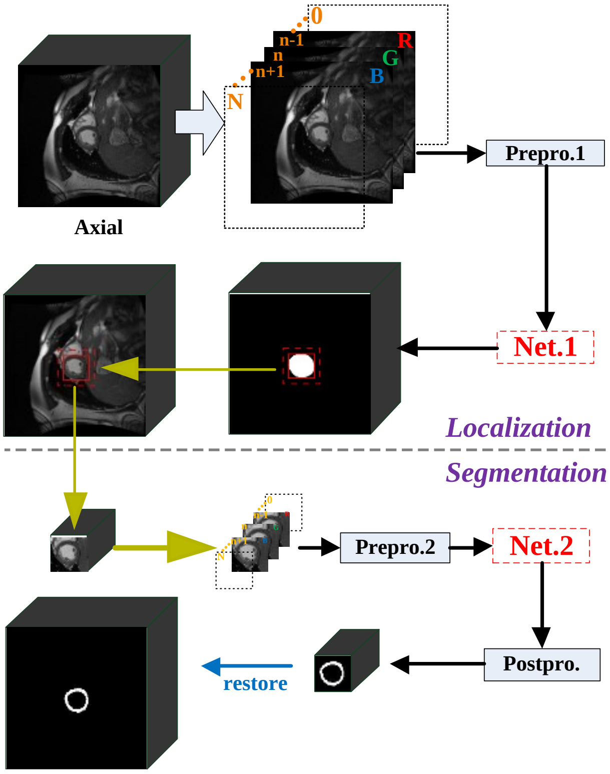


Fig. 1: Global overview of the proposed method (FOANet).

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 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-

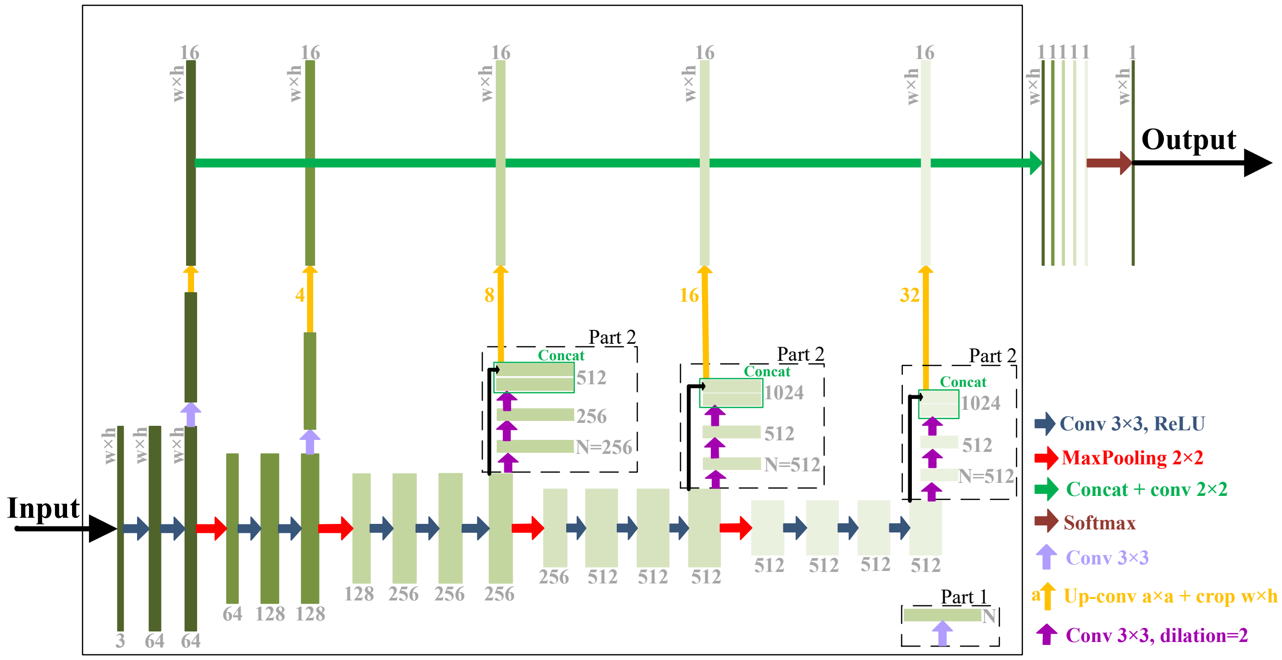


Fig. 2: Architecture of our networks. **Part 1** and **Part 2** correspond to the components of **Net.1** and **Net.2** of Fig. 1, respectively. Because the role of **Net.1** is only to roughly locate the target, using **Part 1** instead of **Part 2** can both reduce model parameters and improve the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

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

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:

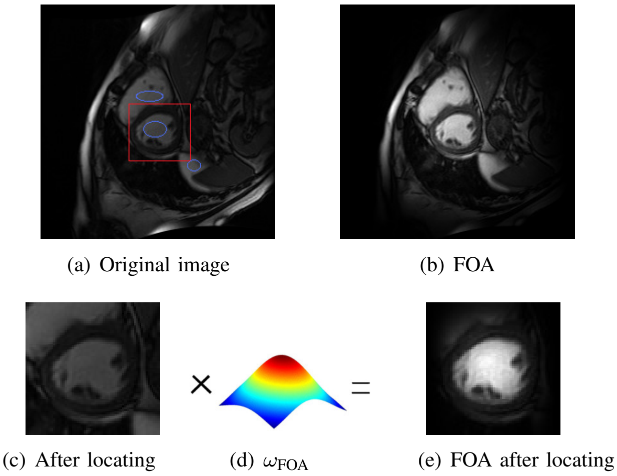
$$\ell_{\text{SSIM}} \; = \; 1 \; - \; \frac{(2\mu_{\text{S}}\mu_{\text{G}} + C_1)(2\sigma_{\text{SG}} + C_2)}{(\mu_{\text{S}}^2 + \mu_{\text{G}}^2 + C_1)(\sigma_{\text{S}}^2 + \sigma_{\text{G}}^2 + C_2)}, \qquad (2)$$

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:

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

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


(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

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

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.

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

[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.

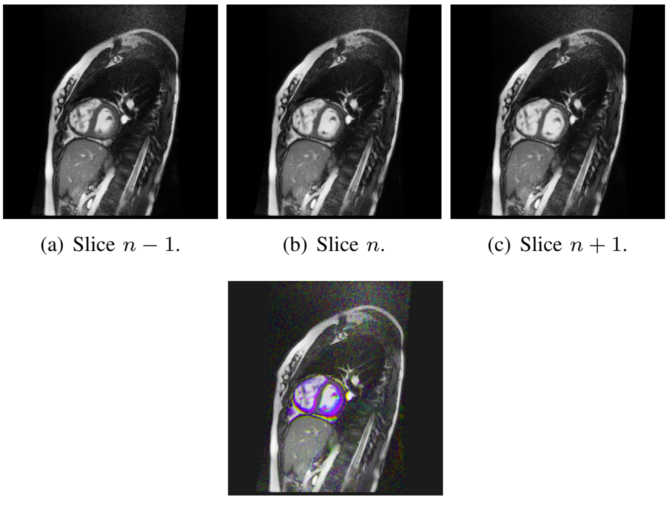
Let us remind what we call *Gauss normalization*: for each $(2D+t)$-image I corresponding to a given patient, we compute $I := (I - \mu)/\sigma$ where μ is the mean of I and σ its standard deviation (σ is assumed not to be equal to zero). There are then two different pre-processing steps as depicted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

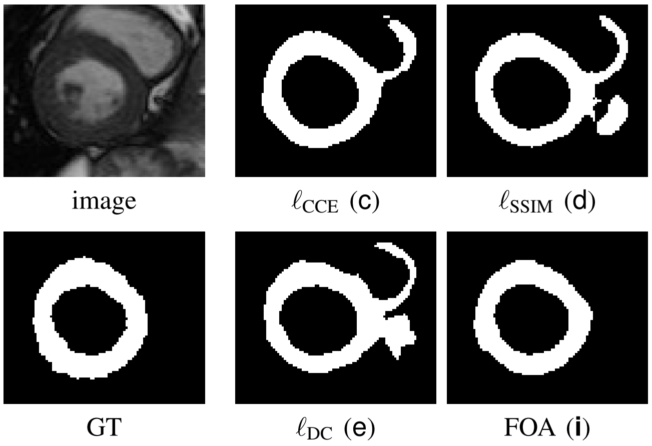
Three measures are used to evaluate our method: DC given in Eq. 3, 95% in the Hausdorff distance (95HD) [24] and Boundary of Dice Coefficient (BDC) to quantitatively evaluate the boundaries. As many diseases appear in the myocardium wall, we chose to quantitatively evaluate the precision of the segmentation 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: $\ell_{\text{CSD}} = \ell_{\text{CCE}} + \ell_{\text{SSIM}} + \ell_{\text{DC}}$



image

ℓ_{CCE} (c)

ℓ_{SSIM} (d)

GT

ℓ_{DC} (e)

FOA (**i**)

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

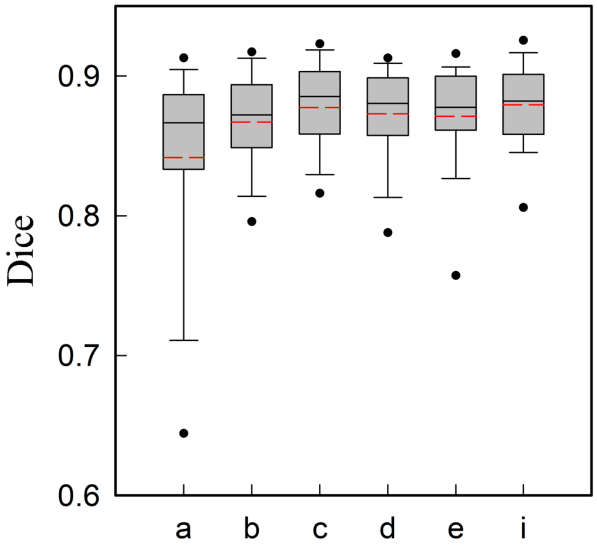


Fig. 6: Box plots of dice scores for the 56 patients. The red dotted line represents the average value, and a, b, c, etc. on the abscissa correspond to the methods of Tbl. I

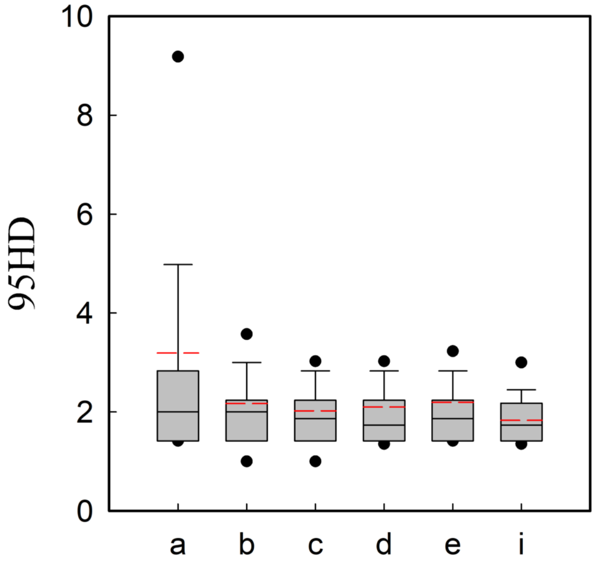


Fig. 7: Box plots of 95HD for the 56 patients. The red dotted line represents the average value, and a, b, c, etc. on the abscissa correspond to Tbl. I

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

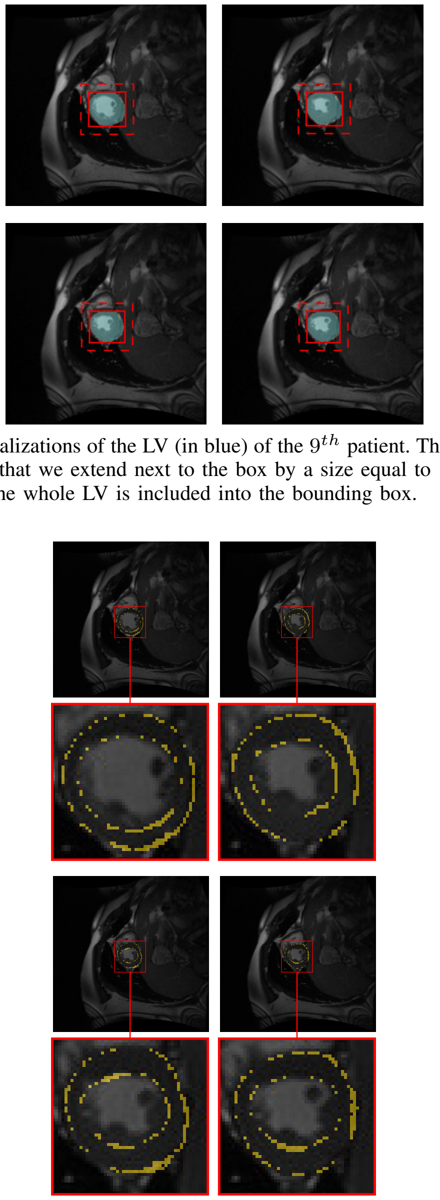
segmented parts are connected to the ground truth, which is very difficult to remove. Therefore, by using our FOA module, we decrease the impact of the surrounding similar tissues, and the segmentation results are better.

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

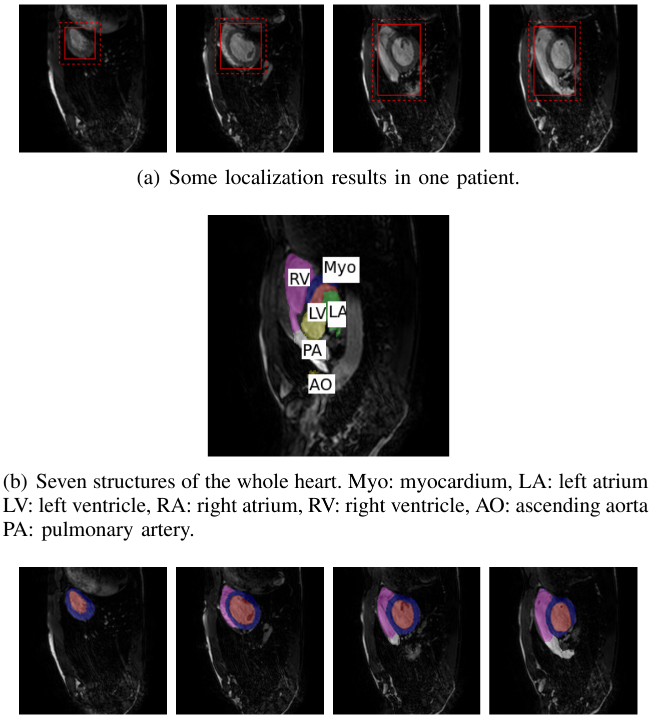
We continued to test our method on the MM-WHS2017 challenge with 5-fold cross-validation and we obtained segmentation results for each class. As we focus in this article on the myocardium segmentation, we will only present our results for this structure. For the comparison with state-of-the-art methods, 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 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 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.



(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, "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.

[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.