Weakly Supervised Whole Cardiac Segmentation via Attentional CNN

Erlei Zhang[1], Minghui Sima[2], Jun Wang[2], Jinye Peng[2(✉)], and Jinglei Li[3(✉)]

[1] Northwest A&F University, No. 22 Xinong Road, Yangling, Shaanxi, China
[2] Northwest University, No. 1, Xuefu Avenue, Xi'an, Shaanxi, China
pjy@nwu.edu.cn
[3] Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China
lijinglei80@126.com

Abstract. Whole-heart segmentation aims to delineate substructures of the heart, which plays an important role in the diagnosis and treatment of cardiovascular diseases. However, segmenting each substructure quickly and accurately is arduous due to traditional manual segmentation being extremely slow, the cost is high and the segmentation accuracy depends on experts' level. Inspired by deep learning, we propose a weakly supervised CNN method to effectively segment the substructure from CT cardiac images. First, we utilize the deformable image registration technology to generate pseudo masks with high confidence for whole heart datasets, which can provide rich feature information to distinguish foreground and background. Meanwhile, the ground truth is used to cut patches containing more heart substructures so that the network can obtain more information about heart substructures. Then, we developed a novel loss function based on the weighted cross-entropy to enforce CNN to pay more attention to the tricky voxels nearby the boundary of cardiac substructures during the training stage. The proposed method was evaluated on MICCAI2017 whole heart CT datasets, and the overall segmentation score of 91.30%.

Keywords: Whole-heart segmentation · Weakly supervised

1 Introduction

The whole heart segmentation is essential for the diagnosis of heart disease. However, the efficiency is limited due to both the annotation of experts and the subjective judgments of doctors. Meanwhile, the segmentation results can only be annotated by doctors and experts, which makes medical images available for research much less than other image datasets. In recent years, deep learning has achieved great success in computer vision and artificial intelligence, which enables the auto segmentation of the substructure of the heart from Computed Tomography (CT) [3]. U-net [9] and Fully Convolutional Network [7] have greatly improved medical image segmentation in terms of accuracy and execution speed, but there exist gradient vanishing and gradient explosion problems when the depth of the network increases. To tackle this problem, Lee et al. [10] added

The original version of this chapter was revised: For the author Jinye Peng a wrong affiliation had been assigned. This has now been corrected. The correction to this chapter is available at https://doi.org/10.1007/978-3-031-14903-0_50

Z. Shi et al. (Eds.): ICIS 2022, IFIP AICT 659, pp. 76–83, 2022.
https://doi.org/10.1007/978-3-031-14903-0_9

depth supervision mechanism into the network, effectively alleviate the problem caused by gradient. Yang et al. [1] applied a deep supervision mechanism to the whole heart segmentation, through integrating DICE loss and cross-entropy loss into the network, they obtained excellent segmentation results. Based on this work, Ye et al. [5] replaced the weighted cross-entropy loss function with the Focal loss function, which makes the model focus on the indistinguishable boundary and improves the Dice accuracy.

For medical images, they contain more background voxels than foreground voxels. Thus, it suffers from the problem of high misclassification. To overcome these limitations, some segmentation frameworks [6, 8] are put forward in recent years. These frameworks, known as cascade networks, are divided into two steps: (1) the first step is to locate the target and simplify the task; (2) the second step is segmentation. Among these frameworks, Payer et al. [8] performed this method on whole heart images and won first place in the MICCAI2017 Whole Heart Segmentation Challenge. However, these frameworks have the disadvantage of excessive or redundant use of parameters, such as repeated extraction of underlying features. Oktay et al. [11] proposed the plug-play Attention Gates (AGS) model, which makes the network automatically focus on relevant areas through training, effectively overcoming the shortcomings of CNNs to some extent. Wu et al. [4] have proposed a WSL (Weakly supervised learning)-based method for brain lesion segmentation. Through weak supervision learning, the network can automatically select the relevant region to suppress the irrelevant image information.

In this paper, we proposed a novel 3D CNN combining WSL learning for cardiac segmentation. We firstly used deformable image registration (DIR) [2] technology to generate pseudo masks of all the CT images for producing weakly supervised information. Then, we utilized that weakly supervised information to guide a novel 3D U-net learning. Furthermore, we developed a novel loss function based on the weighted cross-entropy to enforce CNN to pay more attention to the tricky voxels nearby the boundary of cardiac substructures during the training stage.

The main contributions of this paper are as follows:

(1) We applied traditional medical image registration technology to generate weakly supervised information as the prior knowledge for guiding deep network learning, which not only helps distinguish background and foreground organs but also can be as a data augmentation way avoiding overfitting problems.
(2) We developed an improved weighted cross-entropy loss for enforcing the deep network to pay attention to the missegmented voxels and alleviate the class imbalance problem.

2 Method

2.1 Pseudo Masks

The inputs of the network consist of two parts: one is the original CT image, while another is the pseudo masks that format the one-hot after the background is removed. For the generated pseudo masks, relevant image regions can be automatically selected. Although pseudo masks are not able to segment accurately, they can provide relevant positional features of background and foreground for the region, and effectively extract

heart substructure from background. This paper utilized DIR (deformable image registration) [2] technology to generate pseudo masks for medical images. Set $\{(T_i)\}_{i=1}^{N}$ as N training samples, $\{(V_j)\}_{j=1}^{M}$ as M test samples. There are two training methods, called Model_N-1 and Model_1, as shown in Fig. 2. For a certain training sample T_i, the other N-1 training samples are respectively used as atlas to generate pseudo masks for T_i. In the Model_N-1, we concatenate T_i with its N-1 pseudo masks respectively and put them into deep network for training. In the Model_1, the N-1 pseudo masks of T_i are firstly majority voting to get a final pseudo mask, then we concatenate it with T_i and put them into deep network for training. Thus, similarly, there are two ways to generate test results, called IND and MV, as shown in Fig. 2. IND model is that each training sample is used as atlas to respectively generate pseudo mask for test sample V_j. At testing stage, we concatenate each of N pseudo masks with V_j and pass through the deep network. Then we can obtain N segmentation results for test sample V_j. Finally, we use majority voting method to generate the final segmentation result. MV model is that N pseudo masks of V_j are majority voting to obtain a final pseudo mask, the it is concatenated with V_j and put into the deep network for generating a segmentation result.

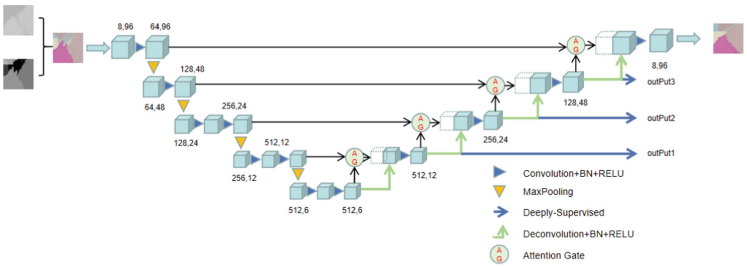

Fig. 1. The framework of the proposed Deep U-net network. In input layer, we concatenated the generated pseudo masks with the cropped patches and placed them into the network for training. The details of pseudo masks generation and patch cropping will be introduced in Sect. 2.1 and 2.2.

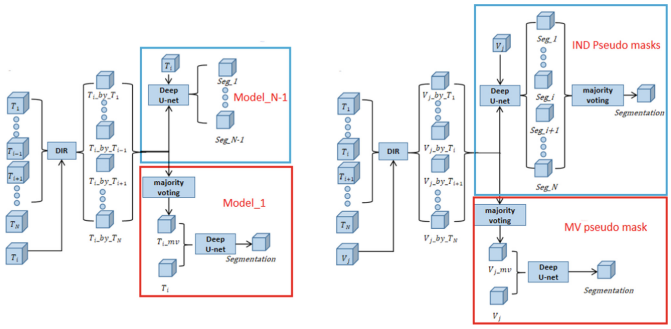

Fig. 2. Two training methods (left) of pseudo masks, two test methods (left) of pseudo masks.

2.2 Deep U-Net Network

In order to better train the deep network, we adopt the method of deep supervision, which increases the output path in different network layers and shortened the backpropagation path of gradient flow. In this paper, three deep supervised branches are introduced in the decoding stage. The output of each branch is the same as that of the main branch, in Fig. 1, out1, out2, and out3 are the three deep supervised branches, and the final total loss is the sum of the losses of each branch and the main branch.

2.3 Improved Weighted Cross-Entropy Loss

The commonly used weighted cross-entropy loss does not perform well for voxels that are difficult to segment. In this paper, we added predicted false negative (FN) and true positive (TP) voxels losses into the weighted cross-entropy to formula the total loss. As shown in Eq. (1).

$$L_{mw}Cross(x, y, z) = -\sum_{c}^{C}\sum_{i=1}^{N} w_c\left[\left(G_c^i + G_{cFN}^i\right)logP_c^i + P_{cTP}^i log\left(1 - P_c^i\right)\right] \quad (1)$$

where G_{cFN}^i is 0 or 1, where 1 indicates that the current voxel belongs to class c but is predicted to be of another class. P_{cTP}^i is 0 or 1, where 1 indicates that the current voxel is predicted to be class c, but is actually something else. $P_c^i(0.005 < P_c^i < 0.995)$ is the probability that the current voxel is class c, and the range is limited to prevent the excessive loss, which is not conducive to network convergence. w_c is the weight coefficient of class c, which can be used to alleviate class imbalance.

MDSC (Multi-Class Dice Similarity Coefficient) based loss function to balance the training for multiple classes [1]. This loss can be defined as:

$$L_{mDSC} = -\sum_{c=1}^{C} \frac{\frac{2}{N}\sum_{i=1}^{N} G_c^i P_c^i}{\sum_{i=1}^{N} G_c^i G_c^i + \sum_{i=1}^{N} P_c^i P_c^i} \quad (2)$$

where N is the number of voxels; G_c^i is a binary value, where 1 indicates the voxels belong to class c, 0 stands for other categories; P_c^i $(0 < P_c^i < 1)$ denotes the probability that the current voxels belong to class c.

After and are added into the network, the new loss function can be defined as follows:

$$L_{out\_x}(d, w) = 100dL_{mDSC} + wL_{mw}Cross \quad (3)$$

where d and w are the weights of different branches, x represents the output of the deep supervised branch, the final loss function, called the Improved Weighted Cross-Entropy (IWCE), is:

$$L_{total} = L_{out\_1}(0.2, 0.3) + L_{out\_2}(0.4, 0.6) + L_{out\_3}(0.8, 0.9) + L_{out\_4}(1.0, 1.0) \quad (4)$$

3 Experimental and Results

3.1 Datasets and Implementation Details

We evaluated our approach with the MICCAI2017 whole-heart CT datasets, which contains 20 publicly available CT data [1]. We randomly selected 10 samples as training samples and the rest as test sets. These data were collected in the actual clinical environment, which was variable and contained some images of poor quality, so the robustness of the algorithm in this paper remains to be verified. Each sample is stacked with multiple 2D images of $512 * 512$ size. All training data were normalized to zero mean and unit variance. Adam is used to optimize network parameters, the number of iterations was 35,000 epochs [5], the batch size was 2, and the initial learning rate was 0.001.

3.2 Patch Selection

Due to the particularity of heart medical images, and the 7 substructures voxels in whole heart CT image account for less. When the random cropped size is 96, the background occupied more than half of the training data, which is not conducive to the better learning prospects of the network. To tackle this problem, we adopted an effective cropped method, which utilized ground truth to crop the patches with less background. For the randomly cropped patches, we calculated the proportion p of the background voxels in the whole patch. If the background proportion p is less than a ($a = 0.5$), this patch will be called the available patch and sent into the network for training, otherwise, the patch will be re-cropped.

3.3 Experimental Results

We took deeply-Supervised U-net [1] as the baseline network, Multi-Depth Fusion [5] is an improvement of the baseline network and Dice score as performance evaluation. In order to the efficiency of the proposed method in this paper, we conducted a series of ablation experiments.

The experimental results of cardiac substructure, pulmonary artery (PUA), ascending aorta (ASA), right ventricular blood chamber (RVBC), right atrial blood chamber (RABC), left ventricular blood chamber (LVBC), left atrial blood chamber (LABC), and myocardium of the left ventricle (MLV) were shown in Table 1. Except for the PUA (Dice score about 82%–86%), we can see that all the methods achieved relatively accurate substructures' segmentation for the whole heart. The reason could be that the shape and appearance of the PUA always has greater variability.

Compared with the baseline method, the proposed the four methods with the pseudo masks can produce better segmentation results in almost substructures of the whole heart. And all the proposed four methods have comparable performance with the advanced Multi-Depth Fusion method. Although, these regions of MLV (has the epicardial surface and the endocardial surface of the left ventricular) and RABC have much larger variation in terms of shapes and heterogeneous intensity of the myocardium and the blood. All the proposed methods outperform the two compared methods on the MLV and RABC. Particularly, "MV + Model_1" achieves the best results on MLV, RVBC, ASA, and PUA.

Table 1. Segmentation accuracy (%) of the state-of-the-art segmentation methods and the proposed four methods. "IND + Model_N-1" indicated that it used Model_N-1 at training stage and IND model at testing stage; "IND + Model_1" indicated that it used Model_1 at training stage and IND model at testing stage; "MV + Model_N-1" indicated that it used Model_N-1 at training stage and MV model at testing stage; "MV + Model_1" indicated that it used Model_1 at training stage and MV model at testing stage. The Bold Font in the proposed four methods means it outperform the Baseline and Multi-Depth Fusion methods. The values with underline mean that they are the best results in the six methods.

Method	MLV	LABC	LVBC	RABC	RVBC	ASA	PUA	Mean
Baseline	87.6	90.5	92.1	86.0	88.6	94.8	82.6	88.93
Multi-Depth Fusion	88.9	91.6	94.4	87.8	89.5	96.7	86.2	90.73
IND + Model_N-1	**89.9**	90.7	94.2	**89.6**	89.4	93.0	**87.0**	90.56
IND + Model_1	**90.2**	90.8	**94.4**	**89.6**	**89.8**	94.0	85.7	90.68
MV + Model_N-1	**89.5**	91.1	94.2	**90.0**	**89.9**	96.5	**86.3**	**91.14**
MV + Model_1	**89.8**	91.3	94.1	**89.9**	**90.0**	**96.9**	**86.9**	**91.30**

3.4 Ablation Experiments

We verify the effectiveness of the proposed IWCE LOSS, patch selection, and pseudo mask modules in the proposed model. We used the best model "MV + Model_1" as the basic model "Model". Then, we ablate or replace each proposed module, respectively. Other experimental conditions are the same as the Table 1.

Table 2 shows the experimental results. We can see that the segmentation results of six substructures become worse after the model without using the Patch Selection module. It proved that the Patch Selection module can select meaningful image patch conducive to the better learning prospects of the network. The third row is the best model using traditional Cross-Entropy loss without using the proposed IWCE loss. We can see that the segmentation results of the almost substructures are slightly worse than the best model. It proved that the proposed loss function takes the class imbalance problem into account and perform well for the voxels, like PUA, that are difficult to segment. The forth row is the model without using pseudo mask information for training, we can see that it achieved comparable performance on five substructures except ASA (reduce ~1%) and PUA (reduce ~3%). One reason is that the pseudo masks generated by simple DIR have lower quality which introduced very limit information for guiding deep network learning on some substructures that are easy to segment. Other reason is that the pseudo masks can provide some useful information, such as location information, for the PUA segmentation.

Table 2. Ablation experiment for the effect of the modules in the proposed MV + model_1 model. "PS" refers to Patch Selection modules; "IWCE" refers to the proposed mixing loss; "pseudo mask" refers to the proposed pseudo mask label modules. "↓" or "↑" denote the increase or decrease of the Dice score (%) compared with the values of "MV + Model_1" method.

Method	MLV	LABC	LVBC	RABC	RVBC	ASA	PUA	Mean
MV + Model_1	**89.8**	**91.3**	**94.1**	**89.9**	**90.0**	**96.9**	**86.9**	**91.30**
Model without PS	89.2↓	90.9↓	92.5↓	89.9	90.2↑	96.5↓	86.4↓	90.90↓
Model without IWCE	89.5↓	91.0↓	94.1	89.6↓	89.1↓	96.2↓	85.6↓	90.82↓
Model without pseudo mask	90.2↑	90.6↓	94.1	89.7↓	90.1↑	95.9↓	83.9↓	90.75↓

Table 3. Generality of the proposed modules. "Baseline" method is the deeply-Supervised U-net [1]; "Baseline PS" is the combination of the baseline method and Patch selection module; "Baseline IWCE" refers to the baseline method whose lose function is replaced for the IWCE loss function; "Baseline Pseudo mask" refers to the baseline method integrates the pseudo mask information during training stage. "↓" or "↑" denote the increase or decrease of the Dice score compared with the values of "Baseline" method.

Method	MLV	LABC	LVBC	RABC	RVBC	ASA	PUA	Mean
Baseline	87.6	90.5	92.1	86.0	88.6	94.8	82.6	88.93
Baseline PS	89.91↑	90.14↓	94.08↑	89.39↑	89.98↑	94.68↓	84.69↑	90.41↑
Baseline IWCE	88.70↑	89.89↓	93.66↑	88.86↑	89.99↑	96.57↑	85.74↑	90.49↑
Baseline Pseudo mask	89.29↑	90.48↓	93.16↑	89.71↑	89.64↑	96.57↑	86.63↑	90.78↑

3.5 Generality Experiments

In order to analysis and discuss the generality of the proposed modules including the Patch Selection, IWCE loss, and pseudo masks, we use the deeply-Supervised U-net [1] as the baseline segmentation network and combine it with the proposed modules respectively. Table 3 shows the experimental results. We can see that the baseline method with each proposed module has a positive effective on most substructures except LABC. Especially, the performance of the baseline with pseudo mask method has significant improvement on PUA. It further proved that the pseudo masks can provide certain prior information which is useful for the hard to segment problem.

4 Conclusion

In this paper, a weakly supervised segmentation method based on CNN is proposed for whole-heart segmentation. We first generate pseudo masks using traditional deformable image registration methods, then perform them on whole-heart data for training. The information provided by pseudo masks is used to distinguish foreground and background. In order to obtain better experimental results, we improved the weighted cross-entropy

loss function and mined the training samples to solve the problems of fuzzy boundary and class imbalance. We performed validation on the MICCAI 2017 whole-heart CT dataset, and the results demonstrate that our method can effectively improve the accuracy of heart segmentation.

Acknowledgements. Guangdong Basic and Applied Basic Research Foundation No. 2022A1515011650, the QinChuangyuan high-level innovation and entrepreneurship talent program of Shaanxi (2021QCYRC4-50). Supported by the International Science and Technology Cooperation Research Plan in Shaanxi Province of China (No. 2022KW-08).

References

1. Yang, X., Bian, C., Yu, L., Ni, D., Heng, P.-A.: Hybrid loss guided convolutional networks for whole heart parsing. In: Pop, M., et al. (eds.) STACOM 2017. LNCS, vol. 10663, pp. 215–223. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0_23
2. Andrade, N., Faria, F.A., Cappabianco, F.A.M.: A practical review on medical image registration: from rigid to deep learning based approaches. In: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images, pp. 463–470. IEEE (2018)
3. Zhuang, X., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med. Image Anal. **58**, 101537 (2019)
4. Wu, K., Du, B., Luo, M., Wen, H., Shen, Y., Feng, J.: Weakly supervised brain lesion segmentation via attentional representation learning. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 211–219. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_24
5. Ye, C., Wang, W., Zhang, S., Wang, K.: Multi-depth fusion network for whole-heart CT image segmentation. IEEE Access **7**, 23421–23429 (2019)
6. Ammar, A., Bouattane, O., Youssfi, M.: Automatic cardiac cine MRI segmentation and heart disease classification. Comput. Med. Imaging Graph. **88**, 101864 (2021)
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
8. Payer, C., Štern, D., Bischof, H., Urschler, M.: Multi-label whole heart segmentation using CNNs and anatomical label configurations. In: Pop, M., et al. (eds.) STACOM 2017. LNCS, vol. 10663, pp. 190–198. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0_20
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. Lee, C. Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial Intelligence and Statistics, pp. 562–570. PMLR (2015)
11. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)