Multi-sequence Cardiac MR Segmentation with Adversarial Domain Adaptation Network

Jiexiang Wang, Hongyu Huang, Chaoqi Chen, Wenao Ma, Yue Huang, and Xinghao Ding[✉]

School of Information Science and Engineering, Xiamen University, Xiamen, China
{wangjx,huanghy,cqchen94,wenaoma}@stu.xmu.edu.cn,
{yhuang2010,dxh}@xmu.edu.cn

Abstract. Automatic and accurate segmentation of the ventricles and myocardium from multi-sequence cardiac MRI (CMR) is crucial for the diagnosis and treatment management for patients suffering from myocardial infarction (MI). However, due to the existence of domain shift among different modalities of datasets, the performance of deep neural networks drops significantly when the training and testing datasets are distinct. In this paper, we propose an unsupervised domain alignment method to explicitly alleviate the domain shifts among different modalities of CMR sequences, *e.g.,* bSSFP, LGE, and T2-weighted. Our segmentation network is attention U-Net with pyramid pooling module, where multi-level feature space and output space adversarial learning are proposed to transfer discriminative domain knowledge across different datasets. Moreover, we further introduce a group-wise feature recalibration module to enforce the fine-grained semantic-level feature alignment that matching features from different networks but with the same class label. We evaluate our method on the multi-sequence cardiac MR Segmentation Challenge 2019 datasets, which contain three different modalities of MRI sequences. Extensive experimental results show that the proposed methods can obtain significant segmentation improvements compared with the baseline models.

1 Introduction

Accurate segmentation of the ventricles and myocardium is fundamental to the diagnosis and treatment of myocardial infarction (MI) [17]. Cardiac MRI sequences are usually used for the MI diagnosis, in particular the T2-weighted MRI detect damaged and ischemic areas, the balanced-Steady State Free Precession (bSSFP) MRI clearly shows the heart structure boundary, and the late gadolinium enhancement (LGE) MRI can enhance infarcted myocardium with distinctive brightness compared to healthy structure [16]. Manual segmentation is time-consuming, so automatic segmentation is significant in the clinic.

J. Wang and H. Huang—Indicates equal contributions.

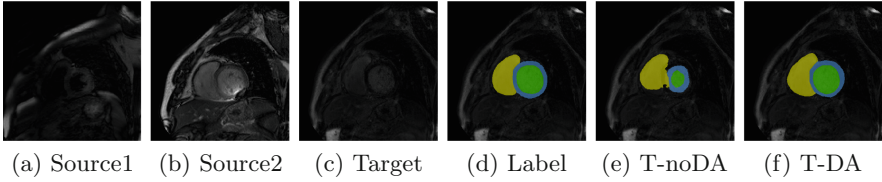(a) Source1   (b) Source2   (c) Target   (d) Label   (e) T-noDA   (f) T-DA

Fig. 1. Performance drops due to domain shift. (a) Original T2-weighted MRI (Source1). (b) Original bSSFP MRI (Source2). (c) Original LGE MRI (Target). (d) LGE MRI annotation (Label). (e) The segmentation results of LGE MRI using an established model trained on T2-weighted and bSSFP MRI data (T-noDA). (f) The segmentation results of an LGE MRI using our model trained on T2-weighted and bSSFP MRI data (T-DA). The yellow region denotes the right ventricle, the green region denotes the left ventricle, and the blue region denotes the myocardium. (Color figure online)

Recently, deep learning network has become a powerful tool for semantic segmentation on heart structures [12,13]. Obviously, the ventricles and myocardium segmentation results can be improved combining the complimentary information from T2-weighted and bSSFP MRI sequences [16]. To save labeling time, sometimes only the T2-weighted and bSSFP MRI sequences and corresponding labels are available. However, a well-trained segmentation model may underperform when being tested on data from different modalities, which is caused by the domain shift (as shown in Fig. 1). Fine-tuning on the target domain data is a simple but efficient method to alleviate the performance drop. But it still requires massive data collection and enormous annotation workload which are impossible for many real-world medical scenarios. For this reason, constructing a general segmentation model suitable for various modalities is promising yet still challenging.

Unsupervised Domain Adaptation (UDA) methods have shown compelling results on reducing the dataset shift across distinct domains. Prior efforts on this problem intended to match the source and target data distributions to learn a domain-invariant representation. For example, Maximum Mean Discrepancy (MMD) was introduced to minimize the distance of source and target feature distributions in Reproducing Kernel Hilbert Space (RKHS) [11]. CycleGAN [15] tackled the image-to-image translation task in a fully unsupervised manner, and thus is capable of reducing the domain shift in the pixel-level. AdaptSegNet [10] solved the unsupervised cross domain segmentation problem by leveraging the domain adversarial training approach. In the context of medical imaging, [3] developed an UDA framework based on adversarial networks for lung segmentation on chest X-rays. [8] improved the UDA framework with Siamese architecture for Gleason grading of histopathology tissue. [5] proposed a domain critic module and a domain adaptation module for the unsupervised cross-modality adaptation problem. These approaches, which based on the domain adversarial training, required empirical feature selection. [2] proposed the synergistic fusion of adaptations from both image and feature perspectives for heart structures

segmentation. However, this approach, which based on image-to-image adaptation, cannot be directly introduced to the multiple source domain adaptation problems due to the presence of multiple domain shifts between different source domains.

In this paper, we propose a domain alignment method for the UDA problem, which helps the established model segment the ventricles and myocardium accurately in the target domain without requiring target labels. Firstly, in order to reduce the domain shift with respect to the image appearance, we propose a histogram match operation for all the data. Secondly, we introduce the domain adversarial training in the output space, which can directly align the predicted segmentation results across different domains. Finally, we further propose a group-wise feature recalibration module (GFRM) to improve the domain adversarial training by integrating multi-level features without requiring manual selection to progressively align the source and target feature distributions. The proposed method is extensively evaluated on the multi-sequence cardiac MR Segmentation (MS-CMRSeg) Challenge 2019 datasets, including bSSFP, LGE and T2-weighted MRI sequences.

2   Method

Figure 2 overviews our segmentation method for ventricles and myocardium in MRI sequences. We use modified 2D attention U-Net with pyramid pooling module as our segmentation backbone architecture [7,14]. To align the distance over feature and output spaces across different domain, feature-level and mask-level discriminator are adopted. Moreover, the group-wise feature recalibration module (GFRM) is introduced to transfer multi-level feature information. The details of the above modules are shown in Fig. 3.
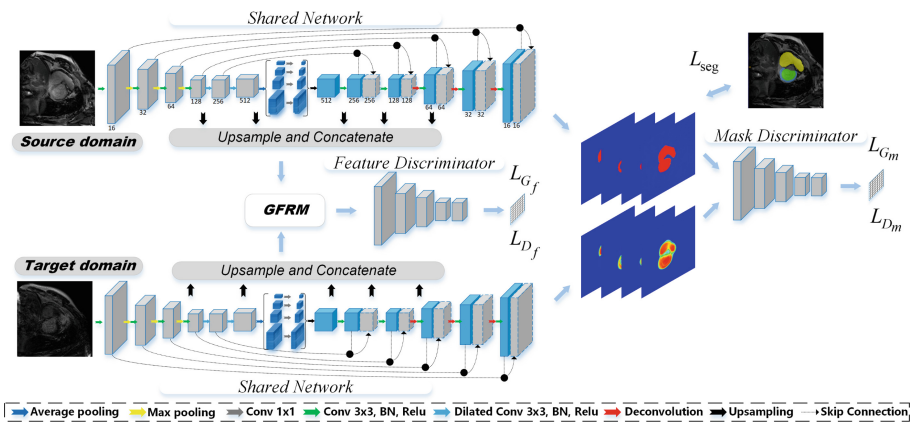

Fig. 2. Schematic view of our proposed framework.

2.1   Network Architecture

Segmentation Network. It is essential to build upon a good baseline model to achieve high-quality segmentation results. Our segmentation network follows the spirit of attention U-Net architecture [7]. In encoder network, we keep the convolution layer as the initial setting. We perform three maxpool operations totally. Dilated convolution is adopted after third maxpool operation to capture large receptive field to alleviate loss of structural information. Inspired by [14], pyramid pooling module is introduced to generate multi-scale features to alleviate the variance of heart size over each patient. In decoder network, we perform three deconvolution operations totally. For further accurate segmentation results, attention gate (as the black dot shown in Fig. 3(a)) is utilized to learn to focus on ventricles and myocardium structures. In attention gate, the features in the encoder part (as the blue rectangle shown in Fig. 3(a)) and decoder part (as the gray rectangle shown in Fig. 3(a)) are first squeezed with 3×3 convolution layer along the channel direction respectively and then added together. After that, we squeeze the features to single channel feature map to form structure attention with 1×1 convolution layer and generate final feature maps by dot product. Finally, we use 1×1 convolution layer with four output channels followed by the *sigmoid* activation function to generate the probability maps. To save computational resources, we share the network with the same parameters between source and target domain.

Fig. 3. Architecture of the sub-networks in our framework. (Color figure online)

Group-wise Feature Recalibration Module. Before we perform group-wise feature recalibration operation, different size features from segmentation network above are expanded and concatenated via upsampling and concatenating operations. The features are send to GFRM. Our GFRM follows the spirit of [9]. Different from the above method, we divide features into four groups corresponding to the segmentation categories to focus on specific heart structures and we recalibrate features in each group (as shown in Fig. 3(b)). GFRM consists of two parts: channel attention part and spatial attention part. In channel attention

part, we first squeeze global spatial information with global average pooling and fully connection layer. Then, we can generate the channel-wise attention features by simple dot product. In the spatial attention part, we first squeeze channel information with 1×1 convolution layer. Then, we can obtain the spatial-wise attention features by simple dot product. The features from channel attention part are added with the features from spatial attention part to generate group-wise recalibrated features. Finally, the features from each group are concatenated to generate final recalibrated features.

Discriminator. The feature-level and mask-level discriminator are based on the multi-level features from GFRM and predicted mask results. We use Patch-GAN as our discriminator [6]. The network consists of 3 convolution layers with stride of 2 and 2 convolution layers with stride of 1. The kernel size of all convolution layers is 4×4 and the corresponding channel number is $64, 128, 256, 256, 1$. Except for the last layer, each convolution layer is followed by a leaky ReLU parameterized by 0.2.

2.2   Hybrid Loss Function for Source Data

Since the labels for source domain are available, we train the segmentation network with a hybrid loss. The vanilla cross-entropy loss with our unbalanced training data leads to low accuracy. We add the Jaccard loss [1] into our loss function. The training objective for source data is

$$\mathcal{L}_{ce}^{s} = -\mathbb{E}_{x_s \sim S}(\sum_{i=1}^{N_s} \sum_{c=1}^{C} y_{s,i,c} \log G(x_{s,i}; \Theta_g)) \tag{1}$$

$$\mathcal{L}_{jac}^{s} = -\mathbb{E}_{x_s \sim S}(\sum_{i=1}^{N_s} \sum_{c=1}^{C} \frac{y_{s,i,c} G(x_{s,i}; \Theta_g)}{y_{s,i,c} + G(x_{s,i}; \Theta_g) - y_{s,i,c} G(x_{s,i}; \Theta_g)}) \tag{2}$$

S represents source domain; For each source image x_s, there is one corresponding annotation y_s; N_s is the number of all source images; $\mathbb{E}_{x_s \sim S}$ means that all x_s are from S; C is the number of all categories; G is segmentation network; Θ_g is the parameters of G; $y_{s,i,c}$ and $G(x_{s,i}; \Theta_g)$ mean the annotation and prediction vectors, respectively. For cross entropy loss, the imbalance of training data leads to a local optimum with inappropriate direction of gradient decreasing, especially in the early stage. The Jaccard loss effectively helps to avoid the local optimum due to its better perceptual quality and scale invariance [1].

2.3   Adversarial Learning for Target Data

In the target domain, due to the lack of annotations, we leverage the adversarial learning to train the segmentation network by minimizing the discrepancy across the source and target domain. Domain adaptation based on both feature and output space is proved to be effective for heart structure segmentation [4].

In our framework, we employ two discriminators. The features input to feature domain discriminator are selected empirically in [4]. To overcome this problem, we propose the GFRM to leverage the full feature spectrum and automatically select prominent features in the feature space. In the segmentation network, each feature scale generates one output feature map in the same dimension via convolution and upsampling operations. The feature maps are further processed by the GFRM to highlight the prominent features and suppress the irrelevant ones. The combined feature maps are then fed to the feature discriminator network for the adversarial learning, where the losses are defined as

$$
\begin{aligned}
\mathcal{L}_{adv_{D_f}} = &- \mathbb{E}_{x_s \sim S} \log D_f(R(G(x_s; \Theta_g); \Theta_r); \Theta_{d_f}) \\
&- \mathbb{E}_{x_t \sim T}(1 - \log D_f(R(G(x_t; \Theta_g); \Theta_r); \Theta_{d_f}))
\end{aligned}
\tag{3}
$$

$$
\mathcal{L}_{adv_{G_f}} = -\mathbb{E}_{x_t \sim T} \log D_f(R(G(x_t; \Theta_g); \Theta_r); \Theta_{d_f})
\tag{4}
$$

T represents target domain; where x_t is target data; $\mathbb{E}_{x_t \sim T}$ means that all x_t are from T; R is the GFRM; Θ_r is the parameters of R; D_f is the feature discriminator; Θ_{d_f} is the parameters of D_f.

In the output space, the segmentation results of target domain should be similar to the ones of source domain. To achieve this, we employ the adversarial learning technique in the output space, where the losses are defined as

$$
\begin{aligned}
\mathcal{L}_{adv_{D_m}} = &- \mathbb{E}_{x_s \sim S} \log D_m(G(x_s; \Theta_g); \Theta_{d_m}) \\
&- \mathbb{E}_{x_t \sim T}(1 - \log D_m(G(x_t; \Theta_g); \Theta_{d_m}))
\end{aligned}
\tag{5}
$$

$$
\mathcal{L}_{adv_{G_m}} = -\mathbb{E}_{x_t \sim T} \log D_f(G(x_t; \Theta_g); \Theta_{d_m})
\tag{6}
$$

where D_m is the mask discriminator; Θ_{d_m} is the parameters of D_m.

Combined with the aforementioned loss, the full objective function

$$
\begin{aligned}
\mathcal{L}_{FULL} = &\lambda_{ce}\mathcal{L}_{ce} + \lambda_{jac}\mathcal{L}_{jac} + \lambda_{D_f}\mathcal{L}_{adv_{D_f}} \\
&+ \lambda_{G_f}\mathcal{L}_{adv_{g_f}} + \lambda_{D_m}\mathcal{L}_{adv_{D_m}} + \lambda_{G_m}\mathcal{L}_{adv_{g_m}}
\end{aligned}
\tag{7}
$$

3   Experiment

Dataset. The validation of the proposed method is performed in the MS-CMRSeg Challenge 2019 dataset covering 45 patients. There are bSSFP, T2-weighted and LGE MRI sequences in each patient data. In one patient data, the slice number and annotation of three MRI modalities are different. We combine labeled bSSFP and T2-weighted MRI sequences as source data, and unlabeled LGE MRI sequences as target data. Experienced experts manually annotated the left ventricle (LV), right ventricle(RV) and myocardium (Myo) as ground truth. We pre-processing the data for domain adaptation. The data is resized and cropped to 400×400 in the center of each slice. In order to eliminate the inconsistency in appearance, we perform histogram match operation on both source and target data, as shown in Fig. 4.
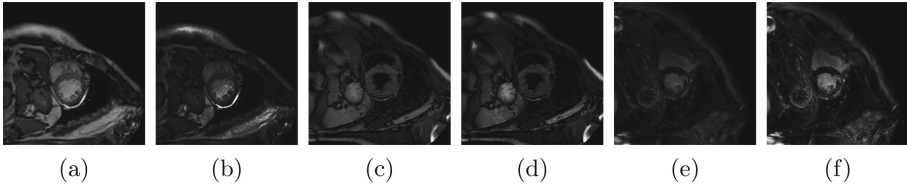
Fig. 4. Visual comparison for histogram match operation: (a) T2-weighted MRI. (b) T2-weighted MRI after histogram match. (c) bSSFP MRI. (d) bSSFP MRI after histogram match. (e) LGE MRI. (f) LGE MRI after histogram match.

Implementation Details. In our experiments, we implement our whole network with PyTorch, using a standard PC with a single NVIDIA 1080Ti. To train the segmentation network, we use the Stochastic Gradient Descent (SGD) optimizer with Nesterov acceleration where the momentum is 0.9 and the weight decay is $1e^-4$. The initial learning rate is set as 0.01 and is decreased to 0.001 after 80 epochs. For training the both feature and mask discriminator, we use Adam optimizer with the fixed learning rate as 0.0002. The weight decay is set as $5e^-5$. We totally trained 150 epochs with a mini-batch size of 8. We set λ_{ce}, λ_{jac}, λ_{G_f}, λ_{D_f}, λ_{G_m} and λ_{D_m} to $0.5, 0.5, 0.05, 1.0, 0.005$ and 1.0. The training time cost only 5 h to converge.
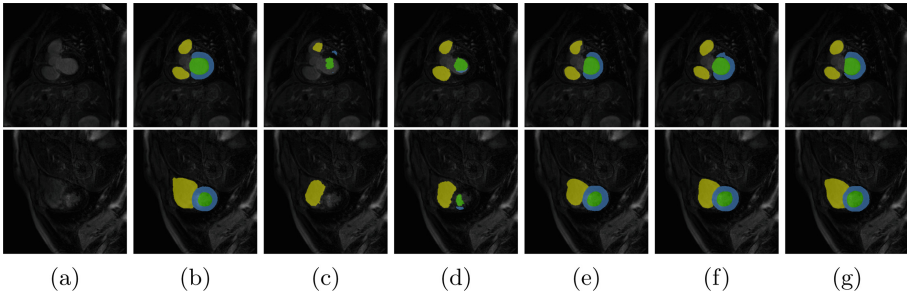

Fig. 5. Visual comparison for the LV, RV, and Myo segmentation results from ablation setting. (a) Original image from source domain. (b) Annotation. (c) S2T. (d) S2T+HM. (e) S2T+HM+MDA. (f) S2T+HM+MDA+FDA. (g) S2T+HM+MDA+FDA+GFRM.

Quantitative and Qualitative Analysis. In order to verify the effectiveness of the proposed method, we adopt Dice coefficient (DSC), Jaccard coefficient (Jac) for further evaluation. We first trained segmentation network on the source data and then test on the target data (S2T). The results in Table 1 shows that the mean Dice in S2T is too slow. As we can see, our method can promote about 36.09% in DSC and 38.38% in Jac than S2T, which indicates that our method can alleviate dataset shift across different domains.

In addition, we examine the effect of the histogram match operation (HM), mask-level adversarial learning (MDA), feature-level adversarial learning (FDA) and GFRM on the performance in the target domain. The result of the ablation study in Table 1 shows that our proposed modules can achieve a better performance than S2T. Figure 5 demonstrates that each proposed module can contribute to alleviate the domain misalignment.

Table 1. Quantitative evaluation of our proposed methods

Method	LV		RV		Myo		Mean	
	DSC [%]	Jac [%]	DSC [%]	Jac [%]	DSC [%]	Jac [%]	DSC [%]	Jac [%]
S2T	50.01	37.41	66.72	51.03	31.69	21.88	49.47	36.78
S2T+HM	59.80	47.66	76.02	62.98	38.13	26.73	57.98	45.79
S2T+HM+MDA	85.67	75.48	86.19	75.89	75.35	60.70	82.40	70.69
S2T+HM+MDA+FDA	88.43	79.68	85.70	75.14	78.43	64.57	84.19	73.13
S2T+HM+MDA+FDA+GFRM	**89.33**	**81.15**	**87.17**	**77.29**	**80.17**	**67.04**	**85.56**	**75.16**

4   Conclusion

In this paper, we proposed an unsupervised domain alignment method for left ventricle (LV), right ventricle (RV) and myocardium (Myo) segmentation from different cardiac MR sequences. We first introduced a segmentation network with hybrid segmentation loss to generate accurate prediction. We alleviate the dataset shift across different domains by leveraging the adversarial learning in both feature and output spaces. The proposed GFRM can enforce the fine-grained semantic-level feature alignment that matching features from different networks but with the same class label. Experiments show that the proposed method can achieve competitive results.

References

1. Berman, M., Rannen Triki, A., Blaschko, M.B.: The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4413–4421 (2018)
2. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: towards cross-modality domain adaptation for medical image segmentation. arXiv preprint arXiv:1901.08211 (2019)

3. Dong, N., Kampffmeyer, M., Liang, X., Wang, Z., Dai, W., Xing, E.: Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 544–552. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_61

4. Dou, Q., et al.: PnP-AdaNet: plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation. arXiv preprint arXiv:1812.07907 (2018)

5. Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.A.: Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 691–697. AAAI Press (2018)

6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)

7. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

8. Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 201–209. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_23

9. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 421–429. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_48

10. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7472–7481 (2018)

11. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)

12. Yang, X., et al.: Combating uncertainty with novel losses for automatic left atrium segmentation. In: Pop, M., et al. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 246–254. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-12029-0_27

13. Yue, Q., Luo, X., Ye, Q., Xu, L., Zhuang, X.: Cardiac segmentation from LGE MRI using deep neural network incorporating shape and spatial priors. arXiv preprint arXiv:1906.07347 (2019)

14. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)

15. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

16. Zhuang, X.: Multivariate mixture model for cardiac segmentation from multi-sequence MRI. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 581–588. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_67

17. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans. Pattern Anal. Mach. Intell. (2018)