Neural Style Transfer Improves 3D Cardiovascular MR Image Segmentation on Inconsistent Data

Chunwei Ma, Zhanghexuan Ji, and Mingchen Gao[(✉)]

Department of Computer Science and Engineering, University at Buffalo,
The State University of New York, Buffalo, USA
{chunweim,zhanghex,mgao8}@buffalo.edu

Abstract. Three-dimensional medical image segmentation is one of the most important problems in medical image analysis and plays a key role in downstream diagnosis and treatment. Recent years, deep neural networks have made groundbreaking success in medical image segmentation problem. However, due to the high variance in instrumental parameters, experimental protocols, and subject appearances, the generalization of deep learning models is often hindered by the inconsistency in medical images generated by different machines and hospitals. In this work, we present StyleSegor, an efficient and easy-to-use strategy to alleviate this inconsistency issue. Specifically, neural style transfer algorithm is applied to unlabeled data in order to minimize the differences in image properties including brightness, contrast, texture, etc. between the labeled and unlabeled data. We also apply probabilistic adjustment on the network output and integrate multiple predictions through ensemble learning. On a publicly available whole heart segmentation benchmarking dataset from MICCAI HVSMR 2016 challenge, we have demonstrated an elevated dice accuracy surpassing current state-of-the-art method and notably, an improvement of the total score by 29.91%. StyleSegor is thus corroborated to be an accurate tool for 3D whole heart segmentation especially on highly inconsistent data, and is available at https://github.com/horsepurve/StyleSegor.

Keywords: Whole heart segmentation · Atrous convolutional network · Neural style transfer

1 Introduction

The segmentation of 3D cardiac magnetic resonance (MR) images is the prerequisite for downstream diagnosis and treatment including heart disease identification and surgical planning. And there has been intensive research on the automatic algorithms for this segmentation problem, for purpose of alleviating the arduous manual labeling. Deep neural networks have made tremendous achievement on this task and many different architectures have been proposed,

such as 3D U-Net [3], VoxResNet [1], 3D-DSN [5], DenseVosNet [11], VFN [10], and their ensemble meta-learner [13], which improved the segmentation performance to the dice score of myocardium at ∼0.833 and that of blood pool at ∼0.939.

However, the current accuracy of 3D cardiovascular MR image segmentation is still not well satisfactory for wider practice due to several issues. First, the morphological variation within the HVSMR data, originated from a variety of congenital heart defects, leads to difficulty in segmentation. Dong et al. [4] proposed an unsupervised domain adaptation network to enforce prediction masks to be similar across domains. However, the shapes of myocardium and blood pool are much more complex than lungs in 2D X-rays images. More important, we have observed non-negligible inter-subject variation within the training and testing images, including brightness, resolution, texture, and signal to noise ratio. In HVSMR data, the training samples are generally of high quality while the quality of the testing samples is relatively low. In a training image (Fig. 1A), the intensity distribution (gray line in Fig. 1D) exhibits three distinguishable peaks whereas the testing image (Fig. 1B and E) shows a substantial overlap of myocardium signal and background signal. This dataset shift phenomena [7] significantly hampered the generalization of deep neural network models. Zhao et al. [12] proposed using learned transforms to generate samples used in data augmentation aiming at one-shot segmentation. In our preliminary experiments, we found that augmenting the training set with images generated from the low-quality domain contributed little to the overall performance.

To address these challenges, we propose StyleSegor, a novel pipeline for 3D MR image segmentation of cardiac and vascular structures. StyleSegor has three main advantages. First, we adopted atrous convolution network with atrous spatial pyramid pooling module as an efficient way to retain as many details of feature maps as possible and achieve better segmentation on subtle structures. Second, we leverage neural style transfer to minimize the inter-subject variation. Every slice sample in the testing data is directly transferred to the same style of a target from the training set. Third, in order to fully utilize both the original and the transformed image data, an ensemble learning scheme is developed through voting of multiple predictions. On the HVSMR 2016 challenge dataset, StyleSegor has demonstrated superior performance compared with other methods, and notably, an improvement of the total score by 29.91%, showing the effectiveness of our strategy.

2   Methods

The complete pipeline of StyleSegor is shown in Fig. 2. The standard ResNet-101 and VGG-16 networks serve as the backbones for segmentation and style transfer, respectively. The network is pre-trained on the combination of images from three orthogonal planes and then fine-tuned separately on images from each plane. Each testing slice goes through style transfer network to generate its transferred counterpart, which is in turn segmented using the fine-tuned segmentation model.
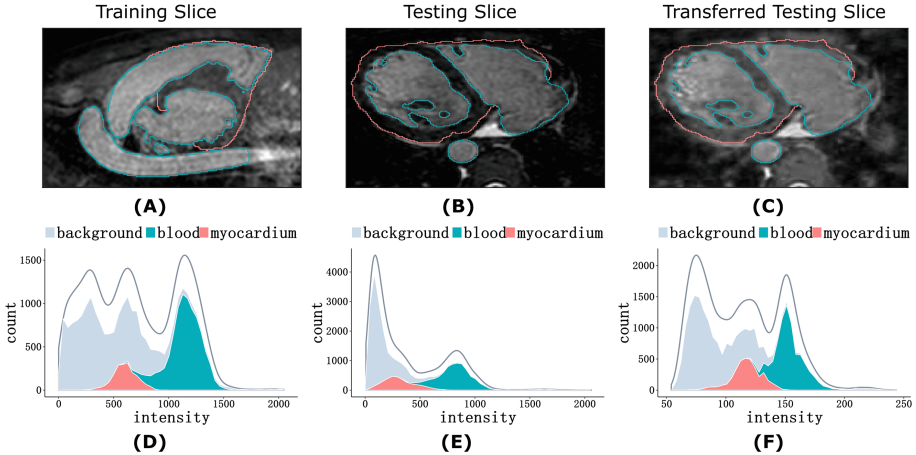
Fig. 1. Two representative slices from training (A) and testing set (B) and the transferred testing slice (C) are shown. Their intensity distributions are presented in D, E, and F, in which the intensity distributions of background, myocardium and blood pool are illustrated by gray, red, and green areas, respectively, and the gray lines show overall intensity distributions. (Color figure online)

2.1 Atrous Convolutional Neural Network for Dense Image Segmentation

For our baseline model, we modified DeepLabv3 [2], the state-of-the-art 2D semantic segmentation network, with ResNet-101 backbone. In order to fully utilize multi-scale information of feature maps elicited from ResNet, a pyramid of atrous convolution layer with various atrous rates $r = (6, 12, 18)$ is constructed on top of the last block of ResNet. Besides the three atrous convolution layers, the features from a 1×1 convolution and a bilinearly upsampled duplication of the input feature map are also considered. These 5 layers compose the atrous spatial pyramid pooling (ASPP) module whose feature maps are all concatenated. Finally, three 1×1 convolution layers are used to generate the final logits. Two batch normalization layers and two dropout layers (dropout rates being 0.5 and 0.1) are inserted between the final 3 convolution layers.

2.2 Neural Style Transfer on Inconsistent Data

Due to the large inconsistency between the training and testing data (Fig. 3 A), we apply a neural style transfer algorithm [6] on all 10 testing samples. Specifically, two types of loss, content loss and style loss are optimized, to change the style of a testing slice x to be similar to that of a target training slice y, while simultaneously impose constraint on the generated image $\hat{y}$ to maintain its content. Formally, given a feature extraction network ϕ that has J layers generating feature maps, the content loss is written as
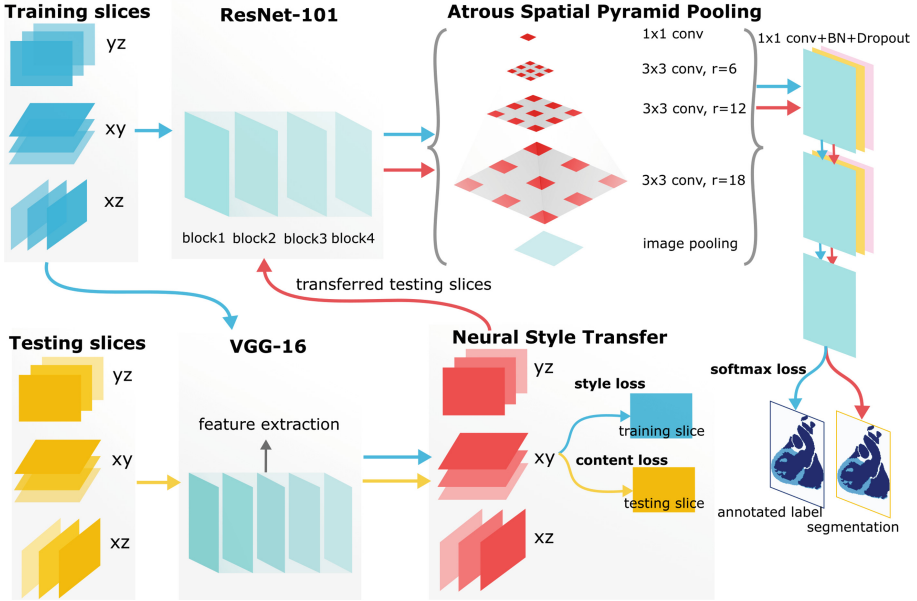
Fig. 2. Schematic illustration of StyleSegor workflow. A modified DeepLabv3 model with ResNet-101 backbone acts as our segmentation network and three models are trained in parallel for three planes $xy, yz, and\ zx$ (blue arrows). Meanwhile, testing slices are transferred to target styles guided by content loss and style loss, and then are fed into the segmentation network (red arrows). (Color figure online)

$$\ell^{\phi}_{content}(\hat{\boldsymbol{y}}, \boldsymbol{x}) = \sum_{j=1}^{J} \frac{1}{C_j H_j W_j} \parallel \phi_j(\hat{\boldsymbol{y}}) - \phi_j(\boldsymbol{x}) \parallel_2^2, \tag{1}$$

where C_j, H_j, W_j are the dimensions of the feature maps in the jth layer.

On the other hand, in order to measure the discrepancy between the generated slice $\hat{\boldsymbol{y}}$ and target slice $\boldsymbol{y}$, Gram matrix, originally designed to capture texture information, is to be computed. The jth Gram matrix for $\boldsymbol{y}$ is

$$G^{\phi}_j(\boldsymbol{y})_{i,k} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(\boldsymbol{y})_{h,w,i} \phi_j(\boldsymbol{y})_{h,w,k}, \tag{2}$$

that is, the i, k position at the jth Gram matrix measures the correlation (inner product) of the ith and the kth feature maps in the jth layer. Subsequently, the style loss is

$$\ell^{\phi}_{style}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \sum_{j=1}^{J} \parallel G^{\phi}_j(\hat{\boldsymbol{y}}) - G^{\phi}_j(\boldsymbol{y}) \parallel_2^2 . \tag{3}$$

The total loss is a weighted combination of content loss ans style loss

$$\ell^\phi_{total}(\hat{\boldsymbol{y}}, \boldsymbol{x}, \boldsymbol{y}) = \alpha\ell^\phi_{content}(\hat{\boldsymbol{y}}, \boldsymbol{x}) + \beta\ell^\phi_{style}(\hat{\boldsymbol{y}}, \boldsymbol{y}), \tag{4}$$

where α and β are user-specified hyper parameters to adjust the relative weights of the two losses. During the style transfer process, stochastic gradient descent (SGD) optimization is directly applied on the generated image $\hat{\boldsymbol{y}}$ starting from the content image $\boldsymbol{x}$.

Until now, a remaining question is, for a given testing slice, how to find the optimal training slice as its target slice. We address this problem in several steps. First, the pairwise similarities of all training and testing samples are measured through the 1st Wasserstein metric

$$W(r, g) = \inf_{\gamma \in \Gamma(r,g)} \mathbb{E}_{(x,y) \sim \gamma} \| x - y \|, \tag{5}$$

where $\Gamma(r, g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are r, g, which measures the work needed to transport from x to y with optimal transport plan. Considering the high difference in intensity ranges across samples, Wasserstein distance is a suitable indicator for sample similarity. Based on these similarities, all samples are clustered using hierarchical clustering algorithm [8], and the training samples reside in one cluster serve as the style library (the first cluster in Fig. 3A). Using our baseline network, the percentages of the three labels within each testing slice are used to measure the distance between two slices, and the slice in the style library with the smallest Euclidean distance to the testing slice is chosen as the target style.

Because in StyleSegor, a full training process is required for every content-style pair, we use VGG-16, a lightweight network, as the feature extraction network ϕ and the feature maps after the $2^{nd}, 4^{th}, 7^{th}, 10^{th}$ convolution layers are used to compute the Gram matrices (see Fig. 2).

2.3   Probabilistic Adjustment and Ensemble Learning

Based on the observation that the signals of myocardium and blood pool tend to be overwhelmed by the background signal (Fig. 3E and H), we perform a probabilistic adjustment step and adjust the score for one label at position i by conditioning on the scores of other labels

$$c(p_k) = \arg\max_{k \in (1,2,3)} p_k \prod_{j \neq k} \left(1 - \frac{e^{p_j}}{\sum_{q \in (1,2,3)} e^{p_q}} \right), \tag{6}$$

where $p_k, k \in (1, 2, 3)$ is the logits output from the network for three labels. For example, the score of myocardium is multiplied by the probabilities of both non-blood pool and non-background (Fig. 3F and I).

In machine learning practice, model ensemble is oft-used to take advantage of multiple models and predictions. Here we adopt a voting scheme to integrate segmentations obtained from both original and transformed images. The final label

at position i is the voting of $c(p_k^{(xy)}), c(p_k^{(yz)}), c(p_k^{(zx)})$ and $c(\sum_{xy,yz,zx} p_k)$ derived from the original images and $c(p_k^{'(xy)}), c(p_k^{'(yz)}), c(p_k^{'(zx)})$ and $c(\sum_{xy,yz,zx} p_k')$ derived from the transferred images (Fig. 3J).

3   Experimental Results

Dataset and Training Process. We evaluate the performance of StyleSegor on HVSMR, the dataset for MICCAI 2016 Challenge on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease. Imaging was done in an axial view on a 1.5T scanner. Ten 3D MR scans, as well as the manually labeled annotations for myocardium and great vessel, are provided for training, but the labels for 10 testing scans are not made publicly available for fair comparison. After carefully investigating the properties of testing images, we observed that the signal of myocardium in testing samples is especially lower than in training samples (see Fig. 1A, D and B, E). The clustering result of training and testing samples based on Wasserstein metric is shown in Fig. 3A, where training samples are marked from 0 to 9 and testing samples from 11 to 19. Clearly, all testing sample reside in the same cluster, which is significantly different from another cluster of training samples. In our style transfer network, the weights of style and content loss α and β are set at 10^6 and 1, respectively, and the optimization terminates after 50 epochs, which typically takes 3s for one content-style pair on a GTX 1080 Ti card. The VGG-16 network is trained on ImageNet dataset.

To fully make use of the slices from three orthogonal planes, all slices are collected for training for 20 epochs with learning rate starting at 0.01. Then the slices derived from xy, yz, and zx planes are used to fine-tune the model separately with learning rate starting from 0.002 for another 20 epochs each. A poly learning rate policy is employed where the starting learning rate is reduced by multiplying $(1 - \frac{epoch}{max\_epoch})$. To accelerate the training process, the segmentation network is pre-trained on COCO dataset. During the training of our baseline model, a series of data augmentation strategy is applied. Each original image is randomly scaled with the rates ranged from 0.5 to 2.0 and a 480×480 patch is cropped then goes through random left-right flipping and random Gaussian blurring. Because the training images are randomly scaled during training, in testing process, each testing image is scaled with scaling rate $= (0.5, 0.75, 1, 1.25, 1.5, 2.0)$ and the accumulated score map is used to produce final segmentation.

A representative testing slice and the transferred slice of it are shown in Fig. 1B and C, while the intensity distribution of background, blood pool, and myocardium are illustrated in Fig. 1E and F. Interestingly, after style transfer, not only the brightness, contrast, texture of the image but also the distribution of the three labels are transformed to be very similar to the training image, and the myocardium signal is smartly elevated.

Quantitative Comparisons. The comparison of StyleSegor and our baseline network, along with other segmentation methods are shown in Table 1, and the
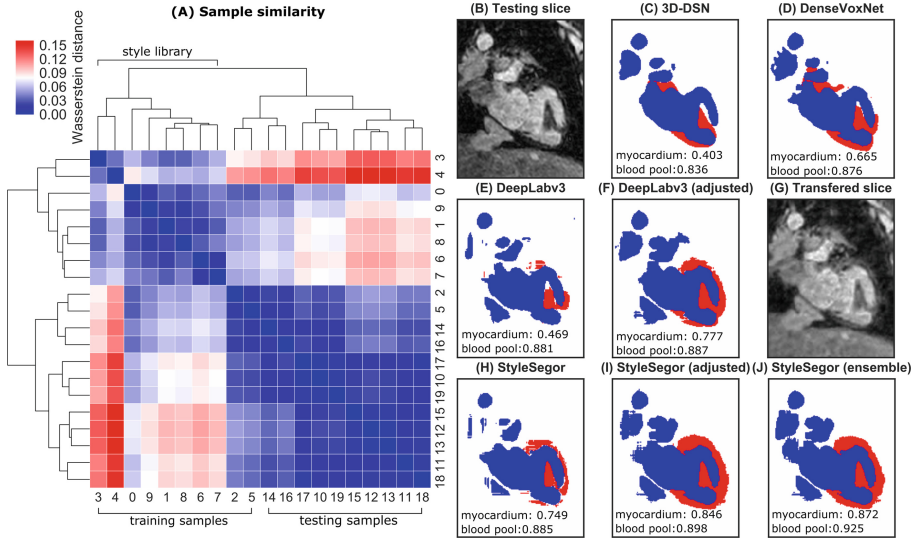
Fig. 3. Clustering result of the 20 samples (A) and segmentation results on a representative testing slice (B to J). The blue and red colors in B to J represent blood pool and myocardium, respectively. The dice scores of the two labels are also shown. The ground truth labels for testing data are not made publicly available. (Color figure online)

visualization of those segmentation results is provided in Fig. 3B to J. After probabilistic adjustment, although our baseline model only performs 2D convolution, it comes up with satisfactory segmentation with dice score of myocardium at 0.808 and that of blood pool at 0.919, and notably, by virtue of the large field of view, it produces the best Hausdorff distance at 3.105 mm compared with previous methods. After style transfer, the segmentation performance of myocardium is promoted to 0.825 and that of blood pool to 0.923, suggesting that with the promotion of myocardium signal, the myocardium structures are better recognized by the same model. However, we notice that after transfer, the Hausdorff distance of myocardium segmentation is enlarged to 4.633 mm, probably caused by false positive prediction of myocardium label brought by style transfer. And this false positive prediction is likely to be eliminated by the ensemble of multiple predictions. As shown in the last row of Table 1, the ensemble result is better than either StyleSegor or DeepLabv3, with dice score of myocardium at 0.839 and that of blood pool at 0.937. Notably, the Hausdorff distances are greatly minimized to 2.832 mm for myocardium and 4.023 mm for blood pool, and the overall score is boosted to 0.304, a 29.91% improvement compared with previously best result, demonstrating StyleSegor's strength to locate the region of interest.

Table 1. Comparison of different methods on HVSMR 2016 dataset. The weights of relative contributions of Dice, Average distance boundary (ADB), and Hausdorff distance to the Overall score are 0.5, −0.25, and −0.03, respectively.

Method	Myocardium			Blood pool			Overall
	Dice	ADB [mm]	Hausdorff [mm]	Dice	ADB [mm]	Hausdorff [mm]	score
3D U-Net [3]	0.694±0.076	1.461±0.397	10.221±4.339	0.926±0.016	0.940±0.192	8.628±3.390	-0.419
3D DSN [5]	0.739±0.072	1.035±0.240	5.248±1.332	0.928±0.014	1.017±0.181	7.704±2.892	-0.162
VoxResNet [1]	0.774±0.067	1.026±0.400	6.572±0.013	0.929±0.013	0.981±0.186	9.966±3.021	-0.202
DenseVoxNet [11]	0.821±0.041	0.964±0.292	7.294±3.340	0.931±0.011	0.938±0.224	9.533±4.194	-0.161
Wolterink et. al [9]	0.802±0.060	0.957±0.302	6.126±3.565	0.926±0.018	0.885±0.223	7.069±2.857	-0.036
VFN [10]	0.773±0.098	0.877±0.318	4.626±2.319	0.935±0.009	0.770±0.098	5.420±2.152	0.108
Zheng et. al [13]	0.833±0.054	**0.681±0.178**	3.285±1.370	**0.939±0.008**	0.733±0.143	5.670±2.808	0.234
DeepLabv3 (baseline)	0.648±0.156	1.234±0.531	5.960±3.921	0.920±0.025	0.983±0.309	7.343±2.999	-0.214
StyleSegor (baseline)	0.744±0.085	1.061±0.322	5.610±2.641	0.923±0.022	1.000±0.285	5.778±2.999	-0.061
DeepLabv3 (adjusted)	0.808±0.057	0.820±0.230	3.105±1.033	0.916±0.018	1.038±0.227	7.887±2.787	0.031
StyleSegor (adjusted)	0.825±0.031	0.934±0.237	4.633±2.241	0.923±0.014	1.073±0.191	7.435±2.649	-0.030
StyleSegor (ensemble)	**0.839±0.037**	0.689±0.140	**2.832±0.660**	0.937±0.014	**0.731±0.182**	**4.023±1.299**	**0.304**

4   Conclusion

In this paper, we present StyleSegor, a novel pipeline for 3D cardiac MR image segmentation. The neural style transfer algorithm automatically transfers the testing images towards the domain of training images, making them easier to be processed by the same model. Our StyleSegor pipeline is also easy to be used in other tasks such as disease detection and classification when data inconsistency is an inevitable issue, e.g., tasks involving datasets collected from different hospitals or institutions.

References

1. Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A.: VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage **170**, 446–455 (2018)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with Atrous separable convolution for semantic image segmentation. In: ECCV, pp. 801–818 (2018)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
4. Dong, N., Kampffmeyer, M., Liang, X., Wang, Z., Dai, W., Xing, E.: Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In: Frangi, A.F., Schnabel, J.A., 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
5. Dou, Q., et al.: 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. **41**, 40–54 (2017)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR, pp. 2414–2423 (2016)

7. Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B.: Covariate Shift and Local Learning by Distribution Matching, pp. 131–160. MIT Press, Cambridge (2009)
8. Müllner, D.: Modern hierarchical, agglomerative clustering algorithms. arXiv preprint arXiv:1109.2378 (2011)
9. Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease. In: Zuluaga, M.A., Bhatia, K., Kainz, B., Moghari, M.H., Pace, D.F. (eds.) RAMBO/HVSMR -2016. LNCS, vol. 10129, pp. 95–102. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52280-7_9
10. Xia, Y., Xie, L., Liu, F., Zhu, Z., Fishman, E.K., Yuille, A.L.: Bridging the gap between 2D and 3D organ segmentation with volumetric fusion net. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 445–453. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_51
11. Yu, L., et al.: Automatic 3D cardiovascular MR segmentation with densely-connected volumetric convnets. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 287–295. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_33
12. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: CVPR, pp. 8543–8553 (2019)
13. Zheng, H., et al.: A new ensemble learning framework for 3D biomedical image segmentation. In: AAAI (2019)