Style-Invariant Cardiac Image Segmentation with Test-Time Augmentation

Xiaoqiong Huang[1,2], Zejian Chen[1,2], Xin Yang[1,2], Zhendong Liu[1,2], Yuxin Zou[1,2], Mingyuan Luo[1,2], Wufeng Xue[1,2], and Dong Ni[1,2](✉)

[1] School of Biomedical Engineering, Shenzhen University, Shenzhen, China
nidong@szu.edu.cn

[2] Medical UltraSound Image Computing (MUSIC) Laboratory, Shenzhen University, Shenzhen, China

Abstract. Deep models often suffer from severe performance drop due to the appearance shift in the real clinical setting. Most of the existing learning-based methods rely on images from multiple sites/vendors or even corresponding labels. However, collecting enough unknown data to robustly model segmentation cannot always hold since the complex appearance shift caused by imaging factors in daily application. In this paper, we propose a novel style-invariant method for cardiac image segmentation. Based on the zero-shot style transfer to remove appearance shift and test-time augmentation to explore diverse underlying anatomy, our proposed method is effective in combating the appearance shift. Our contribution is three-fold. First, inspired by the spirit of universal style transfer, we develop a zero-shot stylization for content images to generate stylized images that appearance similarity to the style images. Second, we build up a robust cardiac segmentation model based on the U-Net structure. Our framework mainly consists of two networks during testing: the ST network for removing appearance shift and the segmentation network. Third, we investigate test-time augmentation to explore transformed versions of the stylized image for prediction and the results are merged. Notably, our proposed framework is fully test-time adaptation. Experiment results demonstrate that our methods are promising and generic for generalizing deep segmentation models.

Keywords: Style transfer · Cardiac image segmentation · Test-time augmentation

1 Introduction

Delineation of the left ventricular cavity (LV), myocardium (MYO), and right ventricle (RV) from cardiac magnetic resonance (CMR) images (multi-slice 2D cine MRI) is a common clinical task to establish the diagnosis. It is of great

X. Huang and Z. Chen – Equally contribute to this work.

E. Puyol Anton et al. (Eds.): STACOM 2020, LNCS 12592, pp. 305–315, 2021.
https://doi.org/10.1007/978-3-030-68107-4_31

interest to develop an accurate automated segmentation method since manual segmentation is tedious and likely to suffer from inter-observer variability. Deep learning cardiac segmentation models have achieved remarkable success based on a large amount of labeled data. However, as shown in Fig. 1, learning-based models often subject to severe performance drop due to testing data that has different distributions from the training data. This is a highly desirable but challenging task that makes deep models robust against the complex appearance shift of testing images [1,13] caused by different sites, scanner vendors, imaging protocols, etc.

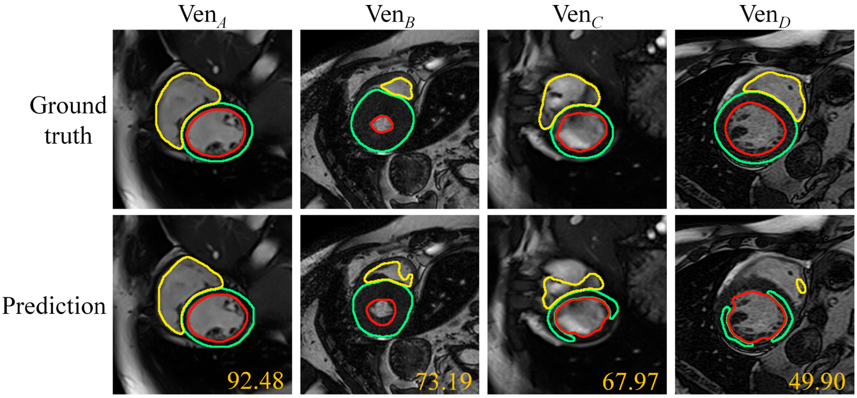

Fig. 1. Illustration of segmentation degradation on cardiac images from four vendors (Ven_A, ..., Ven_D). Green, red and yellow curves represent LV, MYO and RV, respectively. Orange digits denote the average Dice index over three structures. The model trained on the images of Ven_A performs a notable drop on images of other verdors. (Color figure online)

To mitigate the performance degradation, one straightforward choice is data augmentation [10,11]. It can help suppress overfitting but cannot guarantee the generalization ability of deep models. Recently, Domain Adaptation (DA) [4] and Domain Generalization (DG) [5] have been common methods for coping with the appearance shift. As main branches of DA/DG, aligning appearance level or feature level among different domains via adversarial learning were explored. Although DA/DG is attractive, it depends heavily on sufficient data from the target domain or requires enough multiple labeled source data. It is also confined by its domain mapping and may not extend to images from unknown domains. By revisiting the basic definition of appearance shift, style transfer [6] (ST) inspires a new and intuitive way for the problem. ST removes appearance shift by rendering the appearance of the content image as the style image [3,8,9]. Compared to DA, ST is independent on the target domain, retraining-free and suitable for images with unknown appearance shifts. In [9], Ma *et al.* made the early attempt to exploit an online ST to reduce the appearance variation for better cardiac MR segmentation. But such optimization-based ST has high latency and restrains real-time applications. Liu *et al.* [8] proposed an Adaptive Instance Normalization (AdaIN) [7] based ST module for vendor adaption to achieve real-time arbitrary style transfer. However, it directly utilized the pre-trained VGG-16 as the ST backbone, which may be unadaptable for the medical image to retain a more realistic semantic content structure.

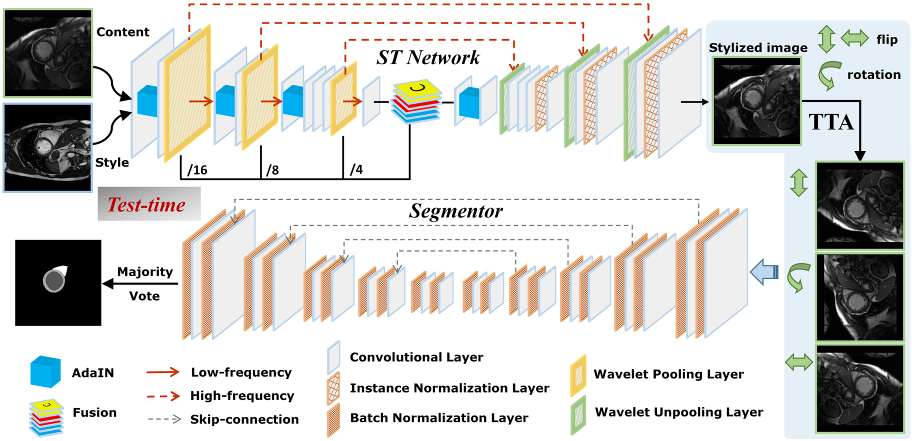

Fig. 2. Schematic view of our proposed framework.

In this paper, based on Wavelet Corrected Transfer network (WaveCT) [14] and WaveCT-AIN [8], we propose an improved ST network to generate style-invariant images for removing appearance shift and test-time augmentation to enhance the segmentation results. Our contribution is three-fold. First, inspired by the spirit of universal style transfer, we develop a zero-shot stylization for testing (content) images to generate stylized images that appearance similarity to the source (style) images. Second, we build up a robust model based on the U-Net structure for cardiac segmentation. Our framework is a two-stage system during testing: we utilize the ST network to generate the stylized image, then feed it into the segmentation model. Third, we investigate test-time augmentation to explore transformed versions of the stylized image for inference, followed by inverse transformation and predictions mergence to get the final segmentation result. In particular, we make two experiments to verify our proposed framework. 1) segmentation model trained on the original dataset and 2) segmentation model trained on the style-unified dataset generated by our zero-shot ST network from the original dataset.

2 Methodology

Figure 2 is the schematic view of our proposed method. The universal 2D U-Net and VGG-16 networks serve as the backbones for segmentation and ST, respectively. We first train the segmentation model on the source data, then develop a ST network to generate stylized images that suitable for the segmentation model. Specifically, the proposed framework is a two-stage system for segmenting images with appearance shifts. In the first stage, the testing image is transferred into a stylized image refer to the source image appearance. In the second stage, the segmentation model trained on source data is performed on the stylized image and get the segmentation result. Moreover, we explore transformed versions of the stylized testing image for prediction by using test-time augmentation and then perform a majority vote to obtain the final segmentation result.

2.1 Cardiac Segmentation Network Design

In this work, we modified the U-Net [12] as our baseline model, the state-of-the-art 2D semantic segmentation network in medical image analysis. Specifically, we use upsampling instead of deconvolution to avoid the grid effect. The output stride of the network is cut to 16 to reduce overfitting. The Batch Normalization layers are inserted after each convolution layer. The segmentation network aims at predicting four-class pixel-wise probabilistic maps for the three cardiac structures (i.e., LV, MYO, RV) and the background. To train the network, we use a composite segmentation loss function L_{seg} which consists of two loss terms:

$$\begin{gathered} L_{ce} = -\sum_{c} y^c log(p^c), \quad L_{Dice} = \sum_{c} 1 - \frac{2|X^c \cap Y^c|}{|X^c| + |Y^c|} \\ L_{seg} = L_{ce} + \lambda L_{Dice} \end{gathered} \tag{1}$$

The first term L_{ce} is a categorical cross entropy loss, where p^c denotes the corresponding predicted probability map of different classes. The second term is a Dice loss to measure the similarity between probability map X^c and ground truth Y^c. We set $\lambda = 0.5$ to balance the contribution of the two losses.

2.2 Zero-Shot Style Transfer

ST enables us to transfer the style of an image called style image to that of an image called the content image, rendering the low-level visual style while preserving its high-level semantic content structure. Inspired by the spirit of ST can remove appearance shift to approach generalize image analysis, we develop a ST network to generate style-invariant images for generalizing segmentation model. Different from the optimization-based or feed-forward approximate stylization, we utilize zero-shot ST to achieve real-time arbitrary stylization without training on any pre-defined styles.

To meet the requirement of universal and stable transfer between any content-style pairs, we adopt the WaveCT network recently used in WCT2 [14] and make improvements to preserve image structure details and render the style features. Different from previous online ST methods [9] used to remove appearance shift which may distort image details, the WaveCT network replaces vanilla max-pooling/unpooling with the Haar wavelet pooling/unpooling layers that maintain the content structure to the great extent. In particular, WaveCT splits the features into low-frequency and high-frequency components via Haar wavelet pooling, then low-frequency information passes the main network and high-frequency information skips to connect between encoder and decoder.

The ST network proposed in this paper is a significant extension of our prior conference paper proposed WaveCT-AIN [8], regarding the following highlighted points. As the ST network depicted in Fig. 2, first, we design a multi-scale feature fusion layer after the encoder, in return mitigate the variation of background area without information in the image. Second, we keep the ST module in the encoder and add an extra AdaIN after the feature fusion. Besides, we enhance the style-invariance by introducing the Instance Normalization (IN) layer into the decoder. In this respect, we focus on rendering the style texture representations in the low-level features and preserves its invariance in the high-level patterns. Third, we simplify the case-specific style image selection strategy more concisely and effectively, which is directly considered selecting the reference style image that as close as possible to the mean and standard deviation of the testing image. Especially, the ST network utilizes the pre-trained VGG-16 as backbone while feature fusion layer and IN layers are embedded, thus it needs to be fine-tuned with image reconstruction task.

2.3 Test-Time Augmentation

Data augmentation significantly improves robustness to appearance shift and can be used as a simple strategy for generalizing model performance. Data augmentation at training time has been commonly used to increase the amount of data for improving performance [11]. Recent works also demonstrated the usefulness of data augmentation directly at test time, for achieving more robust predictions [10]. For the point of data acquisition, a testing image is only one of many possible observations of the underlying anatomy. Therefore, we explore multiple transformed versions of the testing image for robust segmentation. Test-time augmentation includes four procedures: augmentation, prediction, inverse-augmentation, and merging. We firstly consider different transformations on the testing image. For our case, we have already remove appearance shift through the ST network, thus we apply flip and rotation transformations for stylized cardiac images instead of complicated contrast or brightness change. In particular, we make three different transformations on the stylized testing image and inference each version of the testing image, thus four predictions are obtained by the inverse transformation. Then we perform a majority vote to obtain the final segmentation result, that is, once the pixel is predicted twice or more, it will be regarded as the target area.

3 Experimental Results

3.1 Dataset and Implementation Details

Notably, we make two experiments to verify our proposed framework. *Exp.1*) segmentation model trained on the original training dataset, denoted as *SegO* and *Exp.2*) segmentation model trained on the style-unified dataset generated by our zero-shot ST network from the original dataset, denoted as *SegST*.

Table 1. Quantitative comparison results of the *Exp.1*.

Metrics	SegO				STSegO				STSegO-TTA			
	Ven$_A$	**Ven**$_B$	**Ven**$_C$	**Ven**$_D$	**Ven**$_A$	**Ven**$_B$	**Ven**$_C$	**Ven**$_D$	**Ven**$_A$	**Ven**$_B$	**Ven**$_C$	**Ven**$_D$
Dice$_{AVG}$	80.72	86.82	81.20	62.23	82.92	**89.72**	85.27	64.87	**84.70**	89.57	**85.56**	**68.01**
Jac$_{AVG}$	68.93	77.98	69.53	49.81	71.82	**82.04**	75.01	53.49	**74.29**	81.98	**75.45**	**55.80**
HDB$_{AVG}$	19.52	14.24	18.30	44.88	17.54	9.93	13.27	44.53	**14.88**	**9.04**	**12.94**	**38.46**
ASSD$_{AVG}$	1.96	0.89	1.84	12.44	1.85	0.65	1.46	13.17	**1.51**	**0.65**	**1.44**	**7.38**
Dice$_{LV}$	89.14	93.07	85.12	72.97	88.71	**92.90**	**86.46**	**78.21**	**89.75**	92.45	86.39	74.91
Jac$_{LV}$	81.01	87.56	75.56	62.83	80.27	**87.39**	**77.34**	**68.09**	**81.95**	86.92	77.26	64.94
HDB$_{LV}$	13.58	9.55	13.38	33.68	13.28	6.72	12.05	**28.23**	**11.59**	**6.03**	**10.90**	31.42
ASSD$_{LV}$	1.53	0.61	1.91	10.86	1.54	**0.63**	**1.74**	**5.76**	**1.34**	0.68	1.76	8.28
Dice$_{MYO}$	72.71	76.63	73.60	51.76	81.84	**85.69**	83.83	60.76	**83.13**	85.68	**83.92**	**63.19**
Jac$_{MYO}$	57.47	62.64	58.72	36.92	69.49	75.19	72.52	47.01	**71.32**	**75.21**	**72.63**	**48.53**
HDB$_{MYO}$	20.27	21.56	19.15	34.00	16.54	11.36	**14.77**	**27.22**	**14.02**	**10.74**	14.85	30.40
ASSD$_{MYO}$	1.92	1.35	2.00	7.13	1.29	0.59	1.29	**2.67**	**1.09**	**0.55**	**1.28**	3.04
Dice$_{RV}$	80.31	90.77	84.88	61.96	78.21	90.58	85.51	55.66	**81.22**	**90.60**	**86.39**	**65.94**
Jac$_{RV}$	68.32	83.74	74.30	49.69	65.71	83.55	75.18	45.38	**69.60**	**83.79**	**76.47**	**53.94**
HDB$_{RV}$	24.72	11.61	22.38	66.95	22.79	11.71	**12.98**	78.14	**19.03**	**10.35**	13.07	**53.55**
ASSD$_{RV}$	2.43	0.70	1.61	19.34	2.72	0.74	1.35	31.08	**2.09**	**0.72**	**1.27**	**10.83**

Dataset. The framework was trained and evaluated on the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms 2020) dataset [2]. Two subsets of 75 CMR images from vendor A and vendor B (denoted as Ven$_A$ and Ven$_B$) with only ES & ED annotated are provided as training data, respectively. Additionally, 25 unannotated images are also given from Ven$_C$. However, our methods do not use it because we are concentrate on generalizing the model to other more unknown data, not just Ven$_C$. For evaluation, the results are evaluated on not only 50 new studies from each of Ven$_{A,B,C}$, but also 50 else studies from Ven$_D$.

Implementation Details. We obtain 3284 training slices along the anatomical plane from the ES & ED images of Ven$_A$ and Ven$_B$. All slices are resized to 256 × 256. For training the *SegO*, we apply elastic deformations to the available training slices (i.e., random expand, flip, rotation, mirror, contrast change, and brightness change). Whereas the training of *SegST* is not necessary to make contrast or brightness change because its training data already have a particular appearance distribution. We use 3000 cardiac slices to fine-tune the ST Network as the extra feature fusion layer and IN layers are embedded into the pre-trained VGG-16. For the segmentation network, it was trained for 60k iterations with a batch size of 24 and was optimized using the composite loss L_{seg} where Adam optimizer with a learning rate of 10^{-3} initially then decreased to 10^{-5}. We implement all experiments with PyTorch on two GeForce® RTX 2080 Ti GPU.

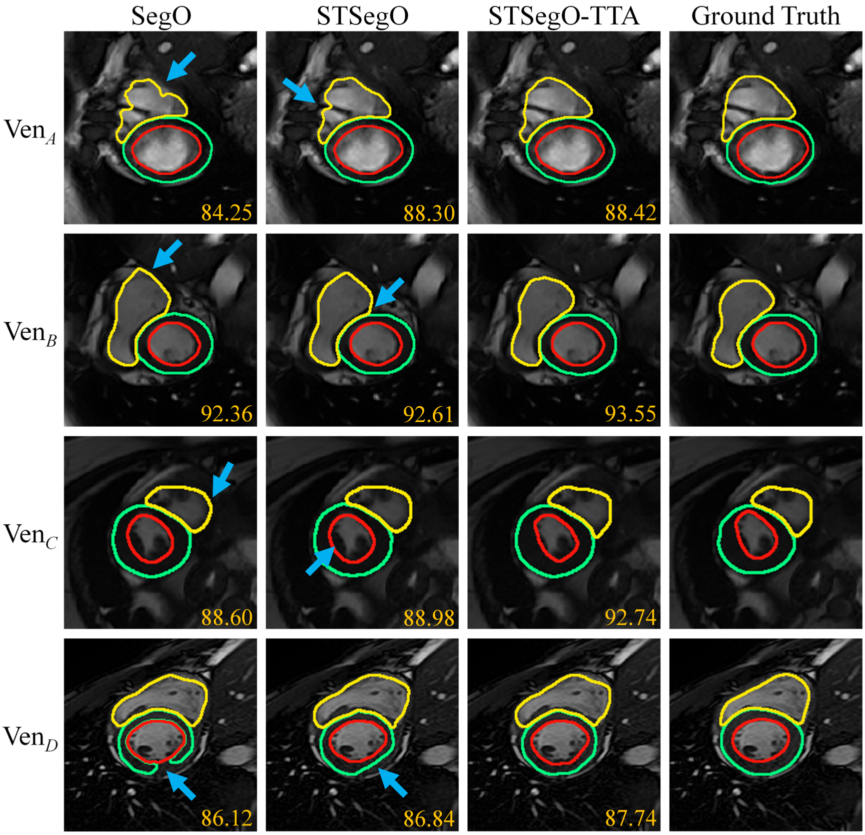

Fig. 3. Visualization of our better 3D segmentation results. From left to right are cases from Ven$_B$, Ven$_C$, Ven$_D$ and Ven$_D$, respectively. Green, red and yellow areas represent LV, MYO and RV, respectively. (Color figure online)

3.2 Quantitative and Qualitative Evaluation

Metrics. To evaluate the accuracy of segmentation performance, we adopt in total 4 indicators including the Dice similarity index (Dice, %), Jaccard

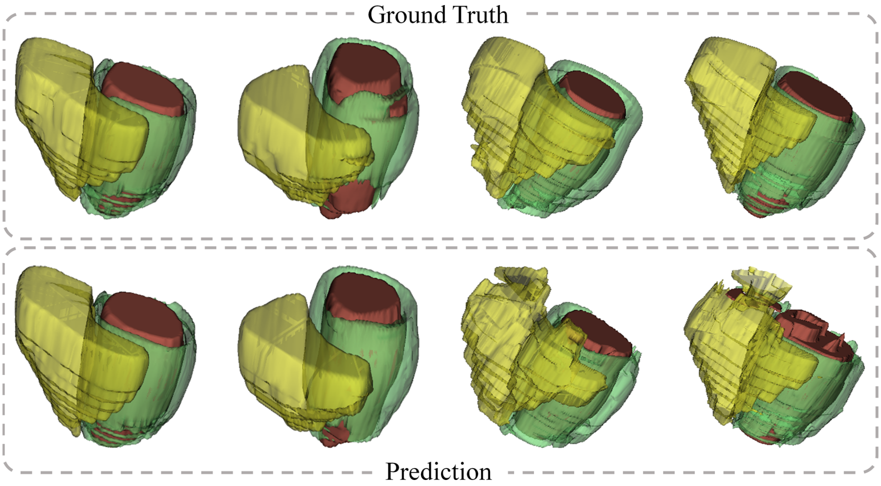

Fig. 4. Visualization of the 2D segmentation results of our proposed methods. Green, red and yellow curves represent the boundary of LV, MYO and RV, respectively. Orange digits denote the average Dice index. The performance is gradually improved from left to right methods, especially in the boundaries of RV. (Color figure online)

similarity index (Jac, %), Hausdorff Distance of Boundaries (HDB, pixel), and Average Symmetric Surface Distance (ASSD, pixel). For ease of comparison, we calculate the average (AVG) per indicator over the three structures (LV, MYO, RV).

Quantitative Results of the *Exp.1*. We first train the segmentation model *SegO* based on the available labeled images from Ven_A and Ven_B. Then we set up a style image library from the training data, which contains 221 slices from the top-20 images via computing the Dice index. Therefore, the testing (content) slice selects the reference style slice from the library through our simple style selection strategy. Subsequently, the content-style pairs feed into our zero-shot ST network to generate the stylized slice for segmentation. This two-stage system is denoted as *STSegO*. Likewise, *STSegO* with test-time augmentation is denoted as *STSegO-TTA*. Table 1 shows the quantitative results based on the 200 test cases correspond to four different vendors. We compare three versions of our proposed *SegO*, *STSegO* and *STSegO-TTA*, respectively. The numbers in bold indicate the best results of multiple vendors among different methods. Both *STSegO* and *STSegO-TTA* get consistent improvements over the pure *SegO*, in which the best results are achieved by the *STSegO-TTA*. It almost improves the Dice index by 5% and the Jaccard index by 6% on average for each vendor. The HDB and ASSD also improve about 5.5 pixels and 1.5 pixels, respectively. Obviously, the segmentation model can be well generalized to the images of Ven_C, but the performance on the Ven_D shows relatively poor. This may be due to the large difference in the data distribution and anatomical structure between images of Ven_D and the source data. Figure 4 visualizes the 2D segmentation results of our proposed methods on unseen cases from four vendors. *STSegO-TTA* produces more anatomically plausible results on the images. Figure 3 visualizes the 3D segmentation results of the *STSegO-TTA* on four unseen cases.

Quantitative Results of the *Exp.2*. Different from the *SegO* trained on the original dataset, *SegST* utilized the style-unified dataset for training, which is generated by our zero-shot ST network from the original dataset. Notably, we randomly select a slice serve as the style slice to generate stylized data, thus the style-unified training data has a particular appearance distribution. Consequently, the ST network only takes over this style slice to achieve stylization during testing. Likewise, *SegST* with test-time augmentation is denoted as *SegST-TTA*. As can be seen from Table 2, *STSeg-TTA* shows improvements over *STSeg*, the Dice index and Jaccard index are both raised about 2% on average for each vendor, and the HDB is improved about 1 pixel. However, the performance of Ven$_D$ shows worse compared with *Exp.1*, which may be caused by the target style slice is not suitable for the images of Ven$_D$. Thus it is crucial to choose a universal style slice to generate the style-unified dataset, which will be our further study.

Table 2. Quantitative comparison results of the *Exp.2*.

Metrics	SegST				SegST-TTA			
	Ven$_A$	**Ven**$_B$	**Ven**$_C$	**Ven**$_D$	**Ven**$_A$	**Ven**$_B$	**Ven**$_C$	**Ven**$_D$
Dice$_{AVG}$	84.19	89.63	**85.74**	53.84	**85.99**	**90.28**	85.23	**58.86**
Jac$_{AVG}$	73.31	82.06	**75.71**	42.64	**76.03**	**82.74**	74.96	**47.90**
HDB$_{AVG}$	16.89	15.46	20.22	**35.56**	**13.62**	**9.28**	**13.39**	47.86
ASSD$_{AVG}$	1.71	0.73	1.44	**14.55**	**1.34**	**0.58**	**1.37**	17.52
Dice$_{LV}$	87.82	92.54	87.04	63.71	**89.64**	**93.83**	**87.50**	**68.34**
Jac$_{LV}$	78.84	86.95	78.23	52.92	**81.85**	**88.68**	**78.87**	**58.54**
HDB$_{LV}$	16.78	14.45	17.34	**48.96**	**10.96**	**6.22**	**11.52**	49.84
ASSD$_{LV}$	1.90	0.76	1.74	**20.87**	**1.33**	**0.47**	**1.49**	21.27
Dice$_{MYO}$	81.04	**86.61**	**84.81**	53.49	**82.92**	86.54	84.70	**56.97**
Jac$_{MYO}$	68.37	**76.81**	**73.96**	40.61	**71.00**	76.44	73.74	**44.49**
HDB$_{MYO}$	16.42	20.60	22.42	**24.78**	**13.46**	**10.11**	**14.08**	30.12
ASSD$_{MYO}$	1.39	0.59	1.28	12.26	**1.07**	**0.49**	**1.01**	**7.68**
Dice$_{RV}$	83.70	89.73	**85.37**	44.31	**85.41**	**90.46**	83.49	**51.26**
Jac$_{RV}$	72.71	82.42	**74.94**	34.40	**75.25**	**83.09**	72.28	**40.66**
HDB$_{RV}$	17.45	**11.32**	20.90	**32.93**	**16.45**	11.49	**14.57**	63.62
ASSD$_{RV}$	1.85	0.84	**1.30**	**10.52**	**1.64**	**0.79**	1.62	23.62

4 Conclusion

In this paper, we proposed a zero-shot ST network to generate style-invariant images for removing appearance shift and test-time augmentation to enhance the segmentation results. By investigating the two experiments *Exp.1* and *Exp.2*, we showed that *SegO* and *STSeg* with their variants present promising performance in segmenting cardiac images across the multi-vendor and multi-cencre dataset.

Acknowledgement. This work was supported by the grant from National Key R &D Program of China (No. 2019YFC0118300), Shenzhen Peacock Plan (No. KQTD2016053112051497, KQJSCX20180328095606003) and National Natural Science Foundation of China (No. 61801296).

References

1. Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., Folk, J.C.: Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit. Med. **1**(1), 1–8 (2018)
2. Campello, V.M., et al.: Multi-centre, multi-vendor & multi-disease cardiac image segmentation. (in preparation)
3. Chen, C., et al.: Unsupervised multi-modal style transfer for cardiac MR segmentation. In: Pop, M., et al. (eds.) STACOM 2019. LNCS, vol. 12009, pp. 209–219. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39074-7_22
4. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE Trans. Med. Imaging (2020)
5. Dou, Q., de Castro, D.C., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. In: Advances in Neural Information Processing Systems, pp. 6450–6461 (2019)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
7. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
8. Liu, Z., et al.: Remove appearance shift for ultrasound image segmentation via fast and universal style transfer. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1824–1828. IEEE (2020)
9. Ma, C., Ji, Z., Gao, M.: Neural style transfer improves 3D cardiovascular MR image segmentation on inconsistent data. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 128–136. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_15
10. Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., Horvath, P.: Test-time augmentation for deep learning-based cell segmentation on microscopy images. Sci. Rep. **10**(1), 1–7 (2020)
11. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621 (2017)

12. 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.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Yang, X., et al.: Generalizing deep models for ultrasound image segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 497–505. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_57
14. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9036–9045 (2019)