A Style Transfer-Based Augmentation Framework for Improving Segmentation and Classification Performance Across Different Sources in Ultrasound Images

Bin Huang[1,7], Ziyue Xu[2], Shing-Chow Chan[3], Zhong Liu[1], Huiying Wen[1], Chao Hou[1], Qicai Huang[1], Meiqin Jiang[1], Changfeng Dong[4], Jie Zeng[5], Ruhai Zou[6], Bingsheng Huang[7(✉)], Xin Chen[1(✉)], and Shuo Li[8]

[1] School of Biomedical Engineering, Shenzhen University Medical School,
Shenzhen University, Shenzhen 518060, China
chenxin@szu.edu.cn

[2] Nvidia Corporation, Bethesda, MD 20814, USA

[3] Department of Electrical and Electronic Engineering,
The University of Hong Kong, Hong Kong, SAR, China

[4] Institute of Hepatology, Shenzhen Third People's Hospital,
Shenzhen 518000, China

[5] Department of Medical Ultrasonics, The Third Affiliated Hospital,
Sun Yat-sen University, Guangzhou 510000, China

[6] State Key Laboratory of Oncology in South China,
Collaborative Innovation Center of Cancer Medicine, Department of Ultrasound,
Sun Yat-sen University Cancer Center, Guangzhou 510000, China

[7] Medical AI Lab, School of Biomedical Engineering,
Shenzhen University Medical School, Shenzhen University, Shenzhen 518000, China
huangb@szu.edu.cn

[8] Department of Biomedical Engineering, Case Western Reserve University,
Cleveland, OH, USA

Abstract. Ultrasound imaging can vary in style/appearance due to differences in scanning equipment and other factors, resulting in degraded segmentation and classification performance of deep learning models for ultrasound image analysis. Previous studies have attempted to solve this problem by using style transfer and augmentation techniques, but these methods usually require a large amount of data from multiple sources and source-specific discriminators, which are not feasible for medical datasets with limited samples. Moreover, finding suitable augmentation methods for ultrasound data can be difficult. To address these challenges, we propose a novel style transfer-based augmentation framework that consists of three components: mixed style augmentation (MixStyleAug), feature augmentation (FeatAug), and mask-based style augmentation (MaskAug). MixStyleAug uses a style transfer network to transform the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_5.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 44–53, 2023.
https://doi.org/10.1007/978-3-031-43987-2_5

style of a training image into various reference styles, which enriches the information from different sources for the network. FeatAug augments the styles at the feature level to compensate for possible style variations, especially for small-size datasets with limited styles. MaskAug leverages segmentation masks to highlight the key regions in the images, which enhances the model's generalizability. We evaluate our framework on five ultrasound datasets collected from different scanners and centers. Our framework outperforms previous methods on both segmentation and classification tasks, especially on small-size datasets. Our results suggest that our framework can effectively improve the performance of deep learning models across different ultrasound sources with limited data.

Keywords: Ultrasound · Segmentation · Classification · Style transfer · Data augmentation

1 Introduction

Classification and segmentation are two common tasks that use deep learning techniques to solve clinical problems [1,2]. However, training deep learning models reliably usually requires a large amount of data samples. Models trained with limited data are susceptible to overfitting and possible variations due to small

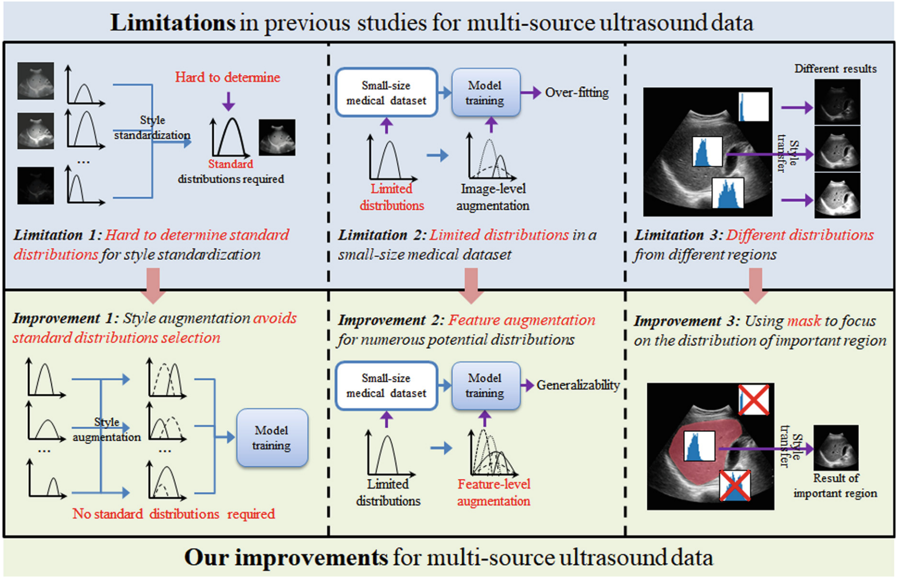
Limitations in previous studies for multi-source ultrasound data
Hard to determine
Style standardization
Standard distributions required
Limitation 1: *Hard to determine standard distributions for style standardization*
Small-size medical dataset
Model training
Over-fitting
Limited distributions
Image-level augmentation
Limitation 2: *Limited distributions in a small-size medical dataset*
Different results
Style transfer
Limitation 3: *Different distributions from different regions*
Improvement 1: *Style augmentation avoids standard distributions selection*
Style augmentation
Model training
No standard distributions required
Improvement 2: *Feature augmentation for numerous potential distributions*
Small-size medical dataset
Model training
Generalizability
Limited distributions
Feature-level augmentation
Improvement 3: *Using mask to focus on the distribution of important region*
Style transfer
Result of important region
Our improvements for multi-source ultrasound data

Fig. 1. Limitations of previous studies and our improvements for multi-source ultrasound data.

sample size, which can lead to poor performance across different sources. Different sources refer to the same modality collected from different scanners. In medical imaging, one of the main reasons for poor performance is the variation in the imaging process, such as the type of scanner, the settings, the protocol, *etc.* [3]. This can cause changes in the intensity distributions of the images [4,5]. While training deep learning models with a large number of high-quality data could potentially address this problem, this approach is often challenging due to limited resources and difficulties in collecting medical images, as well as the manual annotation required by experienced radiologists or experts with professional domain knowledge. Thus, limited labeled data is commonly used to model the classification and segmentation network.

To prevent overfitting and improve generalization, data augmentation [3,6–10] has been proposed to generate more similar but different samples for the training dataset. Very often, this can be done by applying various transformations to the training images to create new images that reflect natural variations within each class. However, the model's performance across different sources heavily depends on the augmentation strategies. Another popular technique is style transfer [11], which adapts the style of test images to match the selected reference images (standard distributions) [4,5,12,13]. However, these methods have a limitation that their performance depends on the quality of the reference images. Moreover, these methods tend to transfer the style of the whole images, which may introduce irrelevant distribution information for medical imaging applications, as shown in Fig. 1. This problem is more severe in ultrasound images due to the presence of acoustic shadow.

To address the above challenges, we propose a novel framework that combines the advantages of data augmentation and style transfer to enhance the model's segmentation and classification performance on ultrasound images from different sources. Our contributions (Fig. 1) are: 1) a mixed style augmentation strategy that integrates the information from different sources to improve the model's generalizability. 2) A feature-based augmentation that shifts the style at the feature level rather than the image level to better account for the potential variations. 3) a mask-based style augmentation strategy that avoids the influence of the irrelevant style information on ultrasound images during the style transfer.

2 Methods

Our proposed framework for ultrasonic image style augmentation consists of three stages, as illustrated in Fig. 2. Stage **A. Mixed style augmentation (MixStyleAug)** integrates the style information from different sources simultaneously. Stage **B. Feature augmentation** transfers the style at the feature level during the training of the multi-task network. Stage **C. Mask-based style augmentation** uses the style information of the region of interest (ROI) in the ultrasound image based on the segmentation results.

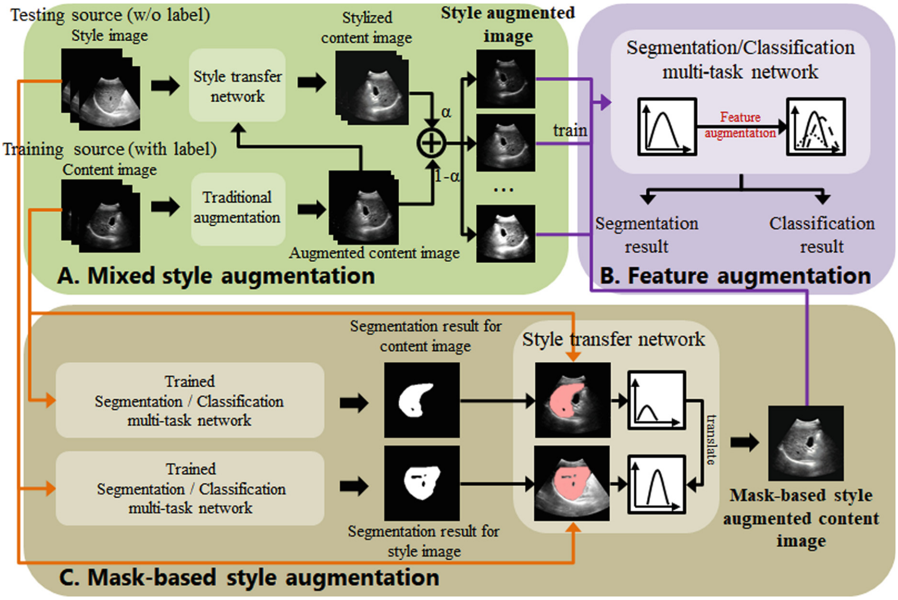
Testing source (w/o label)
Style image
Style transfer network
Stylized content image
Style augmented image
α
1-α
Training source (with label)
Content image
Traditional augmentation
Augmented content image
...
train
A. Mixed style augmentation
Segmentation/Classification multi-task network
Feature augmentation
Segmentation result
Classification result
B. Feature augmentation
Segmentation result for content image
Style transfer network
Trained Segmentation / Classification multi-task network
Trained Segmentation / Classification multi-task network
translate
Segmentation result for style image
Mask-based style augmented content image
C. Mask-based style augmentation

Fig. 2. Overview of our proposed style transfer-based augmentation framework. The whole framework consists of mixed style augmentation, feature augmentation, and mask-based style augmentation.

2.1 Mixed Style Augmentation (MixStyleAug)

To improve the performance of the multi-task network, we design MixStyleAug, combining traditional transformations and style transfer to incorporate image information from target sources during training (Fig. 2A). In this method, the content and the style images are sampled from training and target sources, respectively. Firstly, the traditional augmentation is applied to transform the content image, which can prevent overfitting. The traditional augmentation includes rotation, translation, scaling, and deformation transformations. Next, we translate the style of the augmented content image to that of the style image using the WCT^2 [14] style transfer network, generating a stylized content image. Finally, inspired by AugMix [15], we mix the stylized and augmented content images using random weights to create a style-augmented image that includes information from the training source. MixStyleAug allows the augmented training dataset to implicitly contain information from multiple sources, improving the model's performance across different sources. However, this method requires a large number of available images as reference styles for style augmentation, making it impractical for small-sized datasets.

2.2 Network Architecture and Feature Augmentation (FeatAug)

To address the limitation of MixStyleAug in small-size medical datasets, FeatAug is applied for augmenting image styles at the feature level during the network training (Fig. 2B). In this work, we design a simple multi-task network for simultaneous segmentation and classification, and FeatAug is applied to the feature maps for feature augmentation.

The architecture of our designed multi-task network (Fig. S1 in the *Supplementary Materials*) includes four encoders, four decoders, and a classification head. Each encoder includes two 3 × 3 convolutional layers with padding that are used to fuse the features. Each convolutional layer is followed by a rectified linear unit (ReLU) and a batch normalization (BN) [16]. Max-pooling layer is used to downsample the feature maps for dimension reduction. Through these encoders, the feature maps are generated and fed into the decoders and classification head to generate segmentation and classification results, respectively. Each decoder consists of three 3 × 3 convolutional layers with padding, three BN layers, three ReLUs, and a max-unpooling layer. In the classification head, the feature maps from the encoders are reduced to 128 channels by using a 3 × 3 convolutional layer with padding followed by ReLU and BN layer. Then, a global average pooling is used to downsample the feature maps. Finally, the features are fed into a fully connected layer followed by a sigmoid layer to output the classification result.

Previous studies reported that changing the mean and standard deviation of the feature maps could lead to different image styles [17,18]. Thus, we design a module to randomly alter these values to augment the styles at the feature level. To avoid over-augmentation at the feature level, this module is randomly applied with a 50% probability after the residual connection in each encoder. The module is defined as follows:

$$A' = \frac{A - \mu_A}{\sigma_A} \cdot \left(\sigma_A + \mathcal{N}(\mu, \sigma)\right) + \left(\mu_A + \mathcal{N}(\mu, \sigma)\right) \tag{1}$$

where A indicates the feature map, A' indicates the augmented feature map, μ_A indicates the mean of feature map A, σ_A indicates the standard deviation of feature map A, and $\mathcal{N}(\mu, \sigma)$ indicates a value randomly generated from a normal distribution with mean μ and standard deviation σ. In this study, the μ and σ of the normal distribution were empirically set to 0 and 0.1 according to preliminary experimental results, respectively.

2.3 Mask-Based Style Augmentation (MaskAug)

In general, the style transfer uses the style information of the entire image, but this approach may not be ideal when the regions outside of the ROIs contain conflicting style information as compared to the regions within the ROIs, as illustrated in Fig. 1. To mitigate the impact of irrelevant or even adverse style information, we propose a mask-based augmentation technique (MaskAug) that emphasize the ROIs in the ultrasound image during style transfer network training.

Figure 2C shows the pipeline of MaskAug and the steps are: 1) Content and style images are randomly chosen from training and target sources, respectively. 2) A trained multi-task network, which has been trained for several epochs and will be updated in the later epochs, is used to automatically generate ROIs of these images. 3) The content image, style image and their ROIs are input to the style transfer network. 4) During the style transfer, the intensity distribution of the ROI in the content image is changed to that of the style image. 5) Finally, mask-based style augmented images are produced and these images are then input to the multi-task network for further training.

2.4 Loss Function and Implementation Details

We utilized cross-entropy (CE) as the primary loss function for segmentation and classification during the training stage. Additionally, Dice loss [19] was computed as an auxiliary loss for segmentation. These loss functions are defined as:

$$\mathcal{L}_m = \mathcal{L}_{CE}^{Seg} + \mathcal{L}_{Dice}^{Seg} + \mathcal{L}_{CE}^{Cls} \tag{2}$$

where $\mathcal{L}_{CE}$ denotes CE loss, $\mathcal{L}_{Dice}$ denotes Dice loss, $\mathcal{L}_m$ denotes the loss for the multi-task network optimization, $\mathcal{L}^{Seg}$ denotes the loss computed from the segmentation result, and $\mathcal{L}^{Cls}$ denotes the loss computed from the classification result.

We adopted Pytorch to implement the proposed framework, and the multi-task network was trained on Nvidia RTX 3070 with 8 GB memory. During training, the batch size was set to 16, the maximum epoch number was 300, and the initial learning rate was set to 0.0005. We decayed the learning rate with cosine annealing [20] for each epoch, and the minimum learning rate was set to 0.000001. The restart epoch of cosine annealing was set to 300, ensuring that the learning rate monotonically decreased during the training process. For optimization, we used the AdamW optimizer [21] in our experiments. The whole training takes about 6 h and the inference time for a sample is about 0.2 s.

3 Experimental Results and Discussion

Datasets and Evaluation Metrics. We evaluated our framework on five ultrasound datasets (each representing a source) collected from multiple centers using different ultrasound scanners, including three liver datasets and two thyroid nodules datasets. A detailed description of the collected datasets is provided in Table S1 of the *Supplementary Materials*. We used the dataset with the largest sample size as the training source to prevent overfitting, while the other datasets were the target sources. For each datasets, we randomly split 20% of the samples for test, and used the remaining 80% for training the network. All the results in this study are based on the test set. In the training set, 20% data was randomly selected as validation set. In the data preprocessing, the input images were resized to 224×224 and were normalized by dividing 255.

AUROC is used to evaluate the classification performance. DSC is used to assess the performance of the segmentation. The DSC is defined as:

$$DSC = \frac{2TP}{FP + 2TP + FN} \tag{3}$$

where TP refers to the pixels where both the predicted results and the gold standard are positive, FP refers to the pixels where the predicted results are positive and the gold standard are negative, and FN refers to the pixels where the predicted results are negative and the gold standard are positive.

Table 1. Comparison of segmentation and classification performance of different augmentation methods in five ultrasound datasets in terms of DSC (%) and AUROC (×100%). Training/Target: Training/Target source datasets. MixStyleAug: mixed style augmentation. FeatAug: feature augmentation. MaskAug: mask-based style augmentation. LD: liver dataset. TD: thyroid nodule dataset.

Method	Metric	LD1	LD2	LD3	TD1	TD2
		Training	Target	Target	Training	Target
Traditional Augmentation	DSC	94.5	88.3	89.6	64.0	63.1
	AUROC	86.6	61.3	65.6	72.6	62.3
MixStyleAug	DSC	94.0	87.3	91.1	62.8	65.7
	AUROC	87.9	64.0	68.9	78.1	62.3
MixStyleAug+FeatAug	DSC	94.0	86.9	90.2	63.9	65.2
	AUROC	89.7	66.3	68.8	**85.5**	**64.2**
MixStyleAug+FeatAug+MaskAug	DSC	**94.8**	**89.7**	**91.2**	**77.9**	**65.9**
	AUROC	**92.3**	**67.3**	**69.3**	83.0	62.4

Ablation Study. We evaluated the effects of MixStyleAug, FeatAug, and MaskAug by training a multi-task network with different combinations of these augmentation strategies. Table 1 shows that MixStyleAug improves the segmentation and classification performance on the target sources compared to traditional augmentation. Furthermore, The combination of FeatAug and MixStyleAug improves the classification performance slightly in the liver datasets and significantly in the thyroid nodule datasets. This improvement is due to the style transfer at the feature level, which make the augmented features more similar to the target sources.

Using MaskAug improved both segmentation and classification performance on both training and target sources, compared to the combination of FeatAug and MixStyleAug. This resulted in excellent performance. Figure 3 shows that the mask-based stylized content image has a more similar distribution to the style image than the other images, which helps the model perform better on both training and target sources.

Comparison with Previous Studies. We compared our proposed method with BigAug [3], the style augmentation method by Hesse *et al.* [8], AutoAug [10], and UDA [22] on our collected datasets. Table 2 shows that our method performs excellently on both training and target sources. Unlike BigAug [3], our method uses style augmentation instead of intensity transformations, which avoids a drop in classification performance. Hesse *et al.* [8] only uses training sources for style

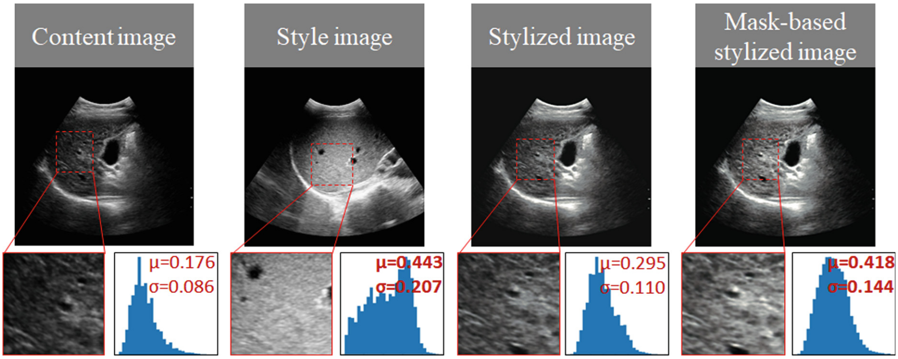

Fig. 3. Illustrations of the conventional style transfer and mask-based style transfer in an ultrasound image. A neural style transfer network is used to translate the content image to the style image, resulting in a stylized image with reference to the style of the entire style image. In contrast, mask-based stylized images are generated with reference to the style of the liver substance in the stylized image. The histogram shows the intensity distribution of the liver region, with μ and σ representing the mean and standard deviation of the liver parenchyma in the ultrasound image, respectively.

Table 2. Segmentation and classification performance of our proposed framework and previous studies in five ultrasound datasets in terms of DSC (%) and AUROC (×100%). Training/Target: Training/Target source datasets. LD: liver dataset. TD: thyroid nodules dataset. UDA: unsupervised domain adaptation.

Method	Metric	LD1	LD2	LD3	TD1	TD2
		Training	Target	Target	Training	Target
BigAug [3]	DSC	93.8	88.0	90.8	54.5	56.4
	AUROC	82.1	59.2	51.9	62.7	**63.9**
Hesse *et al.* [8]	DSC	92.6	86.9	91.3	61.4	62.7
	AUROC	79.6	64.9	68.3	71.7	58.8
AutoAug [10]	DSC	93.9	87.6	**91.4**	68.7	53.5
	AUROC	87.1	65.3	67.8	76.9	39.2
UDA [22]	DSC	94.2	57.7	64.9	65.1	35.2
	AUROC	84.6	61.9	67.9	67.2	55.3
Proposed method	DSC	**94.8**	**89.7**	91.2	**77.9**	**65.9**
	AUROC	**92.3**	**67.3**	**69.3**	**83.0**	62.4

augmentation, which fail to improve performance on target sources, especially in classification tasks, when using a small-sized, single-source training dataset. Our method outperforms AutoAug [10], which relies on large samples to obtain the optimal augmentation strategy. UDA [22] is hard to train with a small-sized dataset due to overfitting and the complex adversarial training.

4 Conclusion

We proposed an augmentation framework based on style transfer method to improve the segmentation and classification performance of the network on ultrasound images from multiple sources. Our framework consists of MixStyleAug, FeatAug, and MaskAug. MixStyleAug integrates the image information from various sources for well generalization, while FeatAug increases the number of styles at the feature level to compensate for potential style variations. MaskAug uses the segmentation results to guide the network to focus on the style information of the ROI in the ultrasound image. We evaluated our framework on five datasets from various sources, and the results showed that our framework improved the segmentation and classification performance across different sources.

References

1. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
2. Shen, D., Wu, G., Suk, H.-I.: Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. **19**(1), 221–248 (2017)
3. Zhang, L., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging **39**(7), 2531–2540 (2020)
4. 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) (2020)
5. Gao, Y., et al.: A universal intensity standardization method based on a many-to-one weak-paired cycle generative adversarial network for magnetic resonance images. IEEE Trans. Med. Imaging **38**(9), 2059–2069 (2019)
6. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
7. Jackson, P.T., et al.: Style augmentation: data augmentation via style randomization. In: CVPR Workshops (2019)
8. Hesse, L.S., et al.: Intensity augmentation to improve generalizability of breast segmentation across different MRI scan protocols. IEEE Trans. Biomed. Eng. **68**(3), 759–770 (2021)
9. Yamashita, R., et al.: Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. IEEE Trans. Med. Imaging **40**(12), 3945–3954 (2021)
10. Cubuk, E. D., et al.: Autoaugment: learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)

11. Jing, Y., et al.: Neural style transfer: a review. IEEE Trans. Visual. Comput. Graph. **26**(11), 3365–3385 (2020)
12. Andreini, P., et al.: Image generation by GAN and style transfer for agar plate image segmentation. Comput. Methods Prog. Biomed. **184**(105268) (2020)
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., et al.: Photorealistic style transfer via wavelet transforms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
15. Hendrycks, D., et al.: Augmix: asimple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019)
16. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
17. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
18. Tang, Z., et al.: Crossnorm and selfnorm for generalization under distribution shifts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
19. Milletari, F., Navab, N., Ahmadi, S.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV) (2016)
20. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations, New Orleans (2019)
22. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 1180–1189 (2015)