Self-adaptive Transfer Learning for Multicenter Glaucoma Classification in Fundus Retina Images

Yiming Bao[1], Jun Wang[1], Tong Li[1], Linyan Wang[2], Jianwei Xu[1], Juan Ye[2], and Dahong Qian[1](✉)

[1] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
dahong.qian@sjtu.edu.cn

[2] The Department of Ophthalmology, The Second Affiliated Hospital of Zhejiang University, College of Medicine, Hangzhou, China

Abstract. The early screening of glaucoma is important for patients to receive treatment in time and maintain eyesight. Deep learning (DL) based models have been successfully used for computer-aided diagnosis (CAD) of glaucoma. However, a DL model pre-trained on certain dataset from one hospital may have poor performance on other hospital data, therefore its applications in the real scene are limited. In this paper, we propose a self-adaptive transfer learning (SATL) strategy to fill the domain gap between multi-center datasets. Specifically, the encoder of a DL model that is pre-trained on the source domain is used to initialize the encoder of a reconstruction model. Then, this reconstruction model is trained using only unlabeled image data from the target domain, which makes the encoder in the model adapt itself to extract useful features both for target domain images encoding and glaucoma classification, simultaneously. Experimental results on a private and two public glaucoma diagnosis datasets demonstrate that the proposed SATL strategy is effective. Also, it meets the real scene application and the privacy protection policy due to its independence from the source domain data.

Keywords: Glaucoma diagnosis · Transfer learning · Multi-center domain adaptation

1 Introduction

Glaucoma is one of the most primary leading causes of blindness [10]. The loss of sight due to glaucoma is irreversible while some other eye diseases such as myopia and presbyopia are not. Thus, early diagnosis of glaucoma for effective treatment and vision conservation matters a lot for patients.

However, the symptoms of glaucoma in the early stage are difficult to perceive. One of the standard methods widely used by eye specialists nowadays is

Y. Bao and J. Wang are co-first authors. J. Ye and D. Qian are co-corresponding authors.

H. Fu et al. (Eds.): OMIA 2021, LNCS 12970, pp. 129–138, 2021.
https://doi.org/10.1007/978-3-030-87000-3_14

the optic nerve head (ONH) assessment [10] in fundus retina images. Whereas, mastering the tricks of performing ONH assessment remains challenging. Therefore, some automatically calculated parameters were presented and popularized as quantitative clinical measurements, such as cup to disc ratio (CRD) which means the ratio of vertical cup diameter to vertical disc diameter in the fundus retina image. Generally, a larger CRD represents a higher possibility of glaucoma and vice verse. However, manually labeling the mask of the cup or disc region is labor-consuming, which makes image-level category labels necessary and reasonable for automatically screening glaucoma.

In the past several years, Deep Learning (DL) based methods have received unprecedented attention and achieved state-of-the-art performance in many fields, including medical image analysis [14]. Glaucoma can be screened from fundus retina images by DL models which are well trained on sufficient data and precise image-level labels [4]. However, DL models trained on one single site cannot be directly generalized and applied to other sites. The distributions of training and testing data are partially different so the pre-trained model may fail to fulfill the diagnosis task.

Commonly, the difference between datasets can be seen as a domain gap. For Example, the discrepancy between images from different dataset can be reflected in many image statistical traits, such as color style, contrast, resolution, and so on. Also, the joint distributions of images and labels may be quite different between the source and the target domain, i.e., $P(x^s, y^s) \neq P(x^t, y^t)$. This is mainly because the margin distributions are different, i.e., $P(x^s) \neq P(x^t)$ even if the conditional distributions, i.e., $P(y^s|x^s)$ and $P(y^t|x^t)$ are similar. Many methods have been proposed to solve this problem. Fine tuning [19] is most widely used in real practical applications. However, fine-tuning is unable to apply when the dataset from a new target domain is completely unlabeled.

To solve the domain adaptation problem, a novel *self-adaptive transfer learning* (SATL) framework is proposed in this paper for glaucoma diagnosis. Specifically, we train a convolutional neural network in the source domain with sufficient labeled data. Then, the feature extraction layers of this trained model is shared as the encoder of a reconstruction network. The reconstruction network is trained in the target domain using only unlabeled data. The encoder is adapted to fit the distribution of target data while maintains the ability for glaucoma diagnosis. The contributions of this paper can be concluded as follows:

(1) To the best of our knowledge, our work is the first to investigate the study of transfer adaptation learning for the classification of glaucoma with multicenter fundus retina images.
(2) Our framework only uses unlabeled date in the target domain and is independent from source domain data, so it has great potential for real scene applications and can meet privacy protection policy for medical data.
(3) Experimental results shows that our framework can preserve most of the classification ability of the off-shelf model and meanwhile improve its classification performance in target domain data. Even totally independent from source domain data, it outperforms other state-of-the-art domain adaptation

methods such as CycleGAN, which heavily relies on source domain data in adaptation stage.

2 Related Works

Transfer adaptation learning (TAL) [20,22] is the most relevant area with the proposed method. It is a combination of transfer learning (TL) and domain adaptation (DA) and can be categorized into three classes, which will be introduced respectively.

Instance Re-weighting Adaptation Learning (IRAL). Methods in this area assign weights to the source domain instances based on their similarity to the target domain instances [13,24]. Via re-sampling or importance weighting, the performance of the trained source classifier in the target domain can be enhanced. However, the estimation of the assigned weights is under a prior-decided parametric distribution assumption [22], which may differ from the true parametric distribution.

Feature Adaptation Learning (FAL). For adapting datasets from multiple domains, methods in this category are widely proposed to find a feature representation space where the projected features from target and source domain follow similar distributions [15,21]. In the past few years, the most famous FAL methods are GAN-based domain adaptation models. However, finding a general feature space for most domains remains challenging. Also, training a GAN-based domain adaptation model needs both source and target domain data, which is more and more impractical in the real scene due to the privacy protection policy for medical data.

Self-supervised Transfer Learning (SSTL). Algorithms in this category focus on training a supervised classifier on the source domain and then transfer its knowledge to the target domain via self-supervised learning [2,3,5,17]. For example, Cheplygina *et al.* [3] investigated a Gaussian texture features-based classification model of chronic obstructive pulmonary disease (COPD) in multicenter datasets. These methods integrate the data information from different domains by extracting some manually designed features from images, which limits the generalization ability of model. Ghifary *et al.* [5] is the most relative literature with our framework. Our method differs from [5] mainly in the network structure. Moreover, we explore application in glaucoma diagnosis in several datasets.

3 Method

The framework of the proposed method is illustrated in Fig. 1. The proposed SATL framework can transfer a pre-trained source classification model to a target domain without using neither source images nor labels.

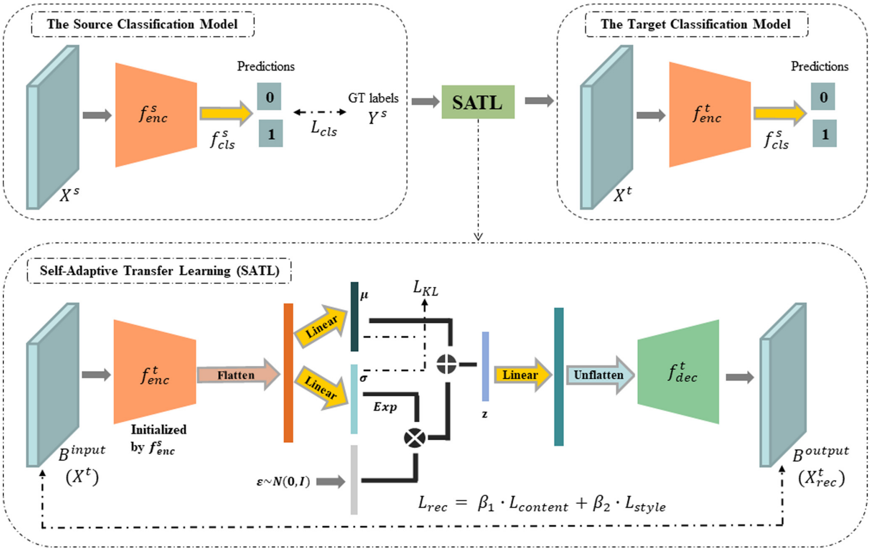


Fig. 1. Illustration of the self-adaptive transfer learning (SATL) strategy, which is independent of the source domain data and more suitable for the real scene applications.

Let $f^s : \mathcal{X}^s \rightarrow \mathcal{Y}^s$ be the source pre-trained classification model and $f^t : \mathcal{X}^t \rightarrow \mathcal{X}^t_{rec}$ be the target reconstruction model. The feature encoder is denoted as $f_{enc} : \mathcal{X} \rightarrow \mathcal{F}$ and the lightweight classification function $f_{cls} : \mathcal{F} \rightarrow \mathcal{Y}$. We denote one more function: an decoder $f_{dec} : \mathcal{F} \rightarrow \mathcal{X}$ in f^t. Then, given an input sample x, f^s and f^t can be formulated as:

$$f^s(x) = f^s_{cls}(f^s_{enc}(x)); f^t(x) = f^t_{dec}(f^t_{enc}(x)) \tag{1}$$

Once $f^t(x)$ is trained, we can build the self-adapted classification model $f^t_{SA}(x)$ for target domain image classification by $f^t_{SA}(x) = f^s_{cls}(f^t_{enc}(x))$

As shown in Fig. 1, the reconstruction model f^t_{dec} is implemented as a variational auto-encoder (VAE), which can compress the image information and sample a latent vector z. The encoder of it f^t_{enc} is initialized by the pre-trained source encoder f^s_{enc}.

The loss function used to optimize the proposed self-adaptive reconstruction model can be represented as:

$$L(f^t_{enc}, f^t_{dec}, x^t) = \alpha \cdot L_{KL} + \beta \cdot L_{rec}, \tag{2}$$

$$L_{KL} = -KL(f^t_{enc}(z|x^t)|f^t_{dec}(z|x^t)), \tag{3}$$

where the first term in the loss function L_{KL} is the KL divergency of the latent vector distribution and the true data distribution. The second term L_{rec} is the reconstruction loss between the output image and the input image.

Instead of using a single MSE loss, we perform a new designed combination of two loss functions following [9]. We argue that the self-adaptive reconstruction model should be guided to reconstruct high-level style information in the target domain images rather than just the pixel-wise texture. Thus, the reconstruction loss function designed in this paper is as:

$$L_{rec} = \beta_1 \cdot \sum_{i,j,k} (B_{ijk}^{output} - B_{ijk}^{input})^2 + \beta_2 \cdot \sum_{m,n} (G_{mn}^{output} - G_{mn}^{input})^2, \tag{4}$$

where B^{output} and B^{input} denote the output and input of the reconstruction model, respectively. i, j, k and m, n represent the position indexes. G^{output} and G^{input} are the Gram matrices of B^{output} and B^{input}. The gram matrix can be calculated as:

$$G = \frac{1}{n_i \times n_j \times n_k} \mathbf{v}\mathbf{v}^{\mathbf{T}}, \tag{5}$$

where $\mathbf{v}$ is the flattened column vector of B^{output} or B^{input}.

4 Experiments and Results

4.1 Datasets

Table 1. The statistical difference between three datasets

Dataset	Domain	Samples	Pos vs. Neg	Avg of image size
LAG (public)	Source/Target	4854	3143:1689	300 × 300
pri-RFG (private)	Source/Target	1881	1013:868	989 × 989
REFUGE (public)	Target only	400	40:360	1062 × 1062

We used two public datasets and one private dataset to validate the proposed SATL framework on glaucoma diagnosis task. The first public dataset is large-scale attention-based glaucoma (LAG) dataset [8] established by Li *et al.*. The second is from the REFUGE challenge [12]. Moreover, we also collected 1881 retina fundus images from one collaborated hospital and built a private dataset (pri-RFG) via labeling all the images by experienced ophthalmologists. The details of the above-mentioned three datasets (LAG, REFUGE, pri-RFG) are summarized and tabulated in Table 1. We can observe that the scales, the average size of images and the ratio of samples in different datasets are quite various, making transfer learning between them challenging. Due to the small number of samples in dataset REFUGE, we just used it as target domain dataset, while LAG and pri-RFG are used for cross-domain evaluation. In other words, we implemented a total of four groups of experiments. Based on the direction from source domain to target domain, they can be represented as LAG → pri-RFG, pri-RFG → LAG, LAG → REFUGE and pri-RFG → REFUGE. When used as a source domain dataset, we separated training and validation set. When used as a target domain dataset, all the images were fed into the reconstruction model to train and adapt the encoder layers.

4.2 Implement Details and Evaluation Metrics

Both the source classification model and the target reconstruction model were implemented using Pytorch (version 1.3.0) and trained on an NVIDIA RTX 2080Ti GPU. We implemented the source classification model as a VGG [16] and optimized it with cross entropy (CE) loss [11]. During the training stage of the source classification model, we set the learning rate as 10^{-6}, weight decay as 5×10^{-4}. All the samples in the source domain were split into training set and validation set using a ratio of 7:3 empirically, following stratified sampling method to ensure that the Pos vs. Neg ratios in each set are similar. At each iteration, a mini-batch of 16 samples were fed into the model. The number of training epochs was set as 50. To avoid the over-fitting issue, the model which achieved the maximum accuracy in the validation set was saved.

During the training stage of the self-adaptive reconstruction model on the target dataset, the learning rate of the encoder was set as 10^{-7} and that of the rest layers was set as 10^{-3}. To avoiding over-fitting on the reconstruction task and losing the ability to extract features that are useful for classification task, the target reconstruction model was trained for only 20 epochs. We empirically set the weights α, β_1 and β_2 in the reconstruction loss function as 0.3, 0.2, 0.5, and the channel number of the latent vector in the model as 32.

Once the target reconstruction model was trained, the self-adapted encoder of it was used as the feature extractor of a target classification model. The last lightweight FC layer of the source classification model played a role as classifier. This new combined target classification model was evaluated on target domain dataset by metrics in terms of Accuracy, Recall, Precision, F1 score and Area Under the ROC Curve (AUC).

4.3 Results and Discussion

As described in Sect. 4.2, based on the three available datasets, there are four executable domain adaptation directions denoted as LAG → pri-RFG, pri-RFG → LAG, LAG → REFUGE, and pri-RFG → REFUGE. For validating the effectiveness of the proposed SATL strategy, on each experiment direction we compared the performance of proposed method (**w/ SATL**) with the source classification model (**w/o SATL**) and a state-of-the-art CycleGAN-based domain adaptation method [23] (**w/ CGAN**). The CycleGAN-based method trains a generator to transfer the target images to the source domain by adversarial learning. The most noteworthy difference between CycleGAN and the proposed SATL strategy is that: our method is completely independent of the source domain data while CycleGAN is not. More specifically, training CycleGAN to perform domain

Table 2. The classification performance of four groups of experiments

Direction	LAG → pri-RFG			pri-RFG → LAG		
Strategy	w/o SATL	w/ CGAN	w/ SATL	w/o SATL	w/ CGAN	w/ SATL
Accuracy	0.799	0.672	**0.856**	0.352	**0.628**	0.579
Recall	0.659	0.422	**0.726**	**1.000**	0.707	0.779
Precision	0.807	**0.923**	0.855	0.352	**0.481**	0.445
F1 score	0.726	0.580	**0.785**	0.521	**0.573**	0.566
Direction	LAG → REFUGE			pri-RFG → REFUGE		
Strategy	w/o SATL	w/ CGAN	w/ SATL	w/o SATL	w/ CGAN	w/ SATL
Accuracy	0.933	0.913	**0.945**	0.240	0.540	**0.580**
Recall	0.425	**0.600**	0.500	**0.975**	0.825	0.850
Precision	0.810	0.558	**0.909**	0.114	0.157	**0.173**
F1 score	0.557	0.579	**0.645**	0.204	0.264	**0.288**

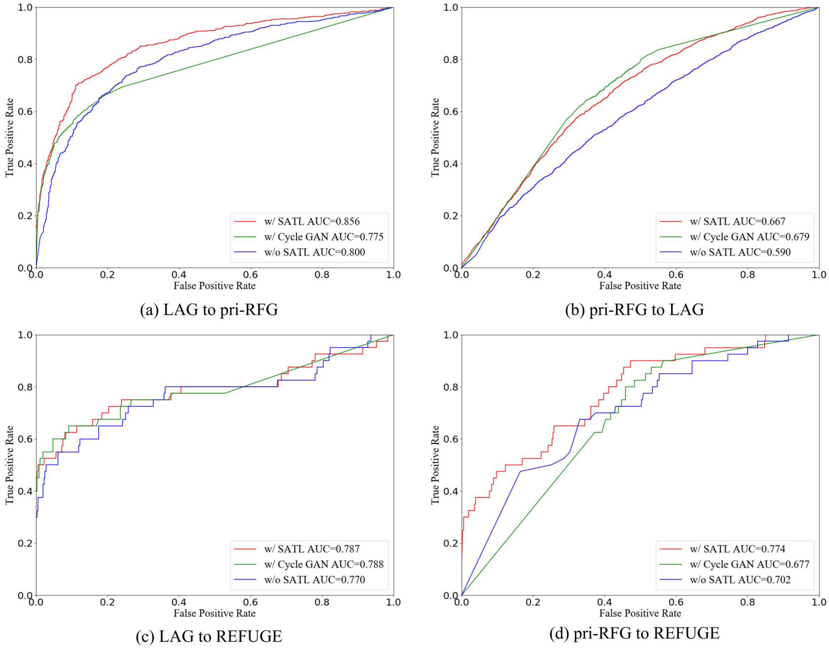
adaptation needs both source and target domain images. On the contrary, the proposed SATL strategy relies on only the target domain unlabeled images.

The experimental results of three strategies are tabulated in Table 2. Moreover, the ROC curves are also plotted and illustrated in Fig. 2. By observing the demonstrated results, two main conclusions can be drawn:

(1) Compared to the source model without SATL, which can be seen as a baseline, the model with SATL outperforms in all four domain adaptation directions in terms of Accuracy and F1 Score. Despite there exist a mass of differences between three used datasets, SATL shows to be effective for self-supervised domain adaptation regardless of the source and target domain data distribution. This phenomenon shows that the proposed SATL is valuable and reliable for the production of pseudo labels in data from a grand-new hospital.

(2) When testing the source model in the target domain images transferred by CycleGAN, the performance is comparable with the proposed SATL strategy in domain adaptation directions of pri-RFG → LAG and LAG → REFUGE. While in directions of LAG → pri-RFG and pri-RFG → REFUGE, the proposed SATL strategy surpasses the CycleGAN by a large margin. This phenomenon demonstrates that SATL is more robust and have more stable generalization ability in different domain adaptation scenes. Note that CycleGAN uses the source domain images in the domain adaptation stage while the proposed SATL does not. Thus, our method which is completely independent of the source domain is more feasible in real scene applications. It can ensure the isolation of multi-center datasets and meet the privacy protection policy.

Discussion. Despite the proposed method improves the performance of the classification model in the target domain via self-supervised training, there still remains some research worth exploring for enhancing the performance. For example, in this paper, we directly trained and validated the source classification model on the source domain. However, it may be a better option to initialize the source classification model by a model pre-trained on large scale nature image datasets such as ImageNet. Besides, the backbone used in this paper is VGG for the convenience of building the reconstruction VAE model. In the future, it can also be replaced by other state-of-the-art backbone such as Inception [18] or SENet [6]. Last but not least, the features adapted by SATL framework in the target domain need to be explore and compare with that before SATL. Further improvement in glaucoma diagnosis may be achieved by learning features which can better represent ONH traits.



(a) LAG to pri-RFG

(b) pri-RFG to LAG

(c) LAG to REFUGE

(d) pri-RFG to REFUGE

Fig. 2. ROC curves of the models evaluated in all four domain adaptation directions.

5 Conclusion

In this paper, we present a self-adaptive transfer learning (SATL) strategy to fill the domain gap between multicenter datasets and perform the evaluation in glaucoma classification based on three fundus retina image datasets. Specifically, a reconstruction model is trained using only target domain unlabeled images.

The encoder of this reconstruction model is initialized from a pre-trained source classification model and self-adapted in the target domain. Experimental results demonstrate that the proposed SATL strategy enhances the classification performance in the target domain and outperforms another state-of-the-art domain adaptation method which even utilizes source domain images for training, as well. In the near future, more efforts will be devoted to exploring how to furthermore lifting the performance of the self-supervised domain adaptation method via designing new reconstruction losses. Moreover, we will extend this strategy to other medical image analysis problems.

Acknowledgments. This work was supported in part by Department of Science and Technology of Zhejiang Province - Key Research and Development Program under Grant 2017C03029 and the Biomedical Engineering Interdisciplinary Research Fund of Shanghai Jiao Tong University under Grant YG2020YQ17.

References

1. Ahn, E., Kumar, A., Fulham, M.J., Feng, D., Kim, J.: Unsupervised domain adaptation to classify medical images using zero-bias convolutional auto-encoders and context-based feature augmentation. IEEE Trans. Med. Imag. **39**, 1 (2020)
2. Cheng, B., Liu, M., Suk, H., Shen, D., Zhang, D.: Multimodal manifold-regularized transfer learning for MCI conversion prediction. Brain Imag. Behav. **9**(4), 913–926 (2015)
3. Cheplygina, V., Pena, I.P., Pedersen, J.H., Lynch, D.A., Sorensen, L., De Bruijne, M.: Transfer learning for multicenter classification of chronic obstructive pulmonary disease. IEEE J. Biomed. Health Inform. **22**(5), 1486–1496 (2018)
4. Fu, H., et al.: Disc-aware ensemble network for glaucoma screening from fundus image. IEEE Trans. Med. Imag. **37**(11), 2493–2501 (2018)
5. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 597–613. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_36
6. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 1 (2019)
7. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. Machine Learning (2013)
8. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: a large-scale database and CNN model. Computer Vision and Pattern Recognition (2019)
9. Li, Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer. Computer Vision and Pattern Recognition (2017)
10. Mary, M.C.V.S., Rajsingh, E.B., Naik, G.R.: Retinal fundus image analysis for diagnosis of glaucoma: a comprehensive survey. IEEE Access **4**, 4327–4354 (2016)
11. Ng, S., Perron, P.: Lag length selection and the construction of unit root tests with good size and power. Econometrica **69**(6), 1519–1554 (2001)
12. Orlando, J.I., et al.: Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. **59**, 101570 (2020)

13. Qi, Q., et al.: Label-efficient breast cancer histopathological image classification. IEEE J. Biomed. Health Inform. **23**(5), 2108–2116 (2019)
14. Ravi, D., et al.: Deep learning for health informatics **21**(1), 4–21 (2017)
15. Shen, Y., et al.: Domain-invariant interpretable fundus image quality assessment. Med. Image Anal. **61**, 101654 (2020)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)
17. Sun, Y., Yang, G., Ding, D., Cheng, G., Xu, J., Li, X.: A GAN-based domain adaptation method for glaucoma diagnosis. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
18. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning, pp. 4278–4284 (2016)
19. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imag. **35**(5), 1299–1312 (2016)
20. Wang, M., Deng, W.: Deep visual domain adaptation: a survey. Neurocomputing **312**, 135–153 (2018)
21. Wang, S., Yu, L., Yang, X., Fu, C., Heng, P.: Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE Trans. Med. Imag. **38**(11), 2485–2495 (2019)
22. Zhang, L.: Transfer adaptation learning: a decade survey. Computer Vision and Pattern Recognition (2019)
23. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, pp. 2242–2251 (2017)
24. Zhu, Q., Du, B., Yan, P.: Boundary-weighted domain adaptive neural network for prostate MR image segmentation. IEEE Trans. Med. Imag. **39**(3), 753–763 (2020)