UNSUPERVISED DOMAIN ADAPTATION FOR CROSS-MODALITY RETINAL VESSEL SEGMENTATION VIA DISENTANGLING REPRESENTATION STYLE TRANSFER AND COLLABORATIVE CONSISTENCY LEARNING

Linkai Peng[1‡], Li Lin[1,2‡], Pujin Cheng[1], Ziqi Huang[1], Xiaoying Tang[1]*

[1]Department of Electronic and Electrical Engineering, Southern University of Science and Technology,
Shenzhen, China
[2]Department of Electrical and Electronic Engineering, The University of Hong Kong,
Hong Kong SAR, China

ABSTRACT

Various deep learning models have been developed to segment anatomical structures from medical images, but they typically have poor performance when tested on another target domain with different data distribution. Recently, unsupervised domain adaptation methods have been proposed to alleviate this so-called domain shift issue, but most of them are designed for scenarios with relatively small domain shifts and are likely to fail when encountering a large domain gap. In this paper, we propose DCDA, a novel cross-modality unsupervised domain adaptation framework for tasks with large domain shifts, e.g., segmenting retinal vessels from OCTA and OCT images. DCDA mainly consists of a disentangling representation style transfer (DRST) module and a collaborative consistency learning (CCL) module. DRST decomposes images into content components and style codes and performs style transfer and image reconstruction. CCL contains two segmentation models, one for source domain and the other for target domain. The two models use labeled data (together with the corresponding transferred images) for supervised learning and perform collaborative consistency learning on unlabeled data. Each model focuses on the corresponding single domain and aims to yield an expertized domain-specific segmentation model. Through extensive experiments on retinal vessel segmentation, our framework achieves Dice scores close to target-trained oracle both from OCTA to OCT and from OCT to OCTA, significantly outperforming other state-of-the-art methods.

Index Terms— Unsupervised Domain Adaptation, Disentangling Representation Style Transfer, Collaborative Consistency Learning, Retinal Vessel Segmentation

1. INTRODUCTION

Optical coherence tomography (OCT) and optical coherence tomography angiography (OCTA) are two novel and non-invasive ophthalmic imaging modalities. OCT has
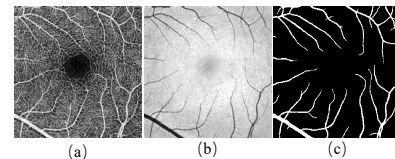
---

‡ These authors contributed equally.

Fig. 1. (a) is an OCTA image, (b) is an OCT image, and (c) is the corresponding RV segmentation label.

micrometer-level resolutions both longitudinally and laterally [1] and OCTA can generate volumetric angiography images in a fast and detailed fashion [2]. These two imaging modalities can both produce detailed visualizations of vascular structures within the retina, exhibiting great potential for improving the diagnosis accuracy of eye-related diseases [3, 4]. Examples of these two modalities are shown in Figure 1. For segmenting retinal vessels (RVs) from OCT and OCTA images for specific clinical purposes, many deep learning methods have been developed [5, 6]. Both 2D and 3D volumes have been utilized to explore various RV segmentation approaches.

In clinical practice, RV segmentation labels are typically annotated on OCTA images due to its relatively clearer visualization of vascular structures. However, segmentation models trained on OCTA images usually perform poorly when tested on OCT images because of domain shift [7]. Thus, it is clinically meaningful to design a RV segmentation framework trained on OCTA images and corresponding labels, but can also work well for segmenting OCT images. In order to solve domain shift related issues, a number of domain adaptation methods have been developed, among which unsupervised domain adaptation (UDA) methods have been gaining greatest popularity in recent years. For instance, Zhou et al. [8] created a segmentation workflow via class-balanced self-training. Tzeng et al. [9] introduced adversarial training into their UDA framework. Further, Tuan-Hung et al. [10] utilized the idea of entropy minimization to design an effective UDA method and Yi-Hsuan et al. [11] improved it by adopting a multi-level adversarial learning scheme. Judy et al. [12] em-

ployed a cycle-consistent adversarial network (CycleGAN). These methods nevertheless contain only one segmentation model, which is one of the main reasons why they fail to converge when the domain gap between training and testing is considerably large. Meanwhile, CycleGAN assumes a one-to-one mapping from source domain to target domain, and thus lacks diversity in the image translation outputs. Moreover, these aforementioned methods are typically designed for addressing small domain gaps, e.g., from GTA5 [13] to Cityscapes [14]. Their performance will decrease remarkably when the data distributions of the source domain and the target domain are quite different, as we will show later.

In such context, we propose a novel UDA framework combining a disentangling representation style transfer (DRST) module and a collaborative consistency learning (CCL) module, which alleviates the aforementioned issues and exhibits superior performance. For DRST, we use a disentangled representation framework under the assumption that images can be decomposed into content components and style codes [15]. Multiple tasks with different objective functions are applied to guide the network to separate apart domain invariants (content components) and domain variants (style codes). Adversarial training is used to train the network to produce images with contents from one domain and styles from the other domain to fool the discriminator. Cycle-consistency is also employed for image reconstruction, which encourages the reconstructed images to be close to the original images. For CCL, we have a source-style segmentation model $\mathcal{F}^{\mathcal{S}}$ and a target-style segmentation model $\mathcal{F}^{\mathcal{T}}$. We first train $\mathcal{F}^{\mathcal{S}}$ on source domain with labeled data, which can be viewed as an expert on the source domain. Then $\mathcal{F}^{\mathcal{S}}$ and $\mathcal{F}^{\mathcal{T}}$ are trained with both original images and transferred images generated by DRST. A collaborative consistency loss is utilized to minimize the distribution differences between the corresponding outputs of the two models.

The main contributions of this paper are three-fold: (1) We propose an innovative UDA framework for RV segmentation across OCT and OCTA images. (2) The UDA framework builds its basis on disentangling representation style transfer and collaborative consistency learning, which can be effectively applied to settings with large domain gaps. We validate our framework on RV segmentation from OCT and OCTA images. (3) Extensive comparison experiments are conducted, both quantitatively and qualitatively, showing that our proposed pipeline achieves superior performance over representative state-of-the-art (SOTA) methods on the publicly-available OCTA-500 dataset [16]. We make our source code available at https://github.com/lkpengcs/DCDA.

2. METHOD

The proposed framework is shown in Figure 2, which consists of a disentangling representation style transfer module and a collaborative consistency learning module.

2.1. Disentangling Representation Style Transfer

2.1.1. Representation Disentanglement

Let $x \in \mathcal{X}, y \in \mathcal{Y}$ respectively be unpaired images from a source domain and a target domain. To separate apart content and style features, we use content encoders $E_{\mathcal{C}}^{\mathcal{X}}$ and $E_{\mathcal{C}}^{\mathcal{Y}}$ to extract content features $\mathcal{C}_{\mathcal{X}}$ and $\mathcal{C}_{\mathcal{Y}}$ from images in the source and target domains. We share weight between the last layer of the two content encoders due to the assumption that images should share a same content latent space. Style codes $\mathcal{S}_{\mathcal{X}}$ and $\mathcal{S}_{\mathcal{Y}}$ can be obtained similarly by style encoders $E_{\mathcal{S}}^{\mathcal{X}}$ and $E_{\mathcal{S}}^{\mathcal{Y}}$ without any weight-sharing strategy since style codes are assumed to be domain-specific. We also use a content discriminator $\mathcal{D}_{\mathcal{C}}$ to discriminate the extracted content features in the two different domains, enforcing $E_{\mathcal{C}}^{\mathcal{X}}$ and $E_{\mathcal{C}}^{\mathcal{Y}}$ to produce encoded content features in a common latent space.

2.1.2. Adversarial Training and Cycle Consistency

Our pipeline also contains two generators, $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$. In the first translation step, $G_{\mathcal{X}}$ uses $\mathcal{S}_{\mathcal{X}}$ and $\mathcal{C}_{\mathcal{Y}}$ to generate source-style images $\hat{x}$ and $G_{\mathcal{Y}}$ uses $\mathcal{S}_{\mathcal{Y}}$ and $\mathcal{C}_{\mathcal{X}}$ to generate target-style images $\hat{y}$. We apply the adversarial loss [17] to both translation directions. Discriminators $\mathcal{D}_{\mathcal{X}}$ and $\mathcal{D}_{\mathcal{Y}}$ are used to discriminate between real images and fake images generated by $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$. Taking source domain as an example, the loss function can be expressed as below

$$
\begin{aligned}
\mathcal{L}_{adv} = {} & \mathbb{E}_{\boldsymbol{x} \sim p_{\mathrm{x}}(\boldsymbol{x})}[\log \mathcal{D}_{\mathcal{X}}(\boldsymbol{x})] \\
& + \mathbb{E}_{\boldsymbol{y} \sim p_{\mathrm{y}}(\boldsymbol{y})}[\log(1 - \mathcal{D}_{\mathcal{X}}(\hat{\boldsymbol{x}}))].
\end{aligned}
\tag{1}
$$

In the second translation step, we use $\hat{x}$ and $\hat{y}$ as inputs for $E_{\mathcal{C}}^{\mathcal{X}}$ and $E_{\mathcal{C}}^{\mathcal{Y}}$. By swapping the style codes again, the two generators $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$ are expected to reconstruct the original images. A cycle-consistency loss [18] is applied to enforce such constraint, namely

$$
\begin{aligned}
\mathcal{L}_{\mathrm{cyc}} = {} & \mathbb{E}_{x \sim p_{\mathrm{x}}(x)} \left[\| G_{\mathcal{X}}(E_{\mathcal{S}}^{\mathcal{X}}(\hat{x}), E_{\mathcal{C}}^{\mathcal{Y}}(\hat{y})) - x \|_1 \right] \\
& + \mathbb{E}_{y \sim p_{\mathrm{y}}(y)} \left[\| G_{\mathcal{Y}}(E_{\mathcal{S}}^{\mathcal{Y}}(\hat{y}), E_{\mathcal{C}}^{\mathcal{X}}(\hat{x})) - y \|_1 \right].
\end{aligned}
\tag{2}
$$

2.2. Collaborative Consistency Learning

Traditional UDA methods use only one segmentation model trained on both source domain and target domain and are likely to converge to non-optimal values when the domain gap is large. Thus, we design a collaborative consistency learning module containing two segmentation models; each model is expected to be an expert on each single domain. We first train our source-style segmentation model $\mathcal{F}^{\mathcal{S}}$ on source domain with the supervision of *ground truth* to get a source-style expert. Then x and $\hat{x}$ are fed into $\mathcal{F}^{\mathcal{S}}$ to get outputs $\mathcal{F}^{\mathcal{S}}_{x}$ and $\mathcal{F}^{\mathcal{S}}_{\hat{x}}$. We input y and $\hat{y}$ into $\mathcal{F}^{\mathcal{T}}$ and obtain $\mathcal{F}^{\mathcal{T}}_{y}$ and $\mathcal{F}^{\mathcal{T}}_{\hat{y}}$. Typically, $\mathcal{F}^{\mathcal{T}}_{\hat{y}}$ is supervised by $\mathcal{F}^{\mathcal{S}}_{x}$ and *ground truth*, while $\mathcal{F}^{\mathcal{T}}_{y}$ is encouraged to be close to $\mathcal{F}^{\mathcal{S}}_{\hat{x}}$.
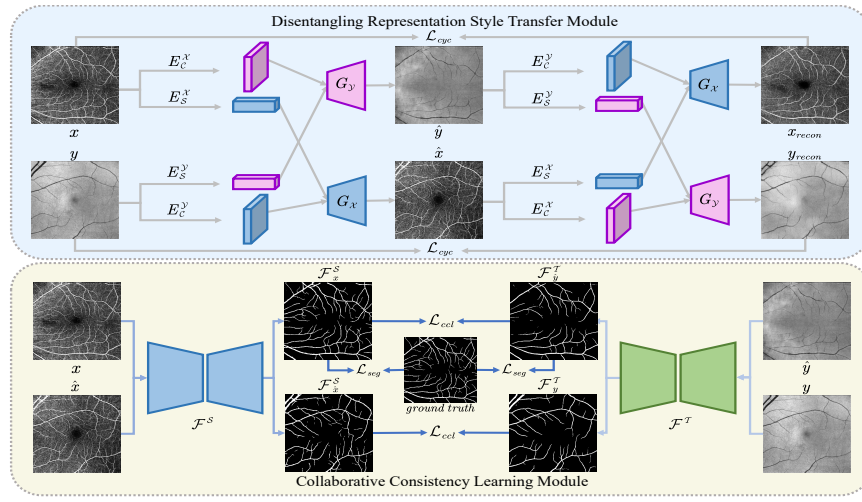
Fig. 2. Schematic demonstration of the architecture of our DCDA framework. The upper part represents the disentangling representation style transfer module and the lower part represents the collaborative consistency learning module.

During training, we set a starting flag τ when the source-style segmentation model $\mathcal{F}^{\mathcal{S}}$ converges to a relatively optimal point, before which only a segmentation loss $\mathcal{L}_{seg}$ is applied. Then we train $\mathcal{F}^{\mathcal{S}}$ and $\mathcal{F}^{\mathcal{T}}$ together and a collaborative consistency loss $\mathcal{L}_{ccl}$ is added to minimize the distribution differences between the corresponding outputs of $\mathcal{F}^{\mathcal{S}}$ and $\mathcal{F}^{\mathcal{T}}$. Available *ground truth* is also used for supervising $\mathcal{F}^{\mathcal{T}}$ via adding another segmentation loss $\mathcal{L}_{seg}$ between $\mathcal{F}^{\mathcal{T}}_{\hat{y}}$ and *ground truth*. The total loss of CCL is formulated as in Eq. 3. For these two models, we use ResNet [19] as the encoder of our U-net [20] architecture to achieve relatively superior performance on RV segmentation.

$$\mathcal{L}_{\text{joint}} = \begin{cases} \mathcal{L}_{\text{seg}}, & \text{epoch} \leqslant \tau, \\ \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{ccl}}, & \text{epoch} > \tau. \end{cases} \quad (3)$$

3. EXPERIMENTS

3.1. Datasets and Image Preprocessing

The OCTA-500 dataset is divided into two subsets according to field of view (FOV). One subset contains 300 samples with $6mm \times 6mm$ FOV, named 6M. The other subset contains 200 samples with $3mm \times 3mm$ FOV, named 3M [16]. We utilize *en-face* OCT images generated by average projection between outer plexiform layer and Bruch's membrane layer and *en-face* OCTA images generated by maximum projection between internal limiting membrane layer and outer plexiform layer. We manually select and form a subset of OCTA-500 (191 from 3M and 254 from 6M), eliminating samples with severe quality issues on either the OCT domain or the OCTA domain. To be noted, the OCT and OCTA images in our subset are unpaired. We resize each image to 384×384. In this subset, 50 images are used for testing and the others are used for training.

3.2. Experimental Setting

We train and evaluate our DCDA pipeline on unpaired UDA tasks in terms of both OCT to OCTA and OCTA to OCT. Such

bi-directional experiments can effectively demonstrate the robustness of our framework. All compared methods and our DCDA are implemented with Pytorch using NVIDIA TITAN RTX GPUs. We use the Adam optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} for all subnetworks included in DRST but no learning rate policy for the segmentation models. We first separately train DRST for 200 epochs and $\mathcal{F}^{\mathcal{S}}$ for 100 epochs. Then we train the whole framework for another 800 epochs. We use Dice loss for $\mathcal{L}_{seg}$ and cross entropy loss for $\mathcal{L}_{ccl}$. The total loss is the summation of $\mathcal{L}_{seg}$ and $\mathcal{L}_{ccl}$. During testing, target domain images are directly inputted to $\mathcal{F}^{\mathcal{T}}$ to get the corresponding predictions.

3.3. Results

All methods are evaluated using two metrics, i.e., Dice[%] and 95% Hausdorff Distance (HD[px]), the results of which are tabulated in Table 1. We compare DCDA with several SOTA UDA models including ADDA [9], ADVENT [10], AdaptSegNet [11], and CyCADA [12]. Apparently, our proposed DCDA achieves the best RV segmentation performance on both OCT to OCTA and OCTA to OCT domain adaptation tasks among all methods. Most compared methods fail in a scenario of such a large domain gap. Intuitively, inverting either OCTA or OCT can reduce their domain gap (see Figure 1). Therefore, we also conduct experiments by inverting the input images of the target domain. Related results are also shown in Table 1. It is obvious that in all directions and all experimental settings, our method achieves Dice scores that are closest to target-trained oracle, significantly outperforming other compared methods.

Representative visualization results are illustrated in Figure 3. It is evident that our framework achieves superior performance on both 3M and 6M images and in both domain adaptation directions.

To evaluate the effectiveness of several key components in DCDA, we conduct ablation studies in Table 2. We compare with our proposed framework without CCL (w/o $\mathcal{F}^{\mathcal{S}}$), without the supervision of ground truth after the starting flag τ (w/o $\mathcal{L}_{seg}$) and without minimizing the distribution differences between the corresponding outputs of $\mathcal{F}^{\mathcal{S}}$ and $\mathcal{F}^{\mathcal{T}}$ (w/o

Table 1. Quantitative evaluations of UDA methods for RV segmentation. We report UDA results both from OCT to OCTA and from OCTA to OCT. * denotes that the input images are inverted. Please note we do not invert input images for Oracle. A two tailed Student's t-test between DCDA and CyCADA identifies p-values that are much smaller than 0.05, indicating statistical significance.

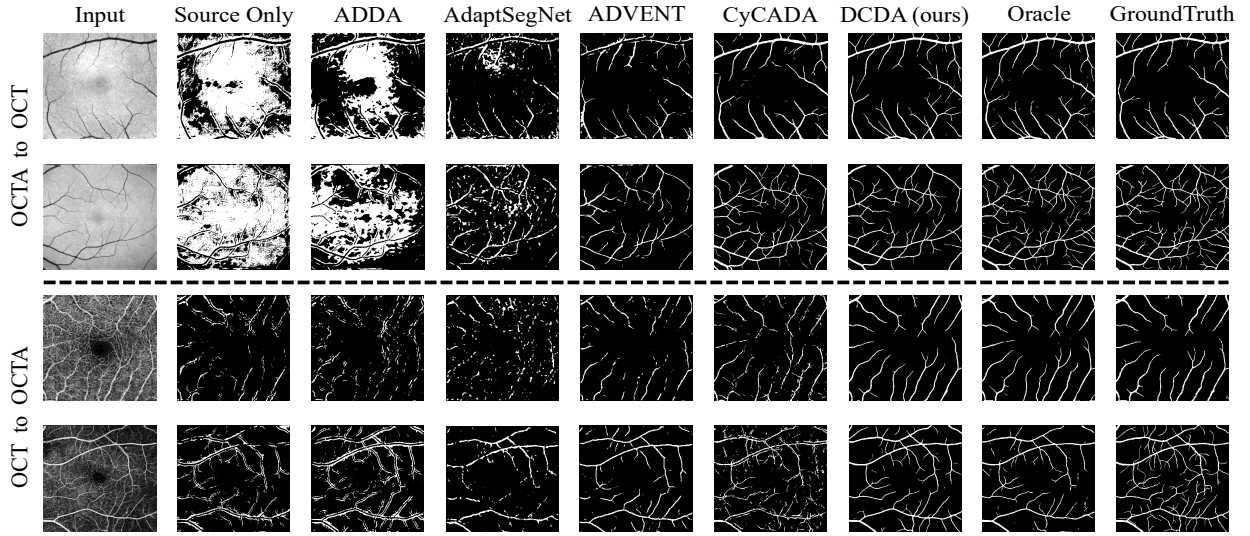| Directions	OCT to OCTA		OCTA to OCT		OCT to OCTA *		OCTA to OCT *	
Metrics	Dice↑	HD↓	Dice↑	HD↓	Dice↑	HD↓	Dice↑	HD↓
Source Only	5.16±1.01	16.29±2.96	5.71±2.04	28.65±12.35	79.37±3.71	16.51±7.12	61.39±15.23	23.99±5.71
ADDA	4.40±1.66	13.72±5.98	4.34±2.47	48.29±31.08	80.14±3.13	14.48±8.02	58.17±16.47	29.38±8.58
AdaptSegNet	6.32±2.49	21.16±4.06	2.65±1.54	34.59±16.41	74.81±3.70	14.89±7.96	58.58±10.73	32.58±6.94
ADVENT	26.22±4.23	15.61±5.47	4.87±2.02	35.69±13.18	78.82±2.90	12.96±5.70	6.11±1.84	45.06±9.03
CyCADA	60.21±5.28	16.52±5.21	18.38±2.67	17.32±4.54	74.94±5.01	18.64±6.94	**79.93±5.52**	**10.71±4.55**
DCDA (ours)	**83.85±3.51**	**9.13±4.49**	**79.28±6.16**	**15.97±6.08**	**84.85±2.47**	**9.41±3.98**	78.92±6.35	15.97±7.46
Oracle	87.48±1.97	6.08±3.73	81.06±5.17	11.94±5.42	87.48±1.97	6.08±3.73	81.06±5.17	11.94±5.42

Fig. 3. Representative visualization results of RV segmentation in two domain adaptation directions. The inputs of the first and the third rows are 3M images and the inputs of the other two rows are 6M images.

$\mathcal{L}_{ccl}$). The results identify the significance of each component in our DCDA framework.

Table 2. Ablation studies of our proposed framework.

Directions	OCT to OCTA		OCTA to OCT	
Metrics	Dice↑	HD↓	Dice↑	HD↓
Source Only	5.16	16.29	5.71	28.65
w/o $\mathcal{F}^{\mathcal{S}}$	55.71	51.39	66.68	27.03
w/o $\mathcal{L}_{seg}$	81.28	13.03	76.99	17.35
w/o $\mathcal{L}_{ccl}$	83.16	11.12	79.27	16.32
Ours	**83.85**	**9.13**	**79.28**	**15.97**
Oracle	87.48	6.08	81.06	11.94

4. CONCLUSION

In this paper, we proposed and validated a novel framework for unsupervised domain adaptation. We disentangled images into content components and style codes and performed style transfer and image reconstruction. Adversarial training and cycle consistency were applied to encourage generators to produce style transferred images as real as possible.

Building on top of this, we further utilized a collaborative consistency learning module to successfully address UDA problems caused by large distribution differences between source domain and target domain. Through extensive quantitative and qualitative experiments, our proposed DCDA method was found to be much more superior than representative SOTA UDA methods, in terms of RV segmentation from OCT and OCTA images.

6. REFERENCES

[1] David Huang, Eric A Swanson, Charles P Lin, Joel S Schuman, William G Stinson, Warren Chang, Michael R Hee, Thomas Flotte, Kenton Gregory, Carmen A Puliafito, et al., "Optical coherence tomography," *Science (New York, NY)*, vol. 254, no. 5035, pp. 1178, 1991.

[2] Talisa E De Carlo, Andre Romano, Nadia K Waheed, and Jay S Duker, "A review of optical coherence tomography angiography (octa)," *International journal of retina and vitreous*, vol. 1, no. 1, pp. 5, 2015.

[3] Alireza Osareh, Majid Mirmehdi, Barry Thomas, and Richard Markham, "Classification and localisation of diabetic-related eye disease," in *European Conference on Computer Vision*. Springer, 2002, pp. 502–516.

[4] Li Lin, Zhonghua Wang, Jiewei Wu, Yijin Huang, Junyan Lyu, Pujin Cheng, Jiong Wu, and Xiaoying Tang, "Bsdanet: A boundary shape and distance aware joint learning framework for segmenting and classifying octa images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 65–75.

[5] Mohammad Shafkat Islam, Jui-Kai Wang, Samuel S Johnson, Matthew J Thurtell, Randy H Kardon, and Mona K Garvin, "A deep-learning approach for automated oct en-face retinal vessel segmentation in cases of optic disc swelling using multiple en-face images as input," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 17–17, 2020.

[6] Linkai Peng, Li Lin, Pujin Cheng, Zhonghua Wang, and Xiaoying Tang, "Fargo: A joint framework for faz and rv segmentation from octa images," in *International Workshop on Ophthalmic Medical Image Analysis*. Springer, 2021, pp. 42–51.

[7] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *European conference on computer vision*. Springer, 2010, pp. 213–226.

[8] Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[9] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[10] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[11] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.

[12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[13] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun, "Playing for data: Ground truth from computer games," in *European conference on computer vision*. Springer, 2016, pp. 102–118.

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[15] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2402–2417, 2020.

[16] Mingchao Li, Yerui Chen, Zexuan Ji, Keren Xie, Songtao Yuan, Qiang Chen, and Shuo Li, "Image projection network: 3d to 2d image segmentation in octa images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3343–3354, 2020.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.