Segmentation-guided Domain Adaptation and Data Harmonization of Multi-device Retinal Optical Coherence Tomography using Cycle-Consistent Generative Adversarial Networks

Shuo Chen[1], Da Ma[2,3,4,1], Sieun Lee[5,6], Timothy T.L. Yu[1], Gavin Xu[1], Donghuan Lu[1,7], Karteek Popuri[1,8]
Myeong Jin Ju [9,10], Marinko V. Sarunic[1,11,12], and Mirza Faisal Beg[1]
[1]School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada
[2] Department of Internal Medicine, Section of Gerontology and Geriatric Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, USA
[3] Center for Biomedical Informatics, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA
[4] Alzheimer's Disease Research Center, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA
[5]Mental Health & Clinical Neuroscience, University of Nottingham, Nottingham, UK
[6]Precision Imaging Beacon, University of Nottingham, Nottingham, UK
[7]Tencent Jarvis Lab, Shenzhen, China
[8] Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada
[9] School of Biomedical Engineering, University of British Columbia, BC, Canada
[10] Department of Ophthalmology & Visual Sciences, The University of British Columbia, Vancouver, BC, Canada
[11] Institute of Ophthalmology, University College London, London, UK
[12] Department of Medical Physics and Biomedical Engineering, University College London, UK

***Background*: Medical images such as Optical Coherence Tomography (OCT) images acquired from different devices may show significantly different intensity profiles. An automatic segmentation model trained on images from one device may perform poorly when applied to images acquired using a different device, resulting in a lack of generalisability. This study aims to address this issue using domain adaptation methods improved by Cycle-Consistent Generative Adversarial Networks(CycleGAN), especially in cases when the ground-truth labels are only available in the source domain.**

***Methods*: A two-stage pipeline is proposed to generate segmentation in the target domain. The first stage involves the training of a state-of-the-art segmentation model in the source domain. The second stage aims to adapt the images from the target domain into the source domain. The adapted target domain images are segmented using the model in the first stage. Ablation tests were performed with a different integration of loss functions, and the statistical significance of these models is reported. Both the segmentation performance and the adapted image quality metrics are evaluated.**

***Results*: Regarding the segmentation Dice score, the proposed *sppg* model achieves a significant improvement of 46.24% compared to no adaptation, and it reaches 87.4% of the upper limit of the segmentation performance. Moreover , the image quality metrics including the FID and KID score indicate that adapted images with better segmentation also have better image qualities.**

***Conclusion*: The proposed method demonstrates the effectiveness of segmentation-driven domain adaptation in retinal imaging processing. It reduces the labour cost of manual labelling, incorporates prior anatomic information to regulate and guide the domain adaptation, and provides insights into improving segmentation qualities in image domains without labels.**

Corresponding author: Shuo Chen (shuo_chen_4@sfu.ca), Da Ma (dma@wakehealth.edu), Mirza Faisal Beg (faisal-lab@sfu.ca)

***Index Terms*—CycleGAN; Domain Adaptation; Optical Coherence Tomography; Retinal segmentation;**

I. Introduction

Optical Coherence Tomography(OCT) has become a widely used retinal imaging modality for its benefits of capturing 3D cross-sectional high-resolution retinal structures with non-invasive and cost-effective techniques (Ikeda and Lam, 2013). Both the retinal thickness change and the presence of retinal fluid were biomarkers for various eye diseases. For example, retinal nerve fibre layer (RNFL) thinning was associated with diabetic retinopathy (DR) and glaucoma, and fluid accumulation can indicate diabetic macular edema (Pekala et al., 2019). Deep learning techniques were more robust and accurate than traditional image processing methods for automated segmentation of such clinically important features in retinal OCT images (Ma et al., 2021; Roy et al., 2017). However, (OCT) images acquired from different devices may show significantly different intensity profiles due to varying properties in the optical systems. Deep neural network (DNN) models trained on a specific dataset might lack generalisability. For instance, a DNN-based segmentation model trained on OCT images acquired from one device using a specific acquisition protocol can experience a significant performance drop when applied to images from different devices and/or protocols. A brute-force solution is to generate manual ground truth data for each device and protocol and re-train the models either locally through the federated-learning framework, (Lo et al., 2021). However, manual segmentation is costly in terms of time and effort, and it is unfeasible to generate precise manual segmentation for all different devices and scanning protocols. Therefore, a universal and robust automated segmentation algorithm is needed to minimize the labour cost while achieving high segmentation quality across various devices.

Domain Adaptation (DA) addressed the domain shifting problem by exploring the mapping function between the distributions of the data from the source domain and data from the target domain (Murez et al., 2018). In the context of medical images such as OCT, the concept of ”domain” represents a specific device with a specific acquisition protocol with which the data were collected. Unsupervised Domain Adaptation (UDA) is widely used in scenarios where no labels are available in the target domain, but the images in both domains share certain intrinsic similarities, i.e. the images from either domain can be synthesized from each other (Toldo et al., 2020). The novel UDA approach, generative adversarial network (GAN), was first introduced in 2014, where the “adversarial” minimax game is played between the generative model (generator) and discriminative model (discriminator) (Goodfellow et al., 2014). The generator tries to adapt the images from the source domain to the target domain, while the discriminator tries to distinguish between the images from the target domain and the generator. The generator and the discriminator are optimized together recursively until both converge. However, the GAN model suffers from several issues. The dynamic equilibrium is difficult to reach due to the unstable oscillations of the two models. The generator may also produce limited variations of synthesized images that can always “fool” the discriminator, which is called “mode collapse” (Che et al., 2017) (Yan et al., 2017). On the contrary, the discriminator can become too successful such that the generator learns nothing. Therefore, the constraints including hyper-parameters, loss functions and model complexity should be rigorously designed. Data harmonization, aiming to eliminate the site-specific bias while maintaining the biological characteristics when aggregating multi-site datasets, has proven to be effective through domain adaptation (Robinson et al., 2020) (Cong et al., 2019). It has been widely used in multi-site multi-scanner MRI datasets with domain variations due to different acquisition protocols (Tian et al., 2022) (Dinsdale et al., 2021).

Cycle-Consistent Adversarial Networks (CycleGAN) was introduced in 2017, which aims to learn bi-directional mapping functions between the source and target domains (Zhu et al., 2017). However, the original CycleGAN suffered from the issue of generating hallucination patterns that hinder the successful application in medical and clinical data (Rahman et al., 2021). Several loss functions were proposed to enforce pixel-wise adaptation and cycle consistency of the reconstructed images, and the details will be discussed in section II. Researchers have dedicated such cross-modality adaptation architecture to various practical applications. Li et al. and Hoffman et al. integrated the semantic consistency concept into the original CycleGAN architecture, which enforces the original image to be segmented consistently as the reconstructed image, and the features of two domains can be distinguished by the discriminator (Li et al., 2019b) (Hoffman et al., 2018). However, the inaccuracy of the segmentation model may confuse the network from learning the correct structural distribution. Hiasa et al. added the gradient consistency constraint to optimize the segmentation performance near the boundaries of the label for CT-MRI adaptation, where the gradient correlation was a commonly used metric in medical registration that measures the cross-correlation between two images (Hiasa et al., 2018). Li et al. adopted a similar strategy by proposing a soft gradient-sensitive loss for the attention of semantic boundaries (Li et al., 2019a). The phase consistency in the Fourier domain was studied and proven to be effective by Yang et al. (Yang et al., 2020). Zeng et al. further added cross-modality segmentation consistency providing a segmentation model for each modality (Zeng et al., 2021). These works showed a great potential of modifying CycleGAN with proper constraints in semantic tasks.

Researchers have also performed several studies regarding the retinal OCT segmentation and domain adaptation using GAN models. He et al. proposed a multi-stage unsupervised domain adaptation network to perform OCT layer segmentation (He et al., 2020b) (He et al., 2020a). A layer segmentation network (ResUNet) was trained first on Spectralis data, then the auto-encoder was trained on Cirrus images to minimize the segmentation error of the reconstructed images, while the weights of the segmentation network were frozen. The authors observed more retinal surface diffusion and layer boundary shifts using CycleGAN compared to the proposed network. However, the comparison might not be fair as the original CycleGAN paper did not adopt segmentation loss. The observed deformation may not be caused by the changes in the adapted anatomy. In fact, for tasks related to only changed in colour and textures, CycleGAN usually performed well with minimal geometric changes (Zhu et al., 2017). The segmentation model was sensitive to pixel intensities, especially in OCT images with a low signal-to-noise ratio (SNR) due to speckle noise. Seebock et al. used CycleGAN to adapt Cirrus images to the Spectralis image domain, where a pre-trained UNet model was available to perform retinal fluid segmentation specifically on Spectralis images (Seebock et al., 2019). The segmentation performance on the adapted images was comparable to the pre-trained segmentation model. However, both of these works did not reveal the potential of GAN due to the limited number of data, unconsolidated constraints of loss functions, etc.

In this paper, we propose a segmentation-guided CycleGAN network aiming to achieve the universal retinal OCT segmentation model. The experiment set assumes that images from the source domain are fully labelled, while the images from the target domain are not labelled. This study has the following contributions:

1) We propose a novel two-staged CycleGAN-based network designed for layer segmentation of retinal OCT images acquired from different devices
2) We incorporate both the ground-truth segmentation information and the pre-trained segmentation model from the source domain to guide the CycleGAN-based domain adaptation.
3) Our approach reduces the hallucination effect that the original CycleGAN network suffered from, and we obtain a segmentation result comparable to the model trained on ground-truth data.

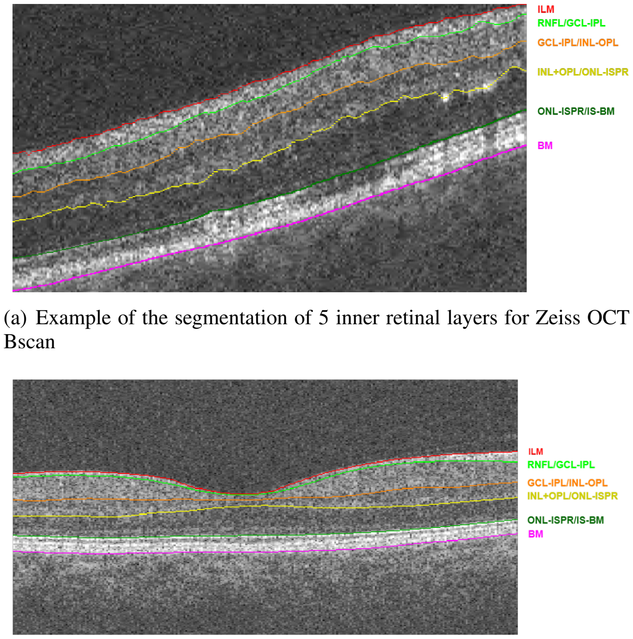
II. Methods

A. Data Acquisition

We adopted two datasets for training and validation purposes. In the first cohort (source domain), 60 OCT volumes of 60 patients acquired from the Zeiss Cirrus 5000 HD-OCT machine (Zeiss Meditec. Inc, Germany) were collected from Vancouver Northshore Clinic. Each volume contained 245 Bscans with a pixel dimension of 1024x245. They were manually labelled with 5 retinal layers, 2 background regions including Vitreous and Choroid-Sclera complex, and fluid. For simplification, the Choroid-Sclera complex will be referred to as Choroid in the following content. More details about the dataset, acquisition protocol and manual segmentation protocol can be found in (Ma et al., 2021). In the second cohort (target domain), 10 OCT volumes of 60 different patients acquired from Topcon 3D OCT 1000 Mk2 machine (Topcon. Inc, Japan) were downloaded from the UK Biobank database and manually labelled with 8 retinal layers (Sir Rory Collins). Each Topcon volume contained 128 Bscans with a resolution of 650x512. All Topcon data were acquired from the healthy patients, i.e. there was no fluid present. We will refer to these two datasets as the Zeiss dataset and Topcon dataset, respectively.

Figure 1 demonstrates examples of the manual segmentation ground truth labels for both the Zeiss dataset and Topcon dataset, respectively. For both datasets, seven regions are defined and labelled (five retinal layers plus two non-reginal regions). The five retinal layers from top to bottom are ILM(Inner Limiting Membrane)-RNFL(Retinal Nerve Fiber Layer), GCL(Ganglion Cell Layer)-IPL(Inner Plexiform Layer) complex, INL(Inner Nuclear Layer)-OPL(Outer Plexiform Layer) complex, ONL(Outer Nuclear Layer)-ISPR(Inner-Segment Photoreceptor Layer) complex, IS(Inner-Segment Layer)-BM(Bruch's Membrane). In addition, the region above ILM (Internal Limiting Membrane) is considered Vitreous, and the region below BM(Bruch's Membrane) is regarded as Choroid. The region above ILM is considered Vitreous, and the region below BM is regarded as Choroid. Each colored trajectory is defined by the boundaries of layers, e.g. RNFL/GCL-IPL means the boundary between the RNFL layer and the GCL-IPL complex.

We applied two common pre-processing steps to optimize the performance (Lee et al., 2017; Ma et al., 2022). Motion correction was performed using phase cross-correlation between adjacent Bscans to calculate both the relative axial and lateral shifts with the first Bscan as a reference, and then 2D spine interpolation was applied to obtain the motion-corrected Bscans. Bounded variation (BV) 3D smoothing was applied to minimize the effect of speckle noise and enhance the contrast of the retinal layer and fluid boundaries. For both Zeiss and Topcon datasets, we cropped the excessive background regions to obtain a Bscan size of 500x231 and 256x512, respectively. It helped both the segmentation and adaptation network to learn more about the retinal layers inside the retinal body. We resized the Bscans from both the Zeiss dataset and the Topcon dataset to a resolution of 512x256 before feeding them into the domain adaptation network.



(a) Example of the segmentation of 5 inner retinal layers for Zeiss OCT Bscan

(b) Example of the segmentation of 5 inner retinal layers for Topcon OCT Bscan

Fig. 1. Demonstration of the segmentation of 5 inner retinal layers for the Zeiss dataset and Topcon dataset, respectively. The corresponding layer boundaries are highlighted and categorized as shown on the right side of the image.

B. Network

The overall pipeline contains two steps, the segmentation network and the device domain adaptation network. The segmentation network was trained first for optimal performance of supervision of the adaptation network. Then, the CycleGAN-based adaptation network was trained while the weight of the segmentation network was frozen. The segmentation loss along with several other auxiliary constraints were combined throughout the training.

1) OCT Segmentation Network

The LF-Unet, a deep neural network integrating U-Net and fully convolutional network(FCN), was used as a pre-trained OCT segmentation network (Ma et al., 2021). As shown in Figure 2, the architecture of the LF-UNet leveraged a symmetrical convolutional network with an encoder-decoder mechanism. Mimicking the original UNet structure, the contracting paths contain 4 down-sampling convolutional blocks to extract high-level features. 2 expansive paths include the original up-sampling path and an FCN path. The output was eventually fed into 3 consecutive dilated convolutional layers for final results.

The network was trained on the Zeiss dataset using the PyTorch Lightning framework with 2-node distributed learning (William Falcon). We set the batch size and learning rate to be 2 and 10^{-4}, respectively. We used Adam optimizer with L2 regularization. We applied early stopping criteria with patience of 3 so that the training was terminated when the validation loss converges. We used the learning scheduler to monitor the validation loss and reduce the learning rate when validation loss plateaued ('ReduceLROnPlateau' command in

PyTorch Lightning). We applied several data augmentations by random horizontal flipping and random spatial rotation with a maximum angle of 10 degrees. We applied a pixel-wise Softmax function for the model output and uses a weighted combination of Dice loss and Cross Entropy Loss to enforce the learning of the fluid under the class imbalance circumstance. We calculated the pixel-wise weight map for each Bscan by assigning more weights to the fluid region and boundary. Given predicted segmentation S_p, ground-truth segmentation S_g and pixel-wise dice weights S_{dw}, class weights S_{cw} the segmentation loss is calculated as:

$$L_{dice} = 1 - \frac{\sum 2 S_p S_g S_{dw}}{\sum S_p S_{dw} \sum S_g S_{dw} + \epsilon} \tag{1}$$

$$L_{ce} = -\sum S_{cw} S_g log(S_p) \tag{2}$$

$$L_{seg} = L_{dice} + L_{ce} \tag{3}$$

To avoid over-fitting and ensure the robustness of the segmentation network performance, 10-fold validation was applied for the entire dataset. Each fold was split into training, validation and testing sets with a ratio of 8:1:1. We applied the volume stratification during the splitting such that the Bscans from any OCT volume were allocated to one set only. It avoided biased evaluation when adjacent Bscans with similar features from the same OCT volume are exposed to multiple training stages. We evaluated the performance of each fold of the model with the average dice score using their corresponding test set. The model with the best average dice score was adopted for the next stage of training of the adaptation network.

2) *Domain Adaptation Network*

Similar to the original CycleGAN architecture, the segmentation-guided adaptation network consists of two generators and two discriminators. As shown in Figure 3, datasets A and B refer to the Zeiss and Topcon dataset, respectively. Image pairs from both datasets were fed into the corresponding generators simultaneously. The output synthetic adapted image pairs were then fed into two branches. The discriminators took a history of the synthesized images to distinguish if the upcoming pairs were real or synthetic. Meanwhile, the reconstructed images were obtained by generators fed with the adapted images. Lastly, the original and reconstructed image pairs were fed into the pre-trained segmentation network to obtain the predicted segmentation pairs.

As shown in Fig 4, the generator block was designed with an encoder-decoder mechanism. We used 2-level convolutional down-sampling to obtain high-level features. It was then fed into 9 consecutive residual blocks(Resblock) to learn cross-domain correlations. Lastly, the adapted latent features were recovered to the original size with Tanh activation. The discriminator block was constructed similarly to VGG16 architecture. Multiple convolutional layers were used for high-level feature discrimination. A single-channel classification map was generated as output.

As shown in Fig 5, we used the LeakyReLU activation function to avoid dying neuron problems for ReLU. To cooperate with this, all the input images were normalized within the range of [-1,1], but they were demonstrated with a re-scaling back to [0,255]. Also, we adopted a dropout layer in Resblock to reduce the effect of over-fitting.

Despite the difference in capacities between the Zeiss and Topcon dataset, we formed the image pairs for adaptation using all available data. For the Zeiss dataset, we adopted the same data split used for the training of the selected segmentation model. For the Topcon dataset, we fixed one volume as a hold-out test set, and perform 9-fold inner cross-validation along with the training and validation dataset from the Zeiss dataset. Specifically, during each fold of training, we split the 9 Topcon volumes with a ratio of 8:1 for training and validation, respectively. Thus, each image pair was formed from Bscans of 6 Zeiss volumes and 1 Topcon volume. For reproducible results, we fixed the combination of volumes and Bscans for image pair formation.

We applied several techniques to optimize the performance of the adaptation network. We applied the same data augmentation as the segmentation network. Since we wanted to adapt the images directly acquired from the OCT devices with few processing steps as mentioned in Section II-A, the intensity-level augmentation may confuse the network of constructing stabilized mapping function. We adopt the tricks used in WassersteinGAN from Arjovsky *et al.* to use RMSProp optimizer for parameter adjustment (Arjovsky et al., 2017). The gradient changes for GAN were usually unstable, so the optimizer using a momentum mechanism may cause gradient clipping. We set the initial learning rate of the discriminator to be 5 times larger than the generator according to Two time-scale update rule (TTUR), which helped the convergence to a local Nash equilibrium and works functionally equivalent to train the discriminator more frequently (Heusel et al., 2017). For losses related to the discriminator, we applied soft/noisy labels with a pixel-wise variation of a normal distribution drawn between [-0.2, 0.2], which helped enhance the generalisability of the whole network. We set a learning rate of 10^{-5} to stabilize the adaptation. The model was trained with a maximum of 100 epochs while the learning rate is linearly decayed after 50 epochs. We applied early stopping criteria that the training stops when the validation segmentation loss of the UKB dataset stops decreasing after 3 epochs. We used the PyTorch Lightning Framework for 2-node distributed parallel training with a batch size of 2.

C. *Domain Adaptation Loss functions*

1) *CycleGAN Losses*

The original CycleGAN paper mentioned three loss functions. GAN loss, or adversarial loss, was designed to evaluate the performance of the discriminator fed with synthetic data. Given the images x in the source domain X (Zeiss dataset) and images y in the target domain Y (Topcon dataset), the generator functions G_{X2Y} and G_{Y2X} for mapping from X to Y and Y to X, respectively, and discriminator functions D_X and D_Y, we have:

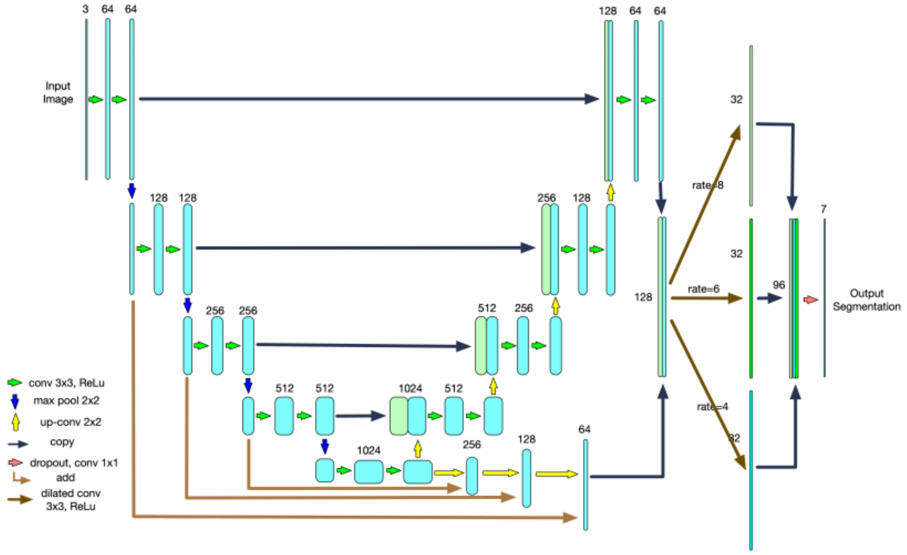


Fig. 2. Illustration of the LF-UNet architecture. It mimics the encoder-decoder design of the original U-Net structure, where the contracting path encoded the high-level features, and the expanding path decoded these features to reconstruct the corresponding semantic information. An extra fully-connected path is added alongside the expanding path to better increase the segmentation accuracy of the retinal layer boundaries. More details are discussed by Ma et al. (Ma et al., 2021) .

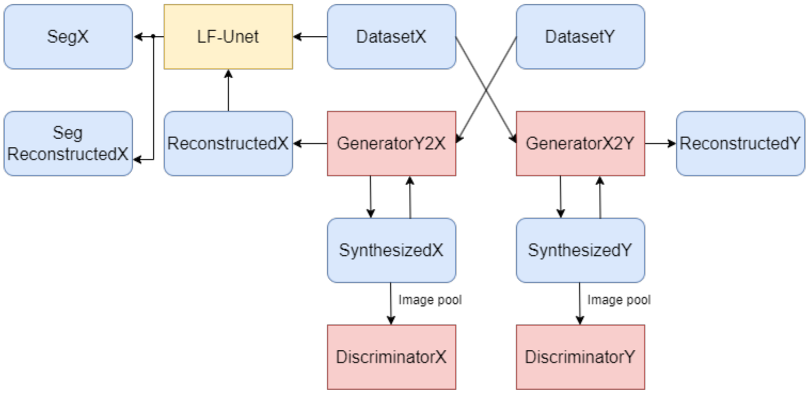


Fig. 3. Workflow for the domain adaptation network. The images from DatasetX and DatasetY are fed into the GeneratorX2Y and GeneratorB2A, respectively. The GeneratorX2Y and GeneratorY2X can then produce adapted images of SynthesizedY and SynthesizedX, respectively. A series of SynthesizedX and SynthesizedY stored in image pools are then fed into the DiscriminatorX and DiscriminatorY accordingly for GAN loss calculations. Meanwhile, the SynthesizedX and SynthesizedY are fed back to the GeneratorX2Y and GeneratorY2X to produce their corresponding reconstructed images ReconststructedY and ReconststructedX, respectively. The images from DatasetX and ReconstructedX images are used by LFUnet to generate the segmentation SegX and SegReconstructedX, which are utilized for semantic loss calculation.

$$L_{GAN}(x,y) = MSE(D_Y(G_{X2Y}(x)),0) + \\ MSE(D_X(G_{Y2X}(y)),0) \quad (4)$$

$$MSE(x,y) = (x-y)^2$$

The adversarial loss can only guide the network with style transfer across domains, but it cannot guarantee the unique mapping of the two domains. Cycle consistency loss was then proposed that the synthetic images should also be reconstructed back to the original images:

$$L_{cyc}(x,y) = \lambda_X \|x - G_{Y2X}(G_{X2Y}(x))\| + \\ \lambda_Y \|y - G_{X2Y}(G_{Y2X}(y))\| \quad (5)$$

Here, the λ_X and λ_Y are tuning parameters from both adapted directions.

To further emphasize the independence of the two adaptation functions, the identity loss was designed such that the original image should remain unchanged if the adaptation function in a reversed direction was applied:

$$L_{idt}(x,y) = \|x - G_{Y2X}(x)\| + \|y - G_{X2Y}(y)\| \quad (6)$$

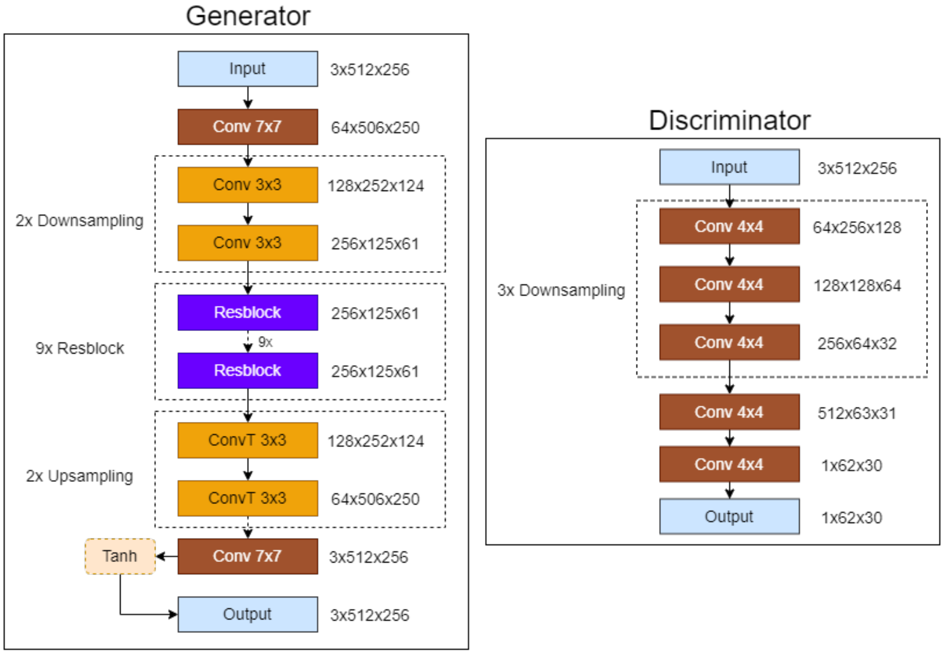


Fig. 4. Generator and Discriminator architecture. The dimension of the output intermediate feature map is shown on the right of each block, which is formatted as $Channel \times Height \times Weight$. The Generator block is implemented as an encoder-decoder architecture. 9 Resblocks are placed at the bottleneck for high-level feature extractions. The Discriminator is implemented with a series of down-sampling convolutions, which ends up with a single-channel classification map filled with probabilities of images being real(1) or synthesized(0). Each element in the classification matrix represents a local region in the original image, which is originally proposed in PatchGAN (Isola et al., 2017)

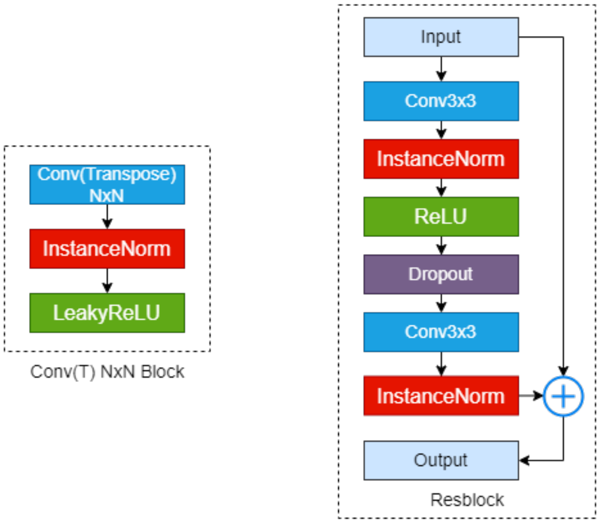


Fig. 5. Convolution block and Resblock used in both Generator and Discriminator blocks. The Conv NxN block is a 3-step operation with the 2D convolution of input size NxN, followed by normalization and LeakyReLU activation layers. The up-sampling operation will replace the standard convolution with transposed convolution. The Resblock is constructed with two consecutive Conv NxN blocks with ReLU activation and a Dropout layer in between.

2) Segmentation Loss

Even though the above loss functions proposed in the original CycleGAN enforced a unique mapping between the two domains, the segmentation model was not guaranteed to perform well onto a such particular solution. As described in Section II-B1, we had a pre-trained segmentation model on the Zeiss dataset. Thus, the semantic consistency loss was designed such that the segmentation model should have comparable performance on the original image x and its reconstructed image. Given the ground-truth labels l for the original image, the segmentation function S, and the same segmentation criterion L_{seg} as Eq. (3), we have:

$$L_{sem}(x, l) = L_{seg}(S(x), l) + L_{seg}(S(G_{Y2X}(G_{X2Y}(x))), l) \quad (7)$$

Since the weights of the segmentation network were frozen during the training of the adaptation network, the first term of the semantic consistency loss should remain relatively stable, and it performed as a “self-awareness” factor of the error caused by the segmentation network itself. The second term will further guide the cyclic adaptation to achieve comparable segmentation.

3) Structure Similarity Loss

Wang et al. proposed Structure Similarity Index Measure (SSIM) to quantify the similarity between two images in luminance, contrast and structure (Wang et al., 2004). Luminance was measured based on the average intensity value, contrast was measured by the standard deviation of the intensity, and structure was calculated based on the normalized correlation between the two images. Based on SSIM, we defined a difference SSIM (DSSIM) loss to maximize the exponentially weighted multiplication of these three metrics between the original image and its reconstruction. Given the mean μ_x μ_y, and standard deviation σ_x σ_y for images x and y, respectively, we have:

$$L_{dssim}(x,y) = (1 - \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)})/2 \quad (8)$$

This expression was obtained by setting the weights for all three losses evenly, and c_1 and c_2 were constants included to prevent the denominator terms from being close to 0.

4) Peak Signal to Noise Ratio Loss

OCT images suffers from speckle noise due to the spatially coherent light source (Bashkansky and Reintjes, 2000). The Peak Signal to Noise Ratio (PSNR) can be characteristic of images from specific OCT devices providing the images were generated consistently. The segmentation model is usually sensitive to the intensity distribution of the image. Therefore, the different PSNR will directly affect the segmentation quality. The PSNR loss was constructed such that the images from the same domain should share the same noise distribution:

$$L_{psnr}(x,y) = -10\log_{10}\frac{max^2(x)}{MSE(x,y)} \quad (9)$$

5) Perceptual Loss

The training of a GAN is time-consuming and difficult to optimize due to the obstacles mentioned in I Introduction. Johnson et al. provided a novel approach that the high-level features of an image should not be lost during adaptation (Johnson et al., 2016). In other words, the original image, synthetic image and reconstructed image are supposed to share structural similarity in high-level convolutional feature extraction layers. In our training, four convolutional layers of a pre-trained VGG16 network from ImageNet were used to extract each level of feature (Simonyan and Zisserman, 2014). It speeded up the converging process by reconstructing the high-resolution image directly from low-resolution features. Given the output of the i^{th} layer of the VGG16 network $f_i(x)$ when fed with image x, and $f_i(y)$ when fed with image y, the size of both input images is (C,H,W), and the perceptual loss is defined as:

$$\begin{aligned} L_{perc}(x,y) = \frac{1}{C \times H \times W} \times \sum_{i=1}^{4}(\|f_i(x) - f_i(G_{X2Y}(x))\| + \\ \|f_i(y) - f_i(G_{Y2X}(y))\| + \|f_i(x) - f_i(G_{Y2X}(G_{X2Y}(x)))\| + \\ \|f_i(y) - f_i(G_{X2Y}(G_{Y2X}(y)))\| + \\ \|f_i(G_{X2Y}(x)) - f_i(G_{Y2X}(G_{X2Y}(x)))\| + \\ \|f_i(G_{X2Y}(y)) - f_i(G_{X2Y}(G_{Y2X}(y)))\|) \quad (10) \end{aligned}$$

Instead of the euclidean distance used in the original paper, we adopted L1 norm for robustness of overall adaptation, since it was common that some low-quality OCT images exist.

6) Gradient Consistency Loss

To further maintain the texture of the source images, especially the layer boundaries where the segmentation network focused on, we proposed gradient consistency loss that enforced a better adaptation of regions with large intensity variations. The assumption was that the adaptation shall not diminish the visually-recognizable edges, especially the boundaries of retinal layers. The edge information was extracted and compared using 1st order image derivative in both x and y direction with 2D Sobel operator:

$$G_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, G_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}$$

Unlike the approach of Li et al. where the attention is emphasized mostly on layer boundaries, we simply used the intrinsic information of the images to maintain more useful information. Let * denote the convolution operation, the gradient consistency loss is constructed as:

$$\begin{aligned} L_{grad}(x,y) = MSE(G_x * x, G_x * G_{X2Y}(x)) + \\ MSE(G_y * y, G_y * G_{Y2X}(y)) \end{aligned} \quad (11)$$

Thus, the overall loss function for the adaptation network is:

$$\begin{aligned} L_{tot} = L_{GAN} + L_{cyc} + \lambda_{idt}L_{idt} + \\ \lambda_{seg}L_{sem} + \lambda_{dssim}L_{dssim} + \lambda_{psnr}L_{psnr} + \\ \lambda_{feat}L_{perc} + \lambda_{grad}L_{grad} \quad (12) \end{aligned}$$

Here, the λ_{idt}, λ_{seg}, λ_{feat}, λ_{dssim}, λ_{psnr}, λ_{grad} along with λ_X and λ_Y mentioned before are tuning hyper-parameters for each corresponding loss term. For all the experiments, we empirically set λ_X and λ_Y to be 20 for cycle consistency loss, 0.5 for λ_{idt}, and all other tuning parameters to be 1.

D. Domain Adaptation Model Evaluation

We evaluated the performance of the domain adaptation on the test sets from the target domain (i.e. domain Y of the Topcon images). Both the segmentation-based and the image-based metrics were evaluated to achieve a comprehensive evaluation of the domain adaptation results that cover task applications.

1) Segmentation-based metrics evaluation

Firstly, the performance of the segmentation network trained on the source domain (Zeiss) was evaluated by computing the segmentation accuracy using the Dice similarity coefficient, or dice score (Equation 13) for each retinal layer labels, between the ground truth manual segmentation in the test set and the automatic segmentation trained from the LF-UNet.

$$Dice = \frac{2(X \bigcap Y)}{|X| + |Y|} \quad (13)$$

where X is the ground truth label maps generated from manual segmentation, and Y is the automatically-generated label maps generated from the proposed whole-body multi-slice segmentation framework.

Secondly, the performance of domain adaptation was further evaluated using each retinal layer label. Dice scores between the auto-segmentation results derived from the first-level segmentation model (LF-UNet) and the automatic segmentation results derived from the domain-adapted images in the source domain were compared.

2) Image-based metrics evaluation

Other than directly evaluating the segmentation performance, we also evaluated the quality of the adapted images, where the variations may not be visible to human eyes. Heusel et al. first proposed the Fréchet Inception Distance (FID) score to evaluate the similarities between two datasets, which was widely used for measuring the performance of GAN via computing the distance between two multidimensional Gaussian distributions (Heusel et al., 2017). It adopted the Inception V3 as a feature extractor and computes the Fréchet distance between the high-level feature maps (Dowson and Landau, 1982) (Szegedy et al.). Given the original and the adapted distribution to be X and Y, μ_X and μ_Y to be the mean of the two distributions, Σ_X and Σ_Y to be the covariance matrices of the two distributions, Tr representing the trace of the matrix, the FID score is formulated as:

$$FID(X,Y) = \|\mu_X - \mu_Y\|^2 - Tr(\Sigma_X + \Sigma_Y - 2\Sigma_X\Sigma_Y) \quad (14)$$

In addition, Gretton et al. proposed a kernel-based method to evaluate the similarity of two-sample distribution using Maximum Mean Discrepancy (MMD) (Gretton et al., 2012). It assumed that two different distributions shall possess different expected values, which can then be utilized to distinguish different empirical datasets (Borgwardt et al., 2006). It maps the L2 distance of two distributions into a universal Reproducing Hilbert Kernel Space (RHKS), providing that the mapping function is a unit ball (Borgwardt et al., 2006). The MMD score requireed the calculation of a polynomial kernel function. Given two data samples x and y, scaling coefficients γ and c, degree of the polynomial kernel n, the polynomial kernel $k(x, y)$ are defined as:

$$k(x,y) = (\gamma \cdot x^T y + c)^n$$

Then, given two data distributions X and Y, the number of samples m and n, for all $x_i \in X, y_i \in Y$, the MMD score is calculated as:

$$MMD = [\frac{1}{m(m-1)} \sum_{i \neq j}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j}^{m} k(y_i, y_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j)]^{\frac{1}{2}}$$

We calculated the mean and standard deviation using the Kernel Inception Distance (KID) via MMD score, which is calculated as:

$$KID(X,Y) = MMD(X,Y)^2 \quad (15)$$

III. Results

We obtained the segmentation results for both stages of the pipeline, i.e. the segmentation network and adaptation network. We also evaluated the quality of the adapted images compared to the original images within the same domain. To maintain consistency with the adaptation network, we evaluated the similarities in the domain of the Zeiss dataset via commonly used metrics mentioned in Section II-D2.

A. Segmentation network

For the segmentation network mentioned in Section II-B1, we applied 10-fold cross-validation to evaluate the segmentation performance. For each fold, the best model was selected with the lowest validation loss. The average dice score was then computed by feeding the corresponding test set data into the selected model. The average dice scores for 8 retinal regions are shown in Table I. The fold 2 model has the highest average dice score among all folds, thus it is used for the next stage of the adaptation network. Notice that the dice score of fluid was also considered as we wanted to minimize the false positive rate of fluid when the segmentation network was applied to Topcon dataset with only healthy patients.

B. Adaptation network

To evaluate the effect of each proposed component in the domain adaptation framework. We performed ablation experiments with 5 different combinations of loss functions, including the original CycleGAN network as our baseline. The other 4 networks add extra constraints onto the baseline network, which are semantic consistency(seg), semantic consistency + perceptual(sp), semantic consistency + perceptual + gradient consistency(spg), and semantic consistency + ssim + psnr + perceptual + gradient consistency($ssppg$). The seg model is equivalent to the method proposed by Hoffman et al. (Hoffman et al., 2018). The spg model is functionally similar to the approach of Li et al. (Li et al., 2019a). The SSIM and PSNR loss were combined for experiments as they both evaluate images at the intensity distribution level. The inference results are directly adopted for calculating the dice score as $1 - L_{dice}$ as mentioned in Eq. (1). Figures 9 and 10 show the dice score for all 7 retinal regions of each adaptation model averaged across all 9 training folds as well as in each fold accordingly. The $ssppg$ model has the best dice score in NFL_IPL, IPL_OPL, OPL_IOS regions, sg has the best dice score in IOS_BM region, and spg has the best dice score in ILM_NFL region. Overall, the $ssppg$ model has the optimal average dice score of 0.8231. Furthermore, all 5 adaptation models outperform the one without the domain adaptation. The segmentation performance of detection of vitreous and choroid is similar for all 5 models.

We applied statistical analysis to compare the segmentation performance of the 5 adaptation models. The one-way Analysis of Variance (ANOVA) with a post-hoc Tukey Honestly Significant Difference (HSD) test is used for pairwise evaluation of means of 5 models. Similar to Figures 9 and 10, we calculated the P-value in terms of layers and training folds,

Dice \ Layer Fold	ILM_NFL	NFL_IPL	IPL_OPL	OPL_IOS	IOS_BM	$Fluid$	$Vitreous$	$Choroid$	$Average$
1	0.8832	0.9354	0.9408	0.9524	0.9329	0.7706	0.9963	0.9908	0.9253
2	**0.9044**	**0.9629**	0.9456	0.9546	0.9477	**0.8342**	0.9904	0.9918	**0.9414**
3	0.9007	0.9622	**0.9541**	**0.9668**	**0.9550**	0.6506	0.9975	0.9939	0.9226
4	0.9008	0.9621	0.9466	0.9514	0.9449	0.6796	**0.9977**	0.9922	0.9219
5	0.8786	0.9390	0.9242	0.9564	0.9392	0.6097	0.9974	0.9911	0.9045
6	0.8941	0.9449	0.9366	0.9480	0.9226	0.7005	0.9974	0.9878	0.9165
7	0.8885	0.9387	0.9104	0.9010	0.9052	0.3764	0.9956	0.9873	0.8629
8	0.9042	0.9509	0.9208	0.9216	0.9354	0.4365	0.9969	0.9910	0.8821
9	0.8853	0.9446	0.9296	0.9299	0.9433	0.5356	0.9968	0.9919	0.8946
10	0.8932	0.9296	0.9219	0.9538	0.9533	0.5996	0.9969	**0.9942**	0.9053

TABLE I

Mean dice score of 7 retinal regions and fluid for the different fold of validations. The best model is selected based on the average dice scores among all regions. The best dice score for each retinal region is highlighted. Fold 2 is chosen with the highest average dice score of 0.9414

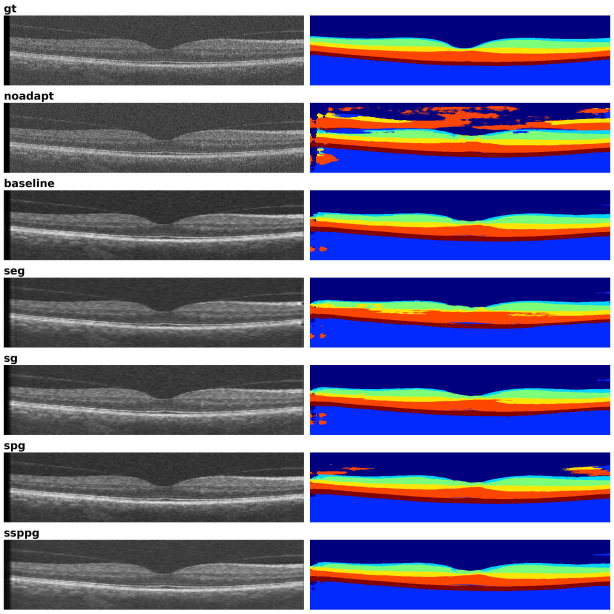


Fig. 6. Sample results of Bscan near central foveal region. Each row represents the adapted image generated by a specific model and its corresponding segmentation generated by the segmentation network.

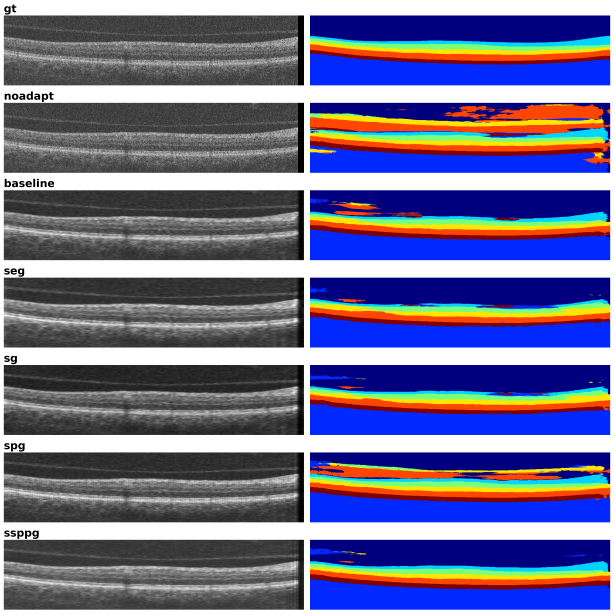


Fig. 7. Sample results of adapted images with hallucination. Each row represents the adapted image generated by a specific model and its corresponding segmentation generated by the segmentation network.

which are shown in Tables II & III. Table II shows the result of the overall mean segmentation performance for each of the fold. The segmentation performance between $baseline$ and seg models, plus sg and spg models are statistically significant by more than half of the training folds, while the seg and sg models plus spg and $ssppg$ models pairs reveals the opposite. Combined with Figure 9, Table III shows the results for each of the retinal layers that are averaged across all folds. The seg model statistically improves the segmentation performance in $Vitreous$ and IOS_BM regions compared to the $baseline$ model. seg and sg models are statistically identical. spg model has statistically better performance than sg in IPL_OPL and $Choroid$ regions, but worse in OPL_IOS, IOS_BM and $Vitreous$. $ssppg$ model outperforms spg model in IPL_OPL and OPL_IOS regions, but worse in ILM_NFL region.

Figures 6 & 7 show the successful examples of adapted images and corresponding segmentation. The first column shows the original UKB image, image without adaptation, adapted images from $baseline$, seg, sg, spg and $ssppg$, respectively. The second column shows the corresponding segmentation for each experiment set. The segmentation results are obtained directly from the model's output without post-processing steps. Figure 6 shows a example Bscan near the central foveal region. We observe a better segmentation performance near the foveal pit for $ssppg$ model compared to other models, and $ssppg$ model is more robust to produce minimal noisy segmentation. Figure 7 demonstrates the effect of possible hallucination of retinal detachment potentially due to background noise. The seg, sg, and spg models all failed to produce clean segmentation of the upper retinal layers(e.g. ILM_NFL) and Vitreous, but the $ssppg$ model mitigates such issue even near the right-hand side region, where the image frame is shifted due to motion correction. Figure 8 shows the failed circumstance where the ILM_NFL and NFL_IPL layers cannot be properly segmented by all models. It might be caused by the central shadowing artifact of the retina as well

P Models Fold	$baseline - seg$	$seg - sg$	$sg - spg$	$spg - ssppg$
1	**0.000**	**0.000**	0.006	0.478
2	0.066	0.145	**0.001**	**0.001**
3	0.908	0.224	0.908	0.283
4	**0.015**	**0.000**	**0.029**	0.835
5	**0.000**	0.423	0.124	**0.000**
6	**0.000**	0.423	0.124	**0.000**
7	0.927	**0.000**	**0.000**	**0.017**
8	0.927	**0.000**	**0.000**	**0.017**
9	**0.000**	0.839	0.258	**0.003**

TABLE II

Statistical analysis of the ablation study to evaluate the overall mean segmentation performance among different adaptation models for each of the folds. ANOVA with post-hoc Tukey HSD tests was performed for 4 pairs of adaptation models in each training fold. The significant level is set to 0.05 and the P values below it are highlighted. 6 out of 9 folds show significant statistical difference between $baseline$ and seg models, 6 out of 9 folds show no variations between seg and sg models, 5 out of 9 folds show the explicit difference between sg and spg models, 5 out of 9 folds show no statistical variations between spg and $ssppg$ models.

P Models Layer	$baseline - seg$	$seg - sg$	$sg - spg$	$spg - ssppg$
ILM_NFL	0.354	0.525	0.525	**0.000**
NFL_IPL	1.000	1.000	0.101	1.000
IPL_OPL	1.000	0.320	**0.028**	**0.003**
OPL_IOS	1.000	1.000	**0.000**	**0.000**
IOS_BM	**0.000**	1.000	**0.000**	0.068
$Vitreous$	0.251	0.251	**0.000**	**0.000**
$Choroid$	**0.009**	0.299	**0.000**	0.245

TABLE III

Statistical analysis of the ablation study to evaluate the overall segmentation performance for each of the retinal layers among different adaptation models, averaged across all the folds. ANOVA with post-hoc Tukey HSD tests was performed for 4 pairs of adaptation models in each retinal region. The significant level is set to 0.05 and the P values below it are highlighted. The difference in segmentation performance between $baseline$ and seg are significant in IOS_BM and $Choroid$. seg and sg model show no statistical difference in all retinal regions. spg model is significantly different with sg model in all regions except ILM_NFL and NFL_IPL. spg and $ssppg$ models have a significant difference in ILM_NFL, IPL_OPL, OPL_IOS and $Vitreous$.

as the noise in the Vitreous region.

We also evaluated the efficiency of the model via total training time under the same early stopping criteria. As mentioned in Section II-B2, the training stops when the validation semantic consistency loss of the UKB dataset converged. Table IV demonstrates the training time in hours for each model in each training fold. We observed that both spg and $ssppg$ models are trained significantly faster than other models, benefiting from the perceptual loss without affecting segmentation qualities.

C. Image metrics

The quantitative measure of the FID score was applied to the original and the adapted Zeiss dataset. We wanted to evaluate the image similarity of the synthesized Zeiss dataset adapted from the Topcon dataset. Due to the limited number of validation data(128 Bscans), we computed the Fréchet distance for 4 dimensionalities of feature maps from Inception V3, which are 64 from the first max pooling layer, 192 from the second max pooling layer before the Inception module A, 768 from the output of the Inception module A, and 2048 from the final average pooling layer (Maximilian Seitzer, 2020). The results are shown in Table V, the spg model and $ssppg$ model obtain the lowest and the second lowest FID scores among all adaptation models. And the FID scores of all adaptation models are significantly smaller than the one without adaptation. Table VI illustrates the evaluation using KID score. We set γ to be 2048 matching the output features of the InceptionV3 model, the bias constant c to be 1 and degree of polynomial to be 3. Results show that the $ssppg$ model has the lowest average KID score compared to other 5 adaptation models in 7 training folds, and sg model has the best KID score in the rest of 2 folds. Similar to results in Table V, all the adaptation models have much better performance than $noadapt$ model.

Discussion

As shown in Section III-A, the segmentation network achieves the state-of-art performance on all retinal layers. It sets the upper limit for the segmentation results on the adapted images. The $ssppg$ model improves the overall dice performance by an average of 46.2% with respect to no adaptation, and its performance reaches 87.4% of the first-stage segmentation model in terms of average Dice score. From Figures 9 and 10, all the adaptation models show similar performance for $Vitreous$ and $Choroid$ layers.

As shown in Figures 6-8, the segmentation network is very sensitive to the domain shifts of the images. The delineation of retinal layers are difficult since the semantic labels are allocated based on both local gradient changes and global feature distributions. The hallucination shown in Figures 7 & 8 is mainly caused by the lateral shifts and unwanted background noise of the original image, which can be induced during

Training time(hours) Model Fold	*baseline*	*seg*	*sg*	*spg*	*ssppg*
1	51.12	60.96	64.08	**52.08**	76.32
2	55.92	61.92	47.04	48.00	**35.04**
3	70.08	54.00	53.04	**25.92**	43.92
4	40.08	49.92	70.08	33.12	34.08
5	70.08	70.08	51.12	**42.96**	54.00
6	34.08	70.08	55.92	41.04	**33.12**
7	54.00	61.92	66.96	**36.00**	39.12
8	54.00	61.92	39.12	**24.00**	64.08
9	54.00	53.04	54.00	55.92	**39.12**
average	53.71	60.43	55.71	**39.89**	46.53

TABLE IV

TOTAL TRAINING TIME IN HOURS FOR EACH MODEL AMONG 9 TRAINING FOLDS. THE TIME IS CALCULATED BASED ON EARLY STOPPING CRITERIA WHERE THE VALIDATION LOSS OF TOPCON DATA SEGMENTATION CONVERGES. THE SMALLEST TRAINING TIME FOR EACH FOLD IS HIGHLIGHTED. BOTH *spg* AND *ssppg* MODELS SIGNIFICANTLY REDUCE THE TRAINING TIME COMPARED TO THE OTHER 3 MODELS.

FID Model Dim	*noadapt*	*baseline*	*seg*	*sg*	*spg*	*ssppg*
64	151304.575	23.196	23.697	23.285	**20.950**	23.639
192	345710.933	41.908	42.629	42.106	**28.957**	36.922
768	450.540	1.949	1.996	1.985	**1.670**	1.828
2048	2175.558	245.028	250.993	250.230	**198.623**	218.106

TABLE V

FID SCORE FOR EACH DIMENSIONALITY OF THE FEATURE MAP. THE LOWEST FID SCORES AMONG ALL MODELS ARE HIGHLIGHTED, WHICH INDICATES BETTER IMAGE SIMILARITIES TO THE ORIGINAL IMAGE. THE SCORES FOR EACH MODEL IS AVERAGED AMONG ALL 9 TRAINING FOLDS EXCEPT FOR *noadapt*. THE *spg* AND *ssppg* MODELS HAVE THE LOWEST AND SECOND LOWEST FID SCORES IN ALL DIMENSIONALITIES.

KID Model Fold	*noadapt*	*baseline*	*seg*	*sg*	*spg*	*ssppg*
1	0.9659 ±0.0375	0.5559 ±0.2082	0.0456 ±0.0614	**0.0060 ±0.0174**	0.0234 ±0.0383	0.0088 ±0.0289
2	0.9659 ±0.0375	0.3541 ±0.2041	0.0429 ±0.0586	0.0082 ±0.0214	0.0147 ±0.0294	**0.0052 ±0.0189**
3	0.9659 ±0.0375	0.2505 ±0.1716	0.0498 ±0.0676	0.0069 ±0.0219	0.0121 ±0.0270	**0.0052 ±0.0105**
4	0.9659 ±0.0375	0.1876 ±0.1691	0.0278 ±0.0495	**0.0056 ±0.0154**	0.0164 ±0.0376	0.0082 ±0.0218
5	0.9659 ±0.0375	0.1244 ±0.1208	0.0220 ±0.0386	0.0035 ±0.0136	0.0109 ±0.0273	**0.0027 ±0.0125**
6	0.9659 ±0.0375	0.1134 ±0.1165	0.0369 ±0.0529	0.0053 ±0.0180	0.0086 ±0.0200	**0.0048 ±0.0146**
7	0.9659 ±0.0375	0.0750 ±0.0936	0.0246 ±0.0473	0.0048 ±0.0146	0.0180 ±0.0346	**0.0046 ±0.0189**
8	0.9659 ±0.0375	0.0704 ±0.0903	0.0270 ±0.0501	0.0042 ±0.0147	0.0143 ±0.0347	**0.0038 ±0.0137**
9	0.9659 ±0.0375	0.0584 ±0.0766	0.0231 ±0.0457	0.0064 ±0.0249	0.0111 ±0.0242	**0.0041 ±0.0111**
Average	0.9659 ±0.0375	0.1989 ±0.1390	0.0333 ±0.0524	0.0057 ±0.0180	0.0144 ±0.0304	**0.0053 ±0.0167**

TABLE VI

KID SCORE OF 5 MODELS WITH ALL 9 FOLDS. THE FINAL ROW SHOWS THE AVERAGE KID SCORE AMONG ALL FOLDS. THE SMALLEST KID SCORE FOR EACH FOLD IS HIGHLIGHTED. THE SCORE FOR *noadapt* IS FIXED FOR ALL FOLDS SINCE THERE IS NO VARIATION AMONG DIFFERENT FOLDS. THE *ssppg* MODEL HAS THE SMALLEST AVERAGED MEAN AND STANDARD DEVIATION OF 0.0053 ±0.0167 COMPARED TO THE OTHER MODELS

data acquisition or pre-processing steps like speckle noise due to scattering and interference of coherent light and motion corrections. For the *seg* model, the undefined shifted region may be mistakenly identified as a topological feature of the retina. The gradient loss significantly mitigates such a problem by enforcing the learning of the "true" local boundaries, which act as a more generalized semantic constraint using the gradient map. However, as mentioned in Section I, the GAN network usually suffers from long training time and complicated parameter tuning. As shown in Table IV, both *spg* and *ssppg* models converge significantly faster than the other models due to the perceptual loss. Unlike the traditional style-transfer networks, the perceptual loss is calculated using a pre-trained classification network(e.g. VGG16) with weights frozen, which significantly speeds up the training process while maintaining the high-level texture of the original images. However, the classification network pre-trained on ImageNet may not precisely extract high-level features from unseen images like retinal OCT. The differentiation of the adjacent retinal layers relies more on pixel-level details than global features. Such noise may represent propagation of the high-level deviations. Therefore, both SSIM and PSNR criterion contribute to noise reduction throughout the image. The discrepancy of two intensity distributions can be reduced via matching both the noise distributions and the pixel information. Even though the time cost increases compared to the *spg* model, the improvement of the segmentation performance is worth the trade-off.

While the segmentation of the retinal OCT is our main interest, we also investigated the relationship between the segmentation performance and the image quality characteristics. Specifically, the adapted images can be re-usable if the segmentation-favoured domain is close to the actual target domain. As illustrated in Section III-C, the images adapted by *spg* and *ssppg* models have lower FID and KID scores than the adaptation models. Both the FID and KID scores tend to decrease along with the complexities of the loss functions. It inferred that the adapted images with better segmentation can

Dice \ Layer / Model	ILM_NFL	NFL_IPL	IPL_OPL	OPL_IOS	IOS_BM	$Vitreous$	$Choroid$	Average
$ssppg\_adapt$	0.5065	0.6177	0.6694	0.5237	0.7049	0.9082	0.7350	0.7476
$Direct$	0.9597	0.9652	0.9309	0.9171	0.9375	0.9987	0.9997	0.9560

TABLE VII

DICE SCORE FOR TWO EXTRA EXPERIMENTS. $ssppg\_adapt$ INDICATES THE GAN-BASED TRANSFER LEARNING USING ONLY ADAPTED UKB IMAGES. $Direct$ REPRESENTS THE DIRECT TRAINING USING ORIGINAL UKB IMAGE AND LABEL PAIRS. THE AVERAGE DICE SCORE IS CALCULATED FOR 7 RETINAL REGIONS. THE $ssppg\_adapt$ MODEL DOES NOT PRODUCE COMPARABLE RESULTS TO OUR ADAPTATION MODELS. THE $Direct$ MODEL HAS BETTER PERFORMANCE, WHICH USES THE TRUE IMAGE AND LABEL PAIRS.

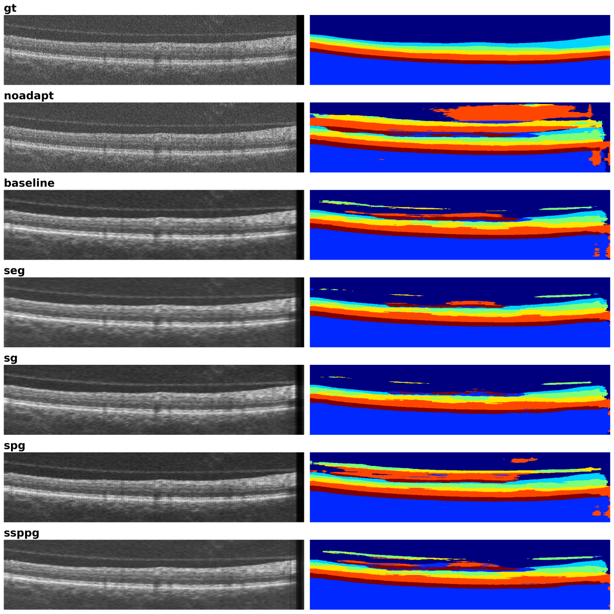


Fig. 8. Sample results of failed case where the upper retinal layers cannot be delineated. Each row represents the adapted image generated by a specific model and its corresponding segmentation generated by the segmentation network.

lead to higher similarities of image domains, i.e. the segmentation model can produce higher quality semantic labels from a latent space similar to the original image distribution.

We also performed a follow-up experiment using the best-performance $ssppg$ model to re-train the segmentation network with only the adapted images, which can be regarded as a GAN-based transfer learning approach. We used the same training settings as mentioned in Section II-B1, but both the training and validation data were the adapted UKB images, i.e. the synthesized UKB images adapted from the Zeiss dataset. The segmentation performance was evaluated by directly feeding the original UKB testing dataset into the re-trained segmentation network. The number of training data is vastly increased compared to the ones used in the adaptation network, but the segmentation performance is not comparable to any of the previous models. The results are shown in Table VII named $ssppg\_adapt$. It revealed that the extra transfer learning can worsen the segmentation performance, and the adapted images can still be distinguishable from the original images. A possible remedy for this experiment would be including the original UKB images along with its pseudo-labels generated by its adapted images adapted from the Zeiss dataset, thus the network can learn from the original UKB data while the pseudo-labels can also contribute to the model training. Nevertheless, the dice scores may be more suitable to amplify the differences than the commonly used image quality metrics, and it is more of clinical usage in various analytical applications.

IV. Limitation and Future Work

There are several limitations to our proposed two-stage pipeline. The corresponding remedies and possible future work are elaborated on in this section.

First, We did not quantitatively evaluate the performance of the fine-tuning strategies. We experienced high failure rates of training when intensity-level transformations were involved, such as noise injection, interpolations due to deformation, modifications of luminance, contrast, etc. Similar issues occurred when we used optimizers with momenta like Adam and SGD. The choices of proper learning rate was crucial for a GAN-based network, especially the integration of several losses were involved. Multiple researches showed that lower learning rates shall always be considered, but the network can still fail to converge even with smaller learning rates and longer training time (Yazıcı et al., 2019). The TTUR concept was rigorously evaluated by Heusel et al., which benefited more practically for testing and learning rate tuning. The soft labels were proposed in WassersteinGAN, which also benefited the LeakyReLU activation in the discriminator module. We experienced a relatively faster convergence for the baseline model using soft labels, but the segmentation performance were roughly unchanged. As all adaptation models were tested using the same strategies, it will not interfere with the final evaluation results. We also performed several experiments adjusting the weighting parameters(λ) of the loss functions both statically and dynamically. Due to the constraints of the resources, we tried few fixed combinations of parameters, also linearly increasing the weight of the certain loss functions as training carries out. Unfortunately, these adjustment did not lead to better performance or even causes structural collapses of the images. Systematic parameter optimization could potentially be performed in future studies like grid search to obtain the optimal settings.

The Topcon dataset only contain OCT images from healthy patients, where the segmentation model was trained the Zeiss dataset with several pathological cases. To avoid false positives of fluid, we ignored the fluid channel from the output of the Softmax activation layer. An extended study can be performed mainly on pathological data, where the performance of fluid segmentation can be evaluated separately. Due to resource constraints, we did not perform experiments on the public external

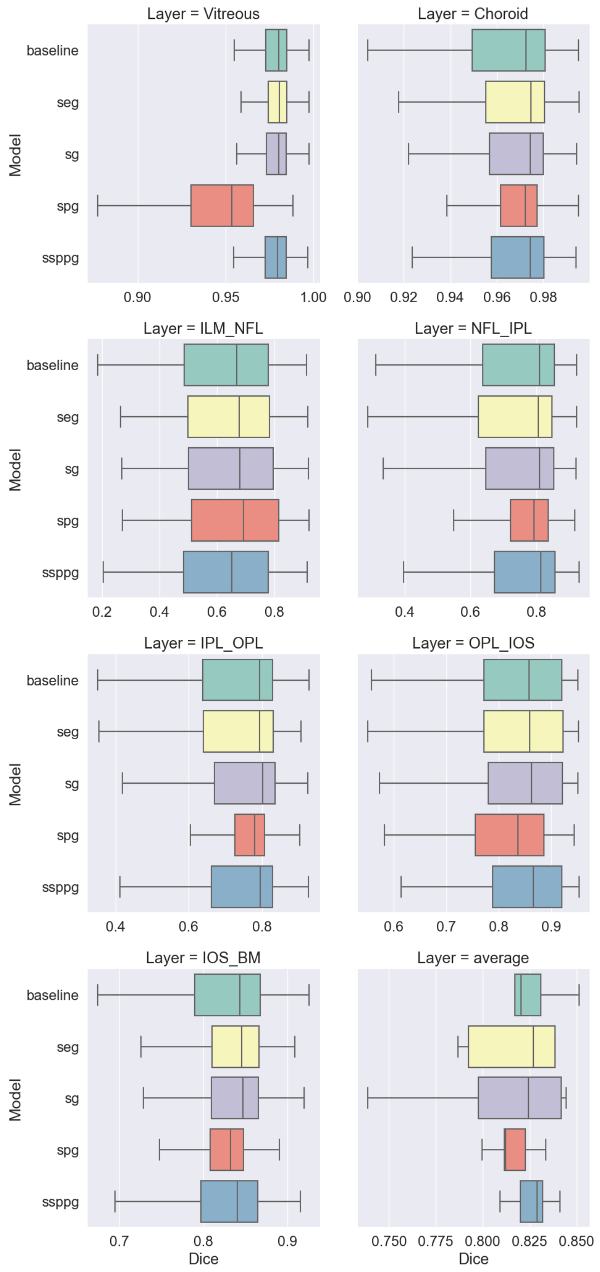


Fig. 9. Box plots of Dice score for 7 retinal regions averaged among 9 training folds. Each subplot represents the distribution of dice scores for each retinal region for all models, and the last subplot represents the averaged dice scores among all training folds. The median dice are marked within each box, and the whisker is calculated as 1.5IQR(Interquartile Range). The *sspg* model has the best average dice score among all models of 0.8231.

dataset, which required a decent number of image and label pairs for both training and validation. However, the structural appearance of SD-OCT images acquired from various devices is visually similar. The successful adaptation with a limited number of UKB images proves the robustness and adaptability of our model. Additionally, the number of retinal layers can be further enriched, which will apply better constraints to the adaptation. Empirically, more semantic constraints shall further improve the performance of the network, providing that our segmentation network is trained successfully.

We investigated the segmentation performance using only Topcon dataset with ground-truth labels used for evaluation of our GAN network. Table VII shows the training results named $Direct$. It outperforms the $sspg$ model as expected. It directly uses the true UKB data and labels. However, our proposed method aimed to generate segmentation when the data in the target domain has no ground-truth labels available. Without such a prerequisite, an extra experiment can be delivered that we can add segmentation constraints using UKB labels too, which shall produce considerably better results compared to the current $sspg$ model.

The segmentation performance was evaluated among raw outputs of different models. A series of post-processing methods can further improve the performance of the models, such as keeping the largest connected components of the labels, hole fillings, proper erosion and dilation, etc. However, such labels still need to be manually corrected for analytical measurements. The comparison of the raw inference results can better reflect the performance of the models.

V. Conclusion

In this study, we proposed a novel two-stage segmentation-guided domain adaptation network based on CycleGAN architecture to achieve effective data harmonization for multi-site OCT data. The proposed approach significantly improves the segmentation results for images of different domains. Once the adaptor module is trained for specific OCT acquisition devices, the preliminary results of good qualities can be produced in real-time to speed up the process of manual corrections. Future work shall be focused on the adaptation of pathological OCT images where the presence of fluid causes explicit deformation or even destruction of the retinal layers. Better performance can be obtained with a follow-up of post-processing steps.

VI. Funding and Acknowledgement

This study was funded by the National Sciences and Engineering Research Council of Canada; Canadian Institutes of Health Research, Compute Canada, Wake Forest University School of Medicine Startup Funding, and Precision Imaging Beacon, University of Nottingham. The sponsor or funding organization had no role in the design or conduct of this research.

VII. Declaration of competing interest

There is no conflict of interest declared from all co-authors

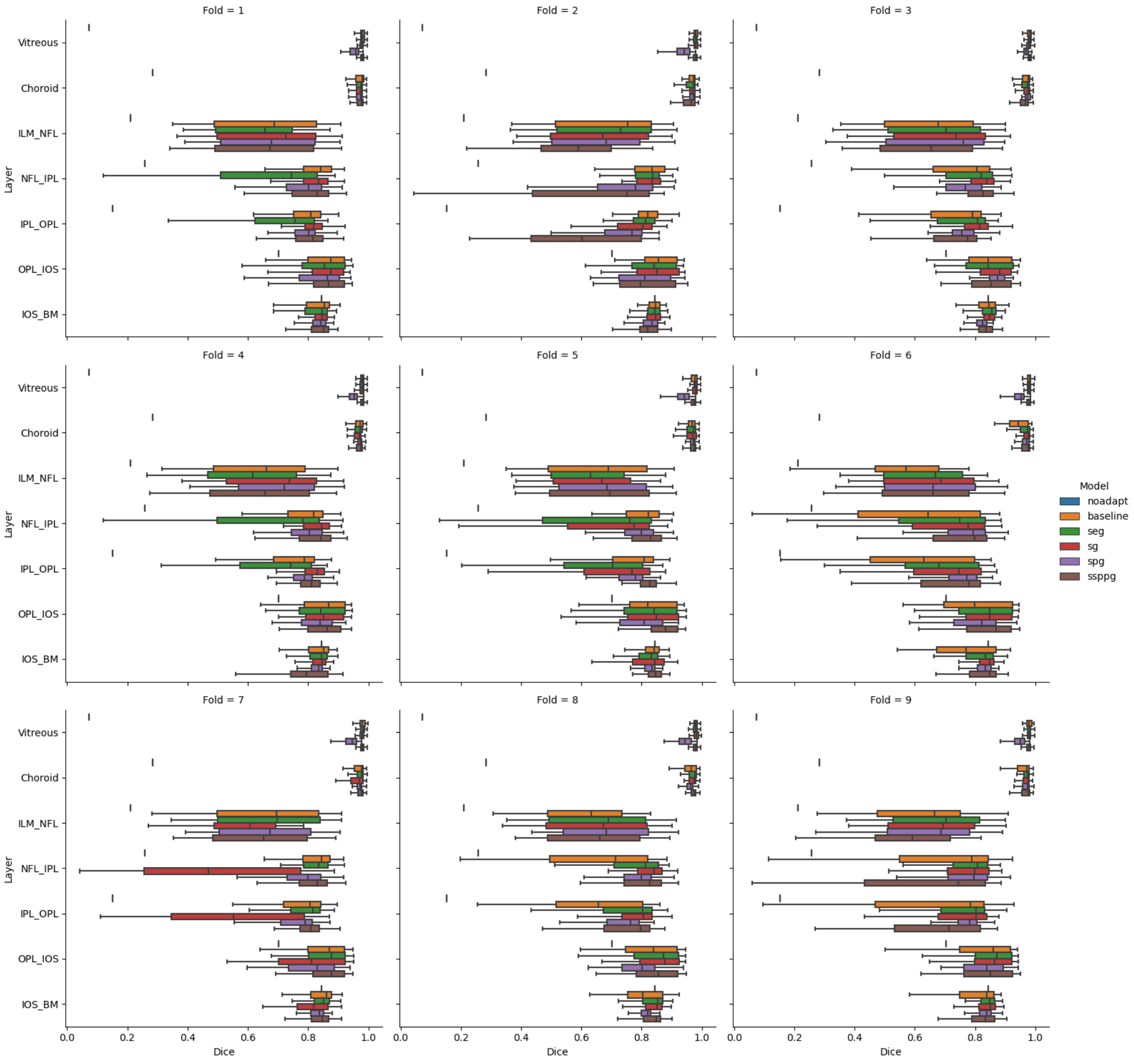


Fig. 10. Box plots of Dice score for 7 retinal regions for each training fold. The dice score of *noadapt* model is constant for all folds shown as a vertical line in each subplot. The median dice are marked within each box, and the whisker is calculated as 1.5IQR(Interquartile Range). The *ssppg* model has the best dice score in folds 5 & 6 of 0.8413 and 0.8093, respectively. The *sg* model has the best dice score in folds 1,3,4, and 8 of 0.8438, 0.8419, 0.8446 and 0.8312, respectively. The *seg* model has the best dice score in folds 7 and 9 of 0.8385 and 0.8300, respectively. The *baseline* model has the best dice score in fold 2 of 0.8515.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. 2017. URL http://arxiv.org/abs/1701.07875.

M. Bashkansky and J. Reintjes. Statistics and reduction of speckle in optical coherence tomography. *Optic letters*, 25: 545–547, 2000.

Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans Peter Kriegel, Bernhard Schölkopf, and Alex J. Smola. Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):1–10, 2006. ISSN 13674803. doi: 10.1093/bioinformatics/btl242.

Tong Che, Yanran Li, Ruixiang Zhang, R Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. Maximum-Likelihood Augmented Discrete Generative Adversarial Networks. 2017. URL http://arxiv.org/abs/1702.07983.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. DoveNet: Deep Image Harmonization via Domain Verification. 11 2019.

Nicola K. Dinsdale, Mark Jenkinson, and Ana I.L. Namburete. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*, 228: 117689, March 2021. ISSN 10538119. doi: 10.1016/j.neuroimage.2020.117689. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811920311745.

D C Dowson and B V Landau. The Frkhet Distance between Multivariate Normal Distributions. Technical report, 1982.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2014. ISSN

15577317. doi: 10.1145/3422622.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. ISSN 15324435.

Yufan He, Aaron Carass, Yihao Liu, Shiv Saidha, Peter A. Calabresi, and Jerry L. Prince. Adversarial domain adaptation for multi-device retinal OCT segmentation. 1131309(March 2020):7, 2020a. ISSN 16057422. doi: 10.1117/12.2549839.

Yufan He, Aaron Carass, Lianrui Zuo, Blake E. Dewey, and Jerry L. Prince. *Self domain adapted network*, volume 12261 LNCS. Springer International Publishing, 2020b. ISBN 9783030597092. doi: 10.1007/978-3-030-59710-8{_}43. URL http://dx.doi.org/10.1007/978-3-030-59710-8_43.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):6627–6638, 2017. ISSN 10495258.

Yuta Hiasa, Yoshito Otake, Masaki Takao, Takumi Matsuoka, Kazuma Takashima, Aaron Carass, Jerry L. Prince, Nobuhiko Sugano, Sato, and Yoshinobu. *Cross-Modality Image Synthesis from Unpaired Data Using CycleGAN Effects of Gradient Consistency Loss and Training Data Size*, volume 11037. Springer International Publishing, 2018. ISBN 9783030005351. doi: 10.1007/978-3-030-00536-8. URL http://dx.doi.org/10.1007/978-3-030-00536-8_4.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. CyCADA: Cycle-Consistent Adversarial Domain adaptation. *35th International Conference on Machine Learning, ICML 2018*, 5:3162–3174, 2018.

Norihiko Ikeda and Stephen Lam. *Optical coherence tomography*. 2013. ISBN 9781461460091. doi: 10.1007/978-1-4614-6009-1{_}15.

Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:5967–5976, 2017. doi: 10.1109/CVPR.2017.632.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Style Transfer and SR. *Eccv*, pages 1–5, 2016. ISSN 0302-9743. URL https://cs.stanford.edu/people/jcjohns/papers/eccv16/JohnsonECCV16.pdf%0Ahttp://arxiv.org/abs/1603.08155.

Sieun Lee, Morgan L. Heisler, Karteek Popuri, Nicolas Charon, Benjamin Charlier, Alain Trouvé, Paul J. Mackenzie, Marinko V. Sarunic, and Mirza Faisal Beg. Age and Glaucoma-Related Characteristics in Retinal Nerve Fiber Layer and Choroid: Localized Morphometrics and Visualization Using Functional Shapes Registration. *Frontiers in Neuroscience*, 11, 2017. ISSN 1662-453X. URL https://www.frontiersin.org/articles/10.3389/fnins.2017.00381.

Peilun Li, Xiaodan Liang, Daoyuan Jia, and Eric P. Xing. Semantic-aware grad-GAN for virtual-to-real urban scene adaption. *British Machine Vision Conference 2018, BMVC 2018*, 2019a.

Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:6929–6938, 2019b. ISSN 10636919. doi: 10.1109/CVPR.2019.00710.

Julian Lo, Timothy T. Yu, Da Ma, Pengxiao Zang, Julia P. Owen, Qinqin Zhang, Ruikang K. Wang, Mirza Faisal Beg, Aaron Y. Lee, Yali Jia, and Marinko V. Sarunic. Federated Learning for Microvasculature Segmentation and Diabetic Retinopathy Classification of OCT Data. *Ophthalmology Science*, 1(4):100069, December 2021. ISSN 2666-9145. doi: 10.1016/j.xops.2021.100069. URL https://www.sciencedirect.com/science/article/pii/S2666914521000671.

Da Ma, Donghuan Lu, Shuo Chen, Morgan Heisler, Setareh Dabiri, Sieun Lee, Hyunwoo Lee, Gavin Weiguang Ding, Marinko V. Sarunic, and Mirza Faisal Beg. LF-UNet – A novel anatomical-aware dual-branch cascaded deep neural network for segmentation of retinal layers and fluid from optical coherence tomography images. *Computerized Medical Imaging and Graphics*, 94:101988, 12 2021. ISSN 0895-6111. doi: 10.1016/J.COMPMEDIMAG.2021.101988.

Da Ma, Meenakshi Kumar, Vikas Khetan, Parveen Sen, Muna Bhende, Shuo Chen, Timothy T L Yu, Sieun Lee, Eduardo V Navajas, Joanne A Matsubara, Myeong Jin Ju, Marinko V Sarunic, Rajiv Raman, and Mirza Faisal Beg. Clinical explainable differential diagnosis of polypoidal choroidal vasculopathy and age-related macular degeneration using deep learning. *Computers in biology and medicine*, 143:105319, 2 2022. ISSN 1879-0534. doi: 10.1016/j.compbiomed.2022.105319. URL http://www.ncbi.nlm.nih.gov/pubmed/35220077.

Maximilian Seitzer. pytorch-fid: FID Score for PyTorch, 8 2020. URL https://github.com/mseitzer/pytorch-fid.

Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to Image Translation for Domain Adaptation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00473.

M. Pekala, N. Joshi, T. Y.Alvin Liu, N. M. Bressler, D. Cabrera DeBuc, and P. Burlina. Deep learning based retinal OCT segmentation. *Computers in Biology and Medicine*, 114, 2019. ISSN 18790534. doi: 10.1016/j.compbiomed.2019.103445.

Aimon Rahman, M. Sohel Rahman, and M. R.C. Mahdy. 3C-GAN: Class-consistent CycleGAN for malaria domain adaptation model. *Biomedical Physics and Engineering Express*, 7(5), 2021. ISSN 20571976. doi: 10.1088/2057-1976/ac0e74.

R. Robinson, Q. Dou, D. C. Castro, K. Kamnitsas, M. de Groot, R. M. Summers, D. Rueckert, and B. Glocker. Image-level Harmonization of Multi-Site Data using Image-and-Spatial Transformer Networks. 6 2020. doi: 10.1007/978-3-030-59728-3{_}69.

Abhijit Guha Roy, Sailesh Conjeti, Sri Phani Krishna Karri, Debdoot Sheet, Amin Katouzian, Christian Wachinger, and Nassir Navab. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical optics express*, 8(8): 3627–3642, 8 2017. ISSN 2156-7085. doi: 10.1364/BOE.8.003627.

Philipp Seebock, David Romo-Bucheli, Sebastian Waldstein, Hrvoje Bogunovic, Jose Ignacio Orlando, Bianca S. Gerendas, Georg Langs, and Ursula Schmidt-Erfurth. Using cyclegans for effectively reducing image variability across OCT devices and improving retinal fluid segmentation. *Proceedings - International Symposium on Biomedical Imaging*, 2019-April:605–609, 2019. ISSN 19458452. doi: 10.1109/ISBI.2019.8759158.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 9 2014.

Sir Rory Collins. UK Biobank. URL https://www.ukbiobank.ac.uk/.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, and Jonathon Shlens. Rethinking the Inception Architecture for Computer Vision. Technical report.

Dezheng Tian, Zilong Zeng, Xiaoyi Sun, Qiqi Tong, Huanjie Li, Hongjian He, Jia-Hong Gao, Yong He, and Mingrui Xia. A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset. *NeuroImage*, 257:119297, 8 2022. ISSN 10538119. doi: 10.1016/j.neuroimage.2022.119297.

Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised Domain Adaptation in Semantic Segmentation: A Review. *Technologies*, 8(2):35, 2020. doi: 10.3390/technologies8020035.

Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. ISSN 10577149. doi: 10.1109/TIP.2003.819861.

William Falcon. PyTorch Lightning. URL https://www.pytorchlightning.ai/.

Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:945–954, 2017. doi: 10.1109/CVPR.2017.107.

Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9008–9017, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00903.

Yasin Yazıcı, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, and Vijay Chandrasekhar. THE UNUSUAL EFFECTIVENESS OF AVERAGING IN GAN TRAINING. *ICLR 2019*, 2019.

Guodong Zeng, Till D Lerch, Florian Schmaranzer, Guoyan Zheng B, and Nicolas Gerber. Domain Adaptation for Cross-Modality. *Miccai*, 4:201–210, 2021. doi: 10.1007/978-3-030-87199-4.

Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob: 2242–2251, 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.244.