A Vessel-Segmentation-Based CycleGAN for Unpaired Multi-modal Retinal Image Synthesis

Aline Sindel, Andreas Maier, Vincent Christlein

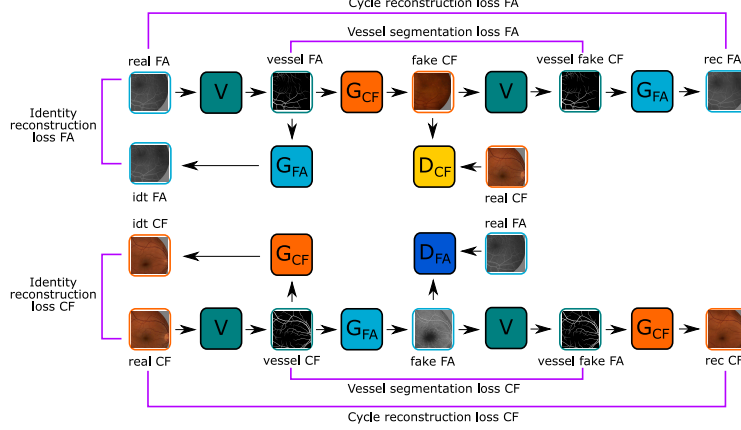Pattern Recognition Lab, FAU Erlangen-Nürnberg
aline.sindel@fau.de

Abstract. Unpaired image-to-image translation of retinal images can efficiently increase the training dataset for deep-learning-based multi-modal retinal registration methods. Our method integrates a vessel segmentation network into the image-to-image translation task by extending the CycleGAN framework. The segmentation network is inserted prior to a UNet vision transformer generator network and serves as a shared representation between both domains. We reformulate the original identity loss to learn the direct mapping between the vessel segmentation and the real image. Additionally, we add a segmentation loss term to ensure shared vessel locations between fake and real images. In the experiments, our method shows a visually realistic look and preserves the vessel structures, which is a prerequisite for generating multi-modal training data for image registration.

1 Introduction

Recent deep learning methods for multi-modal medical image registration require a large amount of training data. Since it is difficult and tedious to obtain precise ground truth from real data, image-to-image translation methods are effective means to synthetically augment multi-modal datasets. In ophthalmology, different imaging systems, such as color fundus (CF), fluorescein angiography (FA), and optical coherence tomography angiography (OCTA) are used for the diagnosis of retinal diseases. Our aim is to generate synthetic multi-modal pairs that can be used to train registration methods in a self-supervised manner. For that the position and shape of the vessels should be preserved by the image-to-image translation method, but the texture and style should be transferred to the other modality. Here, we concentrate on the image-to-image translation between CF and FA images. In CF the vessels are depicted in dark and in FA in light, but by both modalities the fovea is depicted in dark and the optic cup and disc are depicted in light, which needs to be considered by the translation methods.

Conditional generative adversarial networks (cGANs) were explored in literature for the image-to-image translation of CF and FA images. In case of aligned multi-modal images, Pix2Pix [1] based approaches can be used to learn a direct 1-to-1 mapping between both modalities. In this regard, VTGAN [2] is

Fig. 1. Our unpaired image-to-image translation method based on the retinal vessel segmentation in multi-modal fundus images.

introduced for closely but not perfectly aligned CF-FA image pairs, which uses a coarse and a fine generator with attention blocks and a vision transformer as discriminator network. CycleGAN [3] based approaches learn a direct image-to-image translation for unpaired images. For the CF-FA translation task, Li et al. [4] enriches the CycleGAN with structure and appearance encoder networks which are inserted prior to the generator networks and Cai et al. [5] extends the CycleGAN with multi-scale generator and discriminator networks and a quality-aware loss at feature level. In contrast, we extend the CycleGAN by including the vessel segmentation as a shared representation between both domains. There exist a bunch of approaches that generate CF images from extracted vessel segmentations. For instance, the cGAN by Liang et al. [6] adds a class feature loss for diabetic retinopathy grading and a retinal detail loss which is a combination of the reconstruction loss between real and fake image and perceptual loss using a specific layer from the pretrained VGG-19 network. Niu et al. [7] includes pathology specific descriptors into the cGAN to generate CF images with specific pathological features. The real and synthetic images are compared using perceptual and severity losses. By integrating the vessel segmentation into the CycleGAN, we tackle to reduce the domain gap of the vessels between both modalities.

In this paper, we propose VesselCycleGAN, a cGAN based approach for unpaired retinal image-to-image translation based on the vessel segmentation of CF and FA images using cycle consistency. We extend the CycleGAN pipeline by inserting a vessel segmentation UNet before the generator network, which we equip with a UNet vision transformer [8]. With the vessel segmentation network, we modify the identity loss to learn the translation from the vessel segmentation to the real image and we add a segmentation loss to preserve the same vessel structures in the real and generated images and apply our method to two datasets.

2 Materials and methods

2.1 VesselCycleGAN for unpaired retinal image-to-image translation

We incorporate a vessel segmentation network (V) into the CycleGAN [3] framework, as shown in Figure 1, which we place in front of the generator networks, such that those do not learn the direct mapping between the two domains, but the mapping from the vessel segmentation to the particular domain. Cycle-GANs are conditional generative adversarial networks, consisting of two generator $G_{A/B}$ and two discriminator networks $D_{A/B}$, that learn the image-to-image translation between unpaired images from two domains A/B by using adversarial loss, cycle-consistency and identity-consistency losses [3]. The cycle-consistency loss $\mathcal{L}_{\text{cycA}}$ minimizes the difference between the real image A and its reconstruction $\hat{A}$ after passing through the cycle of applying both G_B and G_A, and here in our case V, G_B, V, and G_A:

$$\mathcal{L}_{\text{cycA}}(A) = \lambda_A||A - G_A(V(G_B(V(A))))||_1. \tag{1}$$

Using the identity-consistency loss $\mathcal{L}_{\text{idtA}}$, G_A originally learns the identity mapping of $G_A(A) \stackrel{!}{=} A$, however, in our modified CycleGAN this becomes:

$$\mathcal{L}_{\text{idtA}}(A) = \lambda_A \lambda_{idt}||A - G_A(V(A))||_1, \tag{2}$$

where G_A learns the aligned one-way image translation task of vessel segmentation A to real A. Additionally, we compute the Dice loss between the segmentation of the real A and fake B. The role of $D_{A/B}$ is as in the normal CycleGAN to distinguish unpaired real images A/B from fake images $\tilde{A}/\tilde{B}$ generated using $G_A(V(B))$ or $G_B(V(A))$. As generator network we employ the UNet-ViT [8] which is a four layer UNet ($D = 48$) with a pixel-wise vision transformer bottleneck (with 12 transform encoder blocks). As discriminator network we use the PatchGAN [1] and for the vessel segmentation network a four layer UNet ($D = 16$), which we pretrained for the vessel segmentation task.

2.2 Retinal datasets

We train and test our synthesis method using the CF-FA dataset [9] which consists of 59 pairs of color fundus (CF, 576×720) and fluorescein angiography (FA, 576×720) images from controls (29 image pairs) and from patients with diabetic retinopathy (30 pairs). We split the image pairs into train: 35, val: 10, and test: 14, with equally distributed healthy and non-healthy eyes. For each training image, we extract nine 512×512 patches, which are randomly cropped to 448×448. For the validation and test set, we directly extract up to nine 448×448 patches for each image. This results per modality into train: 315, val: 90 and test: 120 image patches. Since the original image pairs are not aligned, we register the images of the test set (prior to patch extraction) using KPVSA-Net [10].

Secondly, we use the HR fundus dataset [11] which consists of CF images and manual vessel segmentations. To train the vessel segmentation UNet, we use the color and green channel of the fundus images in two resolutions: the center 768×768 region and a downsized version by a factor of 4 to have a similar image size as the CF-FA dataset. During training, we randomly extract 512×512 patches on the fly from the 108 training and 36 validation images. For the image translation task, we use 81 484×484 patches from the CF images of the test set without the ground truth vessel segmentations.
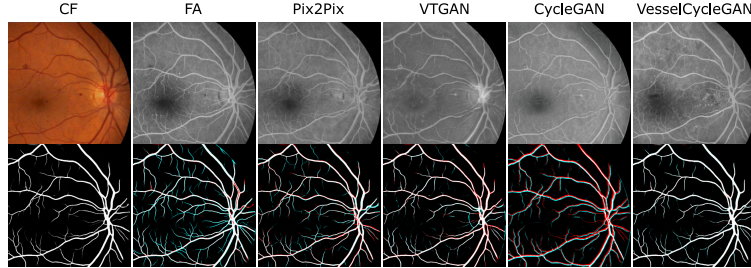

Fig. 2. FA synthesis results and vessel segmentation overlays (vessels segmented by our UNet).
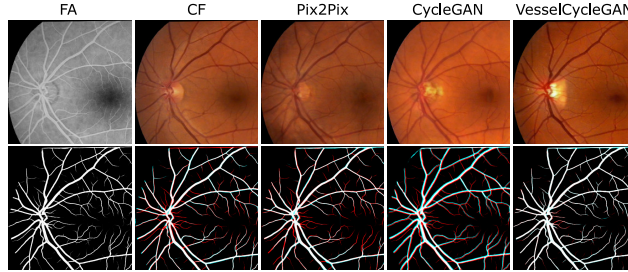

Fig. 3. CF synthesis results and vessel segmentation overlays (vessels segmented by our UNet).
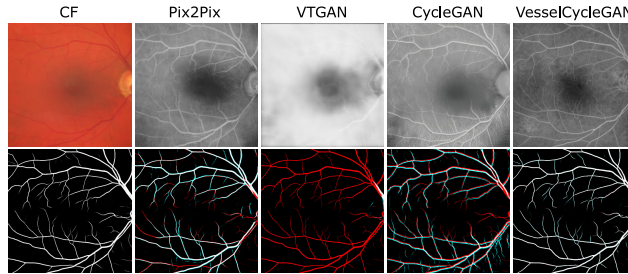

Fig. 4. FA synthesis results of the HRF dataset (vessels segmented by our UNet).

Table 1. Quantitative evaluation for FA and CF synthesis using the CF-FA test images. LPIPS and KID are computed between fake B and registered real B; Dice between the vessels of fake B and real A. Methods with * were trained using registered images.

A-to-B	CF-to-FA			FA-to-CF		
Metrics	LPIPS ↓	KID ↓	Dice ↑	LPIPS ↓	KID ↓	Dice ↑
Pix2Pix* (ResNet9)	*0.3478*	*0.0104*	0.7615	0.3371	0.0227	0.7264
VTGAN*	0.3731	0.0383	0.8035	-	-	-
CycleGAN (ResNet9)	0.4296	0.0388	0.2471	0.3511	0.0182	0.3550
VesselCycleGAN (ResNet9 w/o seg)	0.3893	0.0322	0.8798	0.3507	0.0211	0.8458
VesselCycleGAN (w/o seg)	0.3652	0.0158	0.8471	0.3503	0.0227	0.8567
VesselCycleGAN	0.3685	0.0163	*0.9534*	*0.3329*	*0.0148*	*0.9419*

Table 2. Quantitative evaluation for FA synthesis for the HRF test images using the models trained on the CF-FA dataset. For KID, real FAs are from the CF-FA dataset, since there are none in the HRF dataset. Methods with * are trained using registered CF-FA images.

Metrics	Pix2Pix*	VTGAN*	CycleGAN	VesselCycleGAN
KID ↓ (unaligned fake - real FA)	*0.0295*	0.2010	0.0580	0.0515
Dice ↑ (fake FA - real CF)	0.6943	0.3211	0.2502	*0.9439*

2.3 Experimental details

Prior to the training of our retinal synthesis network, we train the vessel segmentation UNet on the augmented HR fundus dataset by using equally weighted binary cross-entropy and Dice loss for 800 epochs with early stopping, Adam optimizer, a learning rate $\eta = 2 \cdot 10^{-4}$, linear decay of η after 50 epochs, and batch size of 2. Then, we train the generator and discriminator networks of our retina synthesis GAN using Adam solver with a learning rate of $\eta = 2 \cdot 10^{-4}$ for 600 epochs with early stopping, a batch size of 1, $\lambda_{A/B} = 100$, $\lambda_{idt} = 1$, $\lambda_{seg} = 1$. For both tasks, we use online data augmentation (color jittering, horizontal flipping, rotation, and cropping). We compare our method with CycleGAN [3] and Pix2Pix [1] (both: G using ResNet encoder with 9 blocks with instance normalization), and with the VTGAN [2]. Pix2Pix and VTGAN require aligned training data, hence we registered our training and validation image pairs using KPVSA-Net [10]. For our method and CycleGAN, we use unaligned data with randomly sampled patches within the same class (healthy/unhealthy).

3 Results

Table 1 summarizes the quantitative results for the CF-to-FA and FA-to-CF task using similarity (LPIPS, KID) and Dice metrics. For the FA synthesis, Pix2Pix obtained the best LPIPS and KID scores, but has only a relatively low Dice

score of 0.76. Our VesselCycleGAN achieves a bit lower similarity scores, but the highest Dice score of 0.95 and is superior to VTGAN and the default CycleGAN. For the CF synthesis task, our VessselCycleGAN achieves the best LPIPS, KID, and Dice scores. Pix2Pix here only achieves the second best LPIPS score. For both synthesis tasks, the Dice scores of CycleGAN are very low, indicating that the vessel positions have not been preserved in the synthetic image. Moreover, we tested different settings for our VesselCycleGAN to show the advantage of adding the segmentation loss (gain of up to 10 % Dice score) and by using the UNet-ViT instead of the ResNet9 generator network. The qualitative results in Figure 2 and 3 reflect the findings. The structures in the synthetic images generated by VesselCycleGAN and Pix2Pix demonstrate visual similar structures to the real images. In the vessel segmentation overlays, deviating vessels between the content and generated image are marked in red (missed vessels) and cyan (added vessels). Our VesselCycleGAN has the highest overlap in the vessel structure with the content image. Pix2Pix and VTGAN show some small misalignment in the vessel details, as they learn a direct mapping between the domains from the registered data, which can show some small deformations in the vessel structure. Further, we tested the transferability of the trained models for CF images of the HRF dataset, where no ground truth FA images exist. In Figure 4, the synthetic image of VesselCycleGAN depicts the fine structures, the result of Pix2Pix is a bit blurry, very blurry for CycleGAN while VTGAN was not able to obtain a realistic result. Numerically, the KID between the FA images from the CF-FA dataset and the generated FA images of the CF HRF images in Table 2, was best for Pix2Pix and second best for VesselCycleGAN. The Dice score for VesselCycleGAN is close the CF-FA dataset, while the competing methods have lower or very low results.

4   Discussion

We relax the unpaired image-to-image translation of multi-modal fundus images to direct mappings from the vessel segmentation to the other modality within the CycleGAN pipeline. Our method, which is trained with unpaired images, learns the modality specific visual patterns and preserves the vessel locations, and thus can be used to augment training data for multi-modal retinal registration methods. As future work, our vessel-based approach could be extended by also including optical disc segmentations and further methods to control the synthesis of pathological structures could be investigated.

References

1. Isola P, Zhu JY, Zhou T, et al. Image-To-Image Translation With Conditional Adversarial Networks. Proc IEEE CVPR 2017. 2017;.
2. Kamran SA, Hossain KF, Tavakkoli A, et al. VTGAN: Semi-supervised Retinal Image Synthesis and Disease Prediction using Vision Transformers. 2021 IEEE/CVF ICCVW. 2021; p. 3228–3238.

3. Zhu JY, Park T, Isola P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Proc IEEE ICCV 2017. 2017; p. 2242–2251.
4. Li K, Yu L, Wang S, et al. Unsupervised Retina Image Synthesis via Disentangled Representation Learning. SASHIMI 2019. 2019; p. 32–41.
5. Cai Z, Xin J, Wu J, et al. Triple Multi-scale Adversarial Learning with Self-attention and Quality Loss for Unpaired Fundus Fluorescein Angiography Synthesis. IEEE EMBC 2020. 2020; p. 1592–1595.
6. Liang N, Yuan L, Wen X, et al. End-To-End Retina Image Synthesis Based on CGAN Using Class Feature Loss and Improved Retinal Detail Loss. IEEE Access. 2022;10:83125–83137.
7. Niu Y, Gu L, Zhao Y, et al. Explainable Diabetic Retinopathy Detection and Retinal Image Generation. IEEE J Biomed Health Inform. 2022;26(1):44–55.
8. Torbunov D, Huang Y, Yu H, et al. UVCGAN: UNet Vision Transformer Cycle-Consistent GAN for Unpaired Image-to-Image Translation. Proc IEEE/CVF WACV 2023. 2023; p. 702–712.
9. Hajeb Mohammad Alipour S, Rabbani H, Akhlaghi MR. Diabetic Retinopathy Grading by Digital Curvelet Transform. Comput Math Methods Med. 2012;2021:761901.
10. Sindel A, Hohberger B, Maier A, et al. Multi-modal Retinal Image Registration Using a Keypoint-Based Vessel Structure Aligning Network. MICCAI 2022. 2022; p. 108–118.
11. Budai A, Bock R, Maier A, et al. Robust Vessel Segmentation in Fundus Images. Int J Biomed Imaging. 2013;.