JOINT VESSEL SEGMENTATION AND DEFORMABLE REGISTRATION ON MULTI-MODAL RETINAL IMAGES BASED ON STYLE TRANSFER

Junkang Zhang[1], *Cheolhong An*[1], *Ji Dai*[1], *Manuel Amador*[2], *Dirk-Uwe Bartsch*[2], *Shyamanga Borooah*[2], *William R. Freeman*[2], *Truong Q. Nguyen*[1]

[1] Department of Electrical and Computer Engineering, UC San Diego, La Jolla, CA, 92093

[2] Department of Ophthalmology, Jacobs Retina Center at Shiley Eye Institute, UC San Diego,
La Jolla, CA, 92093

ABSTRACT

In multi-modal retinal image registration task, there are two major challenges, *i.e.*, poor performance in finding correspondence due to inconsistent features, and lack of labeled data for training learning-based models. In this paper, we propose a joint vessel segmentation and deformable registration model based on CNN for this task, built under the framework of weakly supervised style transfer learning and perceptual loss. In vessel segmentation, a style loss guides the model to generate segmentation maps that look authentic, and helps transform images of different modalities into consistent representations. In deformable registration, a content loss helps find dense correspondence for multi-modal images based on their consistent representations, and improves the segmentation results simultaneously. Experiment results show that our model has better performance than other deformable registration methods in both quantitative and visual evaluations, and the segmentation results also help the rigid transformation. [1]

Index Terms— Multi-Modal, Retinal Images, Deformable Registration, Vessel Segmentation, Style Transfer

1. INTRODUCTION

Multi-modal image registration is an important task in retinal image analysis. In order to better diagnose ophthalmologic diseases for a patient's eye, retinal images of multiple modalities captured by different imaging systems should be collected and aligned, since various modalities convey complementary information. A conventional routine (*e.g.*, [1]) for this task consists of two steps, *i.e.*, global coarse alignment (*e.g.*, rigid transformation) and local fine alignment (*e.g.*, deformable registration), as shown in Fig. 1. In coarse alignment, a *floating* image is warped globally towards a *reference* image of another modality, where the transformation parameters can be computed by key point detection and feature extraction (*e.g.*, SIFT), feature matching, and parameter estimation (*e.g.*, RANSAC). While in fine alignment, in order to reduce non-rigid matching errors and the errors from the previous step, local patterns in the floating image are further warped to accurately match those in another modality by estimating a registration field, where each individual pixel is assigned an offset vector. In this paper, we will also follow this routine, and mainly handle the fine alignment step by designing a deformable registration model.

One of the challenges in this task is that, it is hard for algorithms to locate and describe mutual information among two modalities, since common retinal patterns (*e.g.*, vessels) appear in different color intensities, contrasts and orientations etc. Previous works tried to solve the problem by designing robust features [2], landmark detectors [3] and line extractors [4] etc. Recently, Li *et al.*[1] extracted phase maps from images' Monogenic Signals [5] for registration, where the phase maps from multi-modal images share similar appearance. Heinrich *et al.*[6] also proposed a modality independent descriptor and matching schemes for rigid and deformable registration which can achieve good performance. However, most methods rely on hand-designed features whose performance is subjected to imaging qualities and modalities.

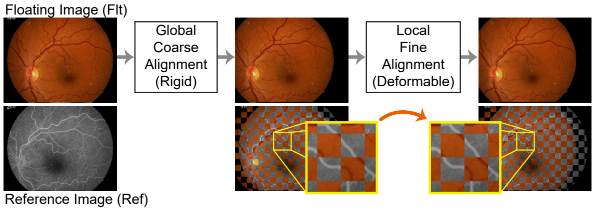

Floating Image (Flt)
Global Coarse Alignment (Rigid)
Local Fine Alignment (Deformable)
Reference Image (Ref)

Fig. 1. The coarse-to-fine routine for registration. The grid-shaped images overlay a warped floating image and a reference image, which visualizes the alignment of vessel patterns in two modalities.

Another challenge for multi-modal registration is lack of dense registration ground truths for training learning-based models. Especially, it is almost impossible to manually label deformation offsets for all pixels in retinal image pairs. Mahapatra *et al.*[7] proposed to directly generate a warped floating image via unsupervised Generative Adversarial Network (GAN), where the consistency constraint of content and cyclic transformation [8] between two input images are applied for training. They also proposed to learn to predict the registration fields with supervision from another existing method, whose performance might be limited. Other ideas for unsupervised training include the adoption of photometric consistency [9, 10], which might not be able to work on multi-modal images.

In this paper, we propose a deformable registration model based on Convolutional Neural Network (CNN) and weakly supervised end-to-end learning, which deals with both mentioned challenges. On one hand, we propose a learning scheme based on style transfer to train a vessel segmentation network without ground truth, in order to extract mutual patterns across two modalities for finding good correspondence. On the other hand, we build a deformable registration model which consists of the segmentation network and a registration field estimation network. And the whole model is trained via end-to-end learning merely using roughly aligned multi-modal images and a style target. To our best knowledge, this is the first CNN-based method to estimate deformation fields for multi-modal retinal images without training on ground truth.

[1] Supplementary materials and codes are available at `https://github.com/JunkangZhang/RetinalSegReg`.

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

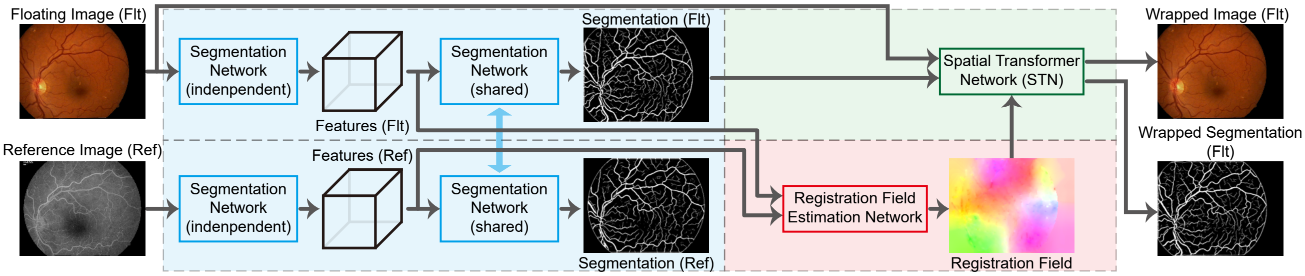

Floating Image (Flt)
Reference Image (Ref)
Segmentation Network (indenpendent)
Segmentation Network (indenpendent)
Features (Flt)
Features (Ref)
Segmentation Network (shared)
Segmentation Network (shared)
Segmentation (Flt)
Segmentation (Ref)
Spatial Transformer Network (STN)
Registration Field Estimation Network
Registration Field
Wrapped Image (Flt)
Wrapped Segmentation (Flt)

Fig. 2. The structure of the proposed model.

2. RELATED WORKS

Deformable registration. Bob *et al.*[9] trained a network for deformable registration field estimation via minimizing the difference between a reference image and a floating image warped by the predicted field. Spatial Transformer Networks [11] (STN) is incorporated as an image warper to keep the network differentiable. Similar ideas can be found in other works [10] and even generic optical flow estimation methods [12, 13], which additionally include smoothness constraints over the estimated field to get better results. However, all these methods assume the two input images share similar appearance or modality and cannot be directly applied in multi-modal scenarios.

Vessel segmentation and retinal image generation. Image segmentation can be viewed as a special case of image generation. Various CNN-based models (*e.g.*, [14]) have been established for retinal vessel segmentation via supervised training. Besides, Sadda *et al.*'s method [15] augments vessel segmentation datasets via CycleGAN [8] which generates retinal images from segmentation maps, yet it cannot produce images for an unseen modality. Hervella *et al.*[16] used CNN to transform a retinal image from its original modality to a target modality, which requires strictly aligned image pairs for training.

Deep style transfer is a popular technique for image generation, where a source image is transformed to share similar styles (*e.g.*, textures) as a target image while maintaining its own content (*e.g.*, object structure). Gatys *et al.*[17] proposed a perceptual loss to directly transform an image iteratively aided by a pretrained network. Johnson *et al.*[18] further finetuned the network on a specified target style to eliminate the iterative prediction process. In this paper, a similar framework is adopted for simultaneous vessel segmentation and deformation field estimation.

3. PROPOSED METHOD

The structure of the proposed deformable registration model is illustrated in Fig. 2. First, two roughly aligned images (*i.e.*, floating I_{flt} and reference I_{ref}) are fed into two segmentation networks H^s_x respectively to obtain vessel segmentation results $I^{seg}_x = \text{H}^s_x(I_x)$, $x \in \{flt, ref\}$. Afterwards, intermediate features extracted from both segmentation networks are propagated into a registration estimation network which predicts a registration field F. Finally, the floating image and its segmentation map can be warped according to the registration field via STN [11].

In the view of learning, our optimization scheme is similar with image style transfer based on perceptual loss [17, 18], which consists of a style loss and a content loss. In our model, the style loss minimizes the style features' differences between the segmentation result and an authentic segmentation map, such that the segmentation prediction looks close to a real one. Meanwhile, the content loss compares two segmentation maps, *i.e.*, the floating map warped by the registration field and the reference map without warping, to ensure that only mutual information in both modalities are extracted and the estimated field can properly warp the floating image.

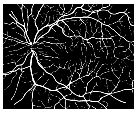


Fig. 3. The style target image I_{style} [19].

3.1. Vessel Segmentation via Style Loss

Many CNN-based vessel segmentation models rely on pixel-wise loss for training. However, existing datasets only have segmentation ground truth for few modalities. Furthermore, a network trained on one dataset might not function on other datasets or modalities due to varying imaging patterns among imaging systems. Instead, in our model, style transfer technique is adopted for training without pixel-wise ground truth. Our method adopts a pretrained CNN as a style feature descriptor to model the global vessel structures. Moreover, the segmentation network is trained by minimizing the style features' difference between the segmentation result and a style target.

According to [17, 18], a pretrained network ϕ is used as the feature extractor. It takes an image I and computes a feature tensor from its j-th layer as $\phi_j(I)$ with shape $c_j \times h_j \times w_j$. In order to find I's global styles (*e.g.*, the distribution of vessel widths, lengths and densities) while ignore their local distributions, a $c_j \times c_j$ Gram matrix $G_j(I)$ can be computed with its (i, k) element as

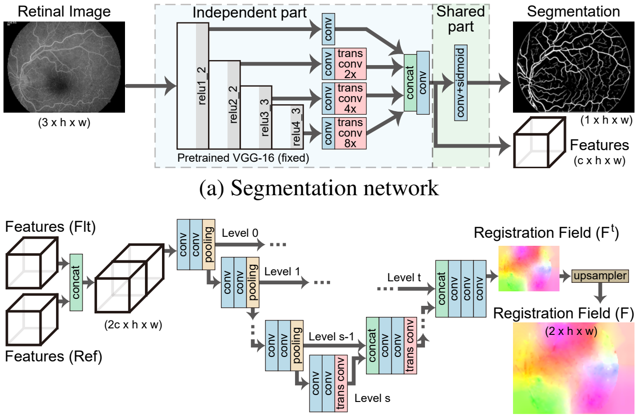
$$G_j(I)_{i,k}=\text{mean}\Big(\text{ele\_prod}\big(\phi_j(I)_{i,\cdot,\cdot},\phi_j(I)_{k,\cdot,\cdot}\big)\Big)/c_j \qquad (1)$$

where $\phi_j(I)_{i,\cdot,\cdot}$ is the i-th slice in $\phi_j(I)$ along the depth dimension, and $\text{ele\_prod}()$ is element-wise production over two input tensors. As a result, the spatial information in $\phi_j(I)$ is removed and only the summaries of styles (*e.g.*, vessels' global structure) are preserved. Finally, in order to minimize the style difference between two images I_1 and I_2, the style loss is defined as the Frobenius norm over the difference of their corresponding Gram matrices as

$$\mathcal{L}^j_{style}(I_1,I_2)=||G_j(I_1)-G_j(I_2)||^2_F. \qquad (2)$$

In our case, I_1 and I_2 are a segmentation prediction I^{seg}_x and a style target I_{style} respectively.

For training, we pick a segmentation map (shown in Fig. 3) from an outside dataset [19] as the style target I_{style} which has no pixel-wise correspondence with any of our training or test images. We only assume that both the style target and our retinal images share similar vessel structural styles, *e.g.*, the tree-like structures, continuously stretching and branching vessel paths with decreasing widths,



(a) Segmentation network

(b) Registration field estimation network

Fig. 4. Network structures.

etc. Besides, we adopt a modified DRIU [14] network for segmentation as shown in Fig. 4 (a), where several new layers are appended after a pretrained VGG-16 [20] network. The pretrained layers are kept fixed and are only used to extract lower-level information (*e.g.*, edges, lines) from retinal images. And the new layers will learn to combine the lower-level patterns into higher-level vessel structures supervised by the style loss. In addition, the VGG-16 network also functions as the style feature descriptors ϕ_j, where results from its 4 layers are used, *i.e.*, $j\in\{\text{relu1\_2,relu2\_2,relu3\_3,relu4\_3}\}$. The single-channel segmentation maps are duplicated into 3 channels to fulfill the requirement of VGG's input shape.

We set up two segmentation networks for floating and reference images respectively. Each network has an independent part and a shared part (*i.e.*, the last layer with sigmoid function) as shown in Fig. 2 and 4 (a). In order to provide mutual retinal information for the following deformation step, features from the independent layers of both networks are propagated into the next step. By sharing the last layer, both independent parts are guided to transform multi-modal images into consistent representations of similar modalities.

3.2. Registration Field Estimation via Content Loss

The features of similar modalities from the segmentation networks are helpful for the registration task. Nevertheless, there is no dense ground truth in retinal registration datasets. So we adopt a weakly supervised scheme to train the deformable registration network, assuming that the corresponding vessels are close to each other but not aligned. In addition, we propose a content loss which compares the warped floating segmentation map and the reference's segmentation, in order to improve the segmentation results I_{flt}^{seg} and I_{ref}^{seg} and predict the registration field F simultaneously.

In detail, the content loss is defined as

$$\mathcal{L}_{content}(I_{flt}^{seg},I_{ref}^{seg},F)=\text{MSE}\big(\text{STN}(I_{flt}^{seg},F),I_{ref}^{seg}\big) \tag{3}$$

where $\text{MSE}()$ is Mean Square Error function, $\text{STN}()$ [11] warps image I_{flt}^{seg} based on registration field F. From the view of the segmentation task, the content loss forces the segmentation results of two modalities to be as close as possible, which could reduce noises and recover missed predictions. Meanwhile, from the view of registration task, the mutual information (*i.e.*, vessels) in different modalities is utilized to find good dense correspondence.

The structure of our registration estimation network is shown in Fig. 4 (b) which is similar with UNet [21]. It takes in and concatenates two feature tensors from the segmentation networks. In its forward path, the features are downscaled s times via pooling layers and then upscaled $s-t$ times via transposed convolutional layers. In the meantime, features in the same scales are concatenated via skip connections during upscaling, so that multi-scale information are integrated for prediction. In order to generate smooth registration fields for both vessel and non-vessel areas, the network makes prediction F^t at a $2^t\times$ $(t>0)$ lower scale. In addition, a smoothness constraint is applied over F^t as

$$\begin{aligned}\mathcal{L}_{smooth}(F^t)=&\text{mean}_{k,i,j}(|F_{k,i,j}^t-F_{k,i+1,j}^t|)+\\&\text{mean}_{k,i,j}(|F_{k,i,j}^t-F_{k,i,j+1}^t|)\end{aligned} \tag{4}$$

where F^t has shape $2\times(h/2^t)\times(w/2^t)$ (h and w are height and width of an input image). Finally, F^t is upsampled via bilinear interpolation into F with shape $2\times h\times w$ to warp the floating image.

3.3. Miscellaneous and Summary

In order to force the segmentation networks to extract edges from both sides of vessels instead of on only one side, *i.e.*, to ensure rotation invariance in vessel segmentation, a self comparison loss is defined as

$$\mathcal{L}_{compare}(I_x)=\text{MSE}\Big(\text{rot}\Big(\text{H}_x^s\big(\text{rot}(I_x)\big)\Big),\text{H}_x^s\big(I_x\big)\Big) \tag{5}$$

where $\text{rot}(I_x)$ rotates the input image by $180°$. We also find that a SSIM [22] loss slightly improves performance, which is defined as

$$\mathcal{L}_{ssim}(I_{flt},I_{ref},F)=1-\text{SSIM}\big(\text{STN}(I_{flt},F),I_{ref}\big) \tag{6}$$

where $\text{SSIM}(\cdot,\cdot)\in[0,1]$ measures structural similarities between the warped floating image and reference image. Finally, the total loss is defined as

$$\begin{aligned}\mathcal{L}_{total}=&\sum_{x,j}\mathcal{L}_{style}^{j}(I_x^{seg},I_{style})+\lambda_1\sum_{x}\mathcal{L}_{compare}(I_x)\\&+\lambda_2\mathcal{L}_{content}+\lambda_3\mathcal{L}_{smooth}+\lambda_4\mathcal{L}_{ssim}\end{aligned} \tag{7}$$

where $x\in\{flt,ref\}$, and λ_i, i=1...4 are weighting parameters for different loss terms.

4. EXPERIMENTS

4.1. Settings

We use a dataset [23] with 59 pairs of multi-modal retinal images with shape 720×576 captured in both Color Fundus (the orange image in Fig. 1) and Fluorescein Angiography (the gray image in Fig. 1). 29 pairs are from healthy eyes and the rest show diseases. We take 30 pairs with odd index in their file names as the training set, and the rest 29 for test. For each pair, we manually label 3 pairs of matching points in both images to roughly align them via affine transformation. These coarse labels actually function as the weak supervision in training our deformation model.

Our model is implemented in PyTorch. The network is trained with $1.8e5$ updates by Adam [24] optimizer with learning rate as $1e{-}3$. We set s=5 and t=2 in the registration network, and empirically set λ_1=λ_2=1e-3, λ_4=1e-5 in the loss function. λ_3 is initially set to $2e$-5, and increased to $5e$-5 after $3e4$ updates to obtain smoother registration fields. During training, the data is augmented by applying random deformation over images, where the deformation fields are generated as $2\times4\times3$ arrays sampled from a normal

Table 1. Evaluation for deformable registration

Method	$Dice$	$Dice_s$
Before warping	0.2744	0.3934
MIND* [6]	0.6019	0.5269
Phase [5] + MIND* [6] (inspired by [1])	0.6178	0.5303
Ours	**0.6546**	**0.5398**

*Only the deformable registration part.

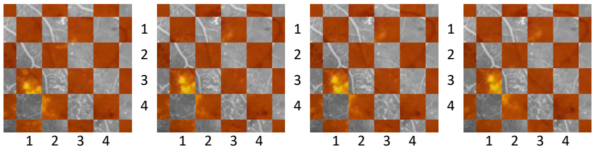


Fig. 5. Deformable registration on a difficult example. From left to right are input pairs, and results from MIND [6], Phase [5] + MIND, and our network respectively. Please zoom in to see details.

distribution (mean 0, standard deviation 5 pixels) and then interpolated to the same resolution as the images. In each update of training, two full-sized multi-modal images of a same eye are fed into the model (*i.e.*, batch size is 1) due to the fashion of style transfer loss and limited GPU memory. Predefined masks are laid over the segmentation results to remove noises outside the imaging circles.

To assess the performance of deformable registration without ground truth, some works (*e.g.*, [1]) use Dice Coefficient $Dice \in [0,1]$ to measure the overlapping degree of vessel structures from the warped floating image and the reference image. $Dice$ is defined as

$$Dice = \frac{2\times|A \bigcap B|}{|A|+|B|} \tag{8}$$

where A and B are binary vessel maps of the warped floating image and reference image respectively, predicted by some third party segmentation method. In this paper, we adopt B-COSFIRE [25] for $Dice$. Besides, we also extend $Dice$ into a soft Dice Coefficient $Dice_s \in [0,1]$ as

$$Dice_s = \frac{2 \cdot \mathrm{sum}\big(\mathrm{ele\_min}(A_p, B_p)\big)}{\mathrm{sum}(A_p)+\mathrm{sum}(B_p)} \tag{9}$$

where $\mathrm{ele\_min}(\cdot,\cdot)$ takes the element-wise minimum value across its two inputs, and A_p and B_p are vesselness probabilities of the floating and reference images. We use Frangi *et al.*'s method [26] to extract vesselness maps from retinal images.

4.2. Deformable Registration Results

We compare our model with two non-CNN methods, *i.e.*, MIND [6], and a scheme combining phase images [5] with MIND inspired by Li *et al.*[1]. For MIND, we only use its deformable registration part. Besides, to simulate the cases where the coarse alignment step produces large errors, we randomly deform the testing images in the same way as treating the training images. For fairness, we use an identical set of random deformations to test different methods.

Table 1 lists the $Dice$ and $Dice_s$ measurements by different methods. As can be seen, our method achieves better performance than other methods, with an increase of 0.0368 in $Dice$ and 0.0095 in $Dice_s$ over Phase [5] + MIND [6]. Fig. 5 shows a hard example where one of the modalities has vague vessels. To check the alignment of vessels, warped floating patches (in orange) and reference patches (in gray) are overlaid in grid view. As can been seen, our method works better in aligning vague and thin vessels, *e.g.*, the orange vessel in $3rd$ row and $3rd$ column of each example. This might be attributed to the global optimization scheme based on style transfer, where vague vessels can still be extracted to fulfill the style loss's constraint. More examples are available in the supplementary materials.

Table 2. Percentage (%) of correctly matched key point pairs

Input image type	Top-50	Top-100	Top-200	Top-400
Original multi-modal	88.34	88.21	87.53	84.11
Phase [5]	93.10	91.97	91.02	87.43
Ours (segmentation)	**98.07**	**97.72**	**97.47**	**97.31**

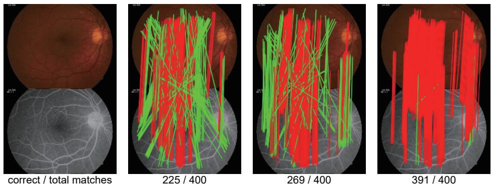


Fig. 6. Matched key point pairs using different input images. Red and greens lines mark correct and incorrect pairs respectively. From left to right are an input pair, and the results using original images, phase images [5] and our segmentation maps respectively.

4.3. Segmentation for Rigid Transformation

In order to evaluate segmentation results without ground truth, we take our segmentation network as an image transformer, and evaluate its performance in a simple rigid transformation workflow. First, dense key points are uniformly sampled at distance d in the two input images to be matched. Then, dense SIFT features are extracted in 64×64 windows centered on each points. Finally, key point features from both images are matched and sorted via brute-force searching based on minimum euclidean distance. For evaluation, we select the top-k pairs of matched key points, and count the number of correctly matched pairs. Considering the sampling distance and errors in manual labels, a matched pair is considered correct if the distance between the point in reference image and its ground truth location is less than d pixels. We set d=8, and use the original images without any warping in this experiment.

Three different types of input images are used for feature extraction, *i.e.*, the original multi-modal retinal images, their corresponding phase images, and the segmentation maps predicted by our network. The percentage of correctly matched points are summarized in Table 2. As can be seen, the success rate of our method has an obvious margin over other methods. When selecting top-400 pairs, our method has a margin of 9% over others. An example shown in Fig. 6 also verifies this result, where our method results in much less errors than other methods in finding correct correspondence.

5. CONCLUSION

In this paper, we handle the deformable registration task over multi-modal retinal images by proposing a joint vessel segmentation and registration model. Dense matching correspondence can be obtained with the help from segmentation task which extracts consistent representations from images of different modalities. The complete model is trained end-to-end under weak supervision aided by the style transfer framework. The experiments verifies the model's performance in both deformable and rigid transformation tasks.

6. REFERENCES

[1] Zhang Li, Fan Huang, Jiong Zhang, Behdad Dashtbozorg, Samaneh Abbasi-Sureshjani, Yue Sun, Xi Long, Qifeng Yu, Bart ter Haar Romeny, and Tao Tan, "Multi-modal and multi-vendor retina image registration," *Biomed. Opt. Express*, vol. 9, no. 2, pp. 410–422, 2018.

[2] J. Chen, J. Tian, N. Lee, J. Zheng, R. T. Smith, and A. F. Laine, "A partial intensity invariant feature descriptor for multimodal retinal image registration," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1707–1718, 2010.

[3] Álvaro S. Hervella, José Rouco, Jorge Novo, and Marcos Ortega, "Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement," *Procedia Computer Science*, vol. 126, pp. 97 – 104, 2018.

[4] M. Hernandez, G. Medioni, Z. Hu, and S. Sadda, "Multimodal registration of multiple retinal images based on line structures," in *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 907–914.

[5] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3136–3144, 2001.

[6] Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Sir Michael Brady, and Julia A. Schnabel, "Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical Image Analysis*, vol. 16, no. 7, pp. 1423 – 1435, 2012.

[7] D. Mahapatra, B. Antony, S. Sedai, and R. Garnavi, "Deformable medical image registration using generative adversarial networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 1449–1453.

[8] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[9] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Marius Staring, and Ivana Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 204–212.

[10] G. Balakrishnan, A. Zhao, M. R. Sabuncu, A. V. Dalca, and J. Guttag, "An unsupervised learning model for deformable medical image registration," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9252–9260.

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*, pp. 2017–2025. 2015.

[12] Jason J. Yu, Adam W. Harley, and Konstantinos G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *Computer Vision – ECCV 2016 Workshops*, 2016, pp. 3–10.

[13] Simon Meister, Junhwa Hur, and Stefan Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," 2018.

[14] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool, "Deep retinal image understanding," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, 2016, pp. 140–148.

[15] Praneeth Sadda, John A. Onofrey, and Xenophon Papademetris, "Deep learning retinal vessel segmentation from a single annotated example: An application of cyclic generative adversarial neural networks," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 2018, pp. 82–91.

[16] Álvaro S. Hervella, José Rouco, Jorge Novo, and Marcos Ortega, "Retinal image understanding emerges from self-supervised multimodal reconstruction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 2018, pp. 321–328.

[17] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, "A neural algorithm of artistic style," *CoRR*, vol. abs/1508.06576, 2015.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, 2016, pp. 694–711.

[19] Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, and Georg Michelson, "Robust vessel segmentation in fundus images," *International journal of biomedical imaging*, vol. 2013, 2013.

[20] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241.

[22] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[23] Shirin Hajeb Mohammad Alipour, Hossein Rabbani, and Mohammad Reza Akhlaghi, "Diabetic retinopathy grading by digital curvelet transform," *Computational and mathematical methods in medicine*, vol. 2012, pp. 1607–1614, 2012.

[24] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[25] George Azzopardi, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov, "Trainable cosfire filters for vessel delineation with application to retinal images," *Medical Image Analysis*, vol. 19, no. 1, pp. 46 – 57, 2015.

[26] Alejandro F. Frangi, Wiro J. Niessen, Koen L. Vincken, and Max A. Viergever, "Multiscale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI'98*, 1998, pp. 130–137.