A SEGMENTATION BASED ROBUST DEEP LEARNING FRAMEWORK FOR MULTIMODAL RETINAL IMAGE REGISTRATION

Yiqian Wang[1], *Junkang Zhang*[1], *Cheolhong An*[1], *Melina Cavichini*[2], *Mahima Jhingan*[2], *Manuel J. Amador-Patarroyo*[2], *Christopher P. Long*[3], *Dirk-Uwe G. Bartsch*[2], *William R. Freeman*[2], *Truong Q. Nguyen*[1]

[1] Electrical and Computer Engineering Department, UC San Diego, La Jolla, California, USA
[2] Jacobs Retina Center, Shiley Eye Institute, UC San Diego, La Jolla, California, USA
[3] School of Medicine, UC San Diego, La Jolla, California, USA

ABSTRACT

Multimodal image registration plays an important role in diagnosing and treating ophthalmologic diseases. In this paper, a deep learning framework for multimodal retinal image registration is proposed. The framework consists of a segmentation network, feature detection and description network, and an outlier rejection network, which focuses only on the globally coarse alignment step using the perspective transformation. We apply the proposed framework to register color fundus images with infrared reflectance images and compare it with the state-of-the-art conventional and learning-based approaches. The proposed framework demonstrates a significant improvement in robustness and accuracy reflected by a higher success rate and Dice coefficient compared to other coarse alignment methods.

Index Terms— Image registration, multimodal, retinal images, convolutional neural networks

1. INTRODUCTION

Retinal imaging plays a vital role in diagnosing and treating ophthalmologic diseases [1]. In order to help the observer confirm their diagnoses, multiple images of the same eye captured by different imaging systems are often collected and aligned. Multimodal image registration is therefore a crucial step to establish a comprehensive representation. Even though widely studied, it is still challenging [2] to detect and describe similar patterns in different modalities, because the same retina structure in different modalities can have disparate color, contrast, resolution, and orientation. Poor image quality is also ubiquitous in clinical applications, which makes multimodal image registration challenging.

Multimodal image registration often follows a coarse-to-fine pipeline (e.g. [2]). The *coarse alignment* step aligns the source image with the target image globally, which is usually done by a linear transformation, such as affine transformation or perspective transformation. The *fine alignment* step is often deformable in nature, which can reduce non-rigid errors after the coarse alignment. If the coarse alignment step is successful, the fine alignment step can proceed to improve local matching precision. However, if the coarse alignment were too far from ground truth, the fine alignment step would not be able to correct errors of the coarse alignment step. Therefore, improving the coarse alignment step is crucial to increase the overall success rate of registration. In this paper, we only focus on the coarse alignment step using perspective transformation.

To estimate the transformation (homography) between source and target images, there are two general approaches. One of them is *area-based* (or intensity-based), which finds transformation parameters to minimize the correlation [3] or mutual information [4] between the registered images. The other is a *feature-based* approach. A conventional feature-based approach has three major steps: vessel extraction, feature detection and description, and outlier rejection. The vessel extraction step extracts edge [5], phase [2] or vessel segmentation [6] in both images. After that, feature points and descriptors like scale invariant feature transformation (SIFT) [7] and speeded up robust features (SURF) [8] can be extracted. The extracted feature descriptors are often matched by the nearest neighbor, and then random sample consensus (RANSAC) [9] is used to reject outlier matches before estimating the transformation matrix.

However, the conventional registration methods are not robust enough for clinical applications, where images can be affected by poor imaging quality and severe diseases. In this paper, we propose a deep learning-based framework for the global registration of multimodal retinal images. The proposed method applies three different networks for vessel segmentation, feature detection and description, and outlier rejection. Rather than designing a single network for registration as in [10] or [11], the proposed method aims to improve progressively each part of the conventional homography estimation pipeline by replacing each algorithm with a network. To the best of our knowledge, this is the first successful completely learning-based method for multimodal retinal image registration.

2. RELATED WORKS

2.1. Vessel extraction

Many multimodal retina image registration approaches first extract vessel information from the source and target images to unify the modality and enhance vessel related features. An edge map [7] or a mean phase image [2] was used to derive implicitly the vessel information, and [6] explicitly obtained the vessel segmentation for more robust matching result.

In deep learning literature, the deep retina image understanding (DRIU) network [12] applied a supervised learning method for the vessel segmentation. Since pixel-wise ground truth is necessary to train the DRIU network, intensive annotation is required. In our previous paper [13], we proposed a style transfer network for vessel segmentation that can be trained without pixel-wise ground truth.

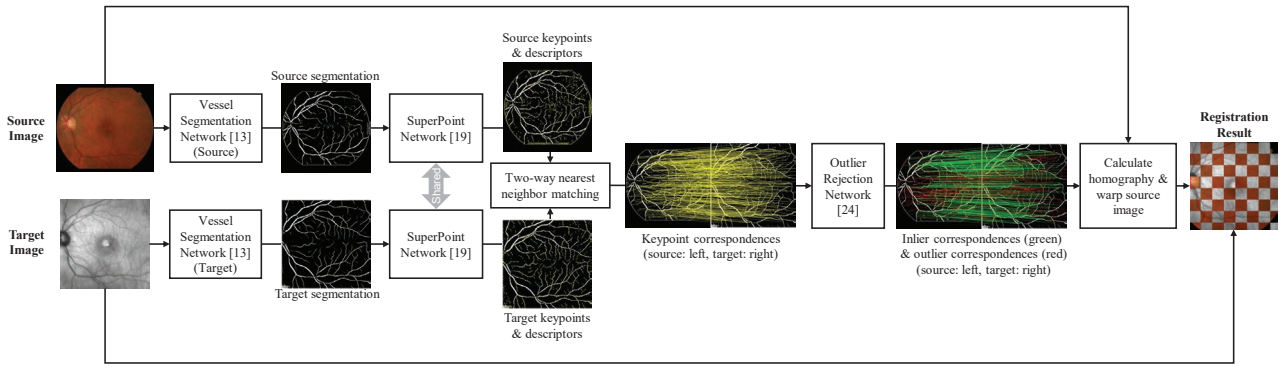

Fig. 1: Block diagram of the proposed registration framework

2.2. Feature detection and description

Traditional feature detection and description method includes scale invariant feature transformation (SIFT) [7] and speeded up robust features (SURF) [8]. Some algorithms only focus on feature detection, such as Harris corner detection [14], and some only focus on feature description, such as histogram of oriented gradients (HOG) [15]. Chen et al proposed a partial intensity invariant feature descriptor (PIIFD) [16] especially for multimodal retinal image registration. Lee et al proposed a low-dimensional step pattern analysis algorithm (LoSPA) [5] to improve the robustness of aligning unhealthy retinal images.

However, the hand-crafted feature detectors and descriptors may not be optimal. There are many recent works or researches in applying deep learning for interest point detection, such as learned invariant feature transform (LIFT) [17], and for feature description, such as universal correspondence network (UCN) [18]. The SuperPoint network [19] is a fully-convolutional network, which can detect interest points and generate corresponding descriptors for a grayscale image of any size.

2.3. Outlier rejection

Random sample consensus (RANSAC) [9] is the most fundamental method for outlier detection. After it randomly selects matching points and counts inlier number iteratively, it derives the model with the largest inlier number. The least median of squares (LMEDS) [20] method computes the median of square error in each iteration, and it is robust when the ratio of inliers is more than 50%. Other alternatives include adaptive outlier rejection based on asymmetric Gaussian mixture model (AGMM) [21] or root mean square error with feature distance (RMSEFD) [22].

In deep learning literature, the differentiable RANSAC (DSAC) [23] was introduced to modify the traditional RANSAC differentiable, which is suitable for end-to-end training, but its performance improvement is marginal. By contrast, Yi et al [24] trained a network to find inlier correspondences, which outperforms RANSAC in a single forward pass.

2.4. Image registration frameworks

In traditional methods, various frameworks can be designed by choosing different combinations of vessel extraction, interest point detector, feature descriptor, and outlier rejection method. In [2], a coarse registration method using a robust mean phase image and dense HOG descriptors was proposed, which demonstrated significant improvement in the registration of multimodal retinal images when compared with previous methods.

There are many newly proposed learning based multimodal retinal image registration methods, such as [13], [25] and [26]. However, they only focus on the deformable registration step with the assumption of affinely aligned image pair. The DLIR network [11] is an end-to-end network for the medical image registration that follows the conventional coarse-to-fine pipeline. However, it was proposed for single-modal cardiac cine MRI and chest CT images. The CNNGeo [10] is an end-to-end network for natural image registration, which also follows the coarse-to-fine pipeline. It first estimates 6 parameters for affine transformation, and then estimates a 12 parameter spline transformation on the coarsely warped image pair. Their network achieved state-of-the-art result when aligning natural images with different instances.

3. PROPOSED METHOD

We propose a multimodal retinal image registration framework, as shown in Fig. 1, that consists of a segmentation network, feature detection and description network, and an outlier rejection network.

3.1. Vessel segmentation network

We use the vessel segmentation network in our previous paper [13] to obtain segmentation map for both source and target images. The vessel segmentation network is a modified DRIU network [12] with pretrained VGG-16 for feature extraction. The network is trained with style transfer technique, which minimizes a style loss [27] that measures the difference of style features between the segmentation result and a style target. In this way, the network can be trained without pixel-wise segmentation ground truth, which requires intensive annotation.

As shown in Fig. 1, the source and target images are passed through two vessel segmentation networks separately, and the output segmentation maps are two single-channel grayscale images with bright vessels and dark background.

3.2. Feature detection and description network

The feature detection and description network uses the pretrained SuperPoint model [19]. The SuperPoint network takes a single channel $H \times W$ grayscale image as input, and outputs a $H \times W$ heatmap for interest point locations, and a $256 \times H \times W$ tensor for descriptors.

The interest point heatmap is post-processed by non-max suppression thresholded at 5 pixels, and keypoints are detected above the confidence threshold at 0.15. Then the corresponding descriptor

vectors are matched by two-way nearest neighbor algorithm, where the nearest neighbor matching from source to target must be the same as the nearest neighbor matching from target to source. After this step, we obtain the keypoint correspondences between source and target segmentation maps, as illustrated in Fig. 1.

3.3. Outlier rejection network

The outlier rejection network adopts a similar structure as in [24], but is larger and uses the perspective transformation matrix for supervision. The outlier rejection network is composed of 12 residual blocks with 256 dimensional weight-sharing perceptrons in each layer. The weight-sharing perceptrons process each correspondence independently and similarly, which makes the network invariant to permutations of the input correspondence. Besides batch normalization, the network uses a novel context normalization to synthesize contextual information across all correspondences.

As shown in Fig. 1, the network takes N pairs of correspondences with $N \times 4$ elements as input $\mathbf{x}$, where $\mathbf{x}_i = [x_i, y_i, x_i', y_i']$, $1 \leq i \leq N$, and outputs a $N \times 1$ vector $\mathbf{s}$ containing a probability score $s_i \in [0, 1)$ for each correspondence. Then a modified 4-point algorithm is used to calculate a 3×3 matrix $\mathbf{M}$ based on the coordinates and the scores.

The transformation matrix has 9 elements and 8 degrees of freedom. The traditional 4-point algorithm calculates a transformation matrix from $N \geq 4$ pairs of correspondence by solving a singular value problem. Denoting the source coordinates with (x, y) and the target coordinates with (x', y'), a $2N \times 9$ matrix $\mathbf{A}$ can be constructed by stacking the following pattern column-wise.

$$\begin{bmatrix} -x & -y & -1 & 0 & 0 & 0 & xx' & yx' & x' \\ 0 & 0 & 0 & -x & -y & -1 & xy' & yy' & y' \end{bmatrix} \quad (1)$$

Denote the vectorized transformation matrix $\mathrm{Vec}(\mathbf{M})$, the problem is to find $\mathrm{Vec}(\mathbf{M})$ that minimizes $||\mathbf{A}\mathrm{Vec}(\mathbf{M})||$, and its solution is the corresponding eigenvector of the smallest eigenvalue of $\mathbf{A}^T\mathbf{A}$. However, for the modified 4-point algorithm, we instead minimize $||\mathbf{WA}\mathrm{Vec}(\mathbf{M})||$, where $\mathbf{W}$ is a $2N \times 2N$ diagonal matrix of the output scores $\mathbf{W} = \mathrm{diag}([s_1, s_1, \ldots, s_N, s_N])$.

The loss function for training the outlier rejection network is the weighted sum of a classification loss, matrix regression loss and Dice loss, and the only supervision is the ground-truth transformation matrix between the two images. The classification loss is a balanced cross-entropy loss. Denote the output for the i-th correspondence at the last linear layer $o_i(\mathbf{x})$, and its label as an inlier $y_i \in \{0, 1\}$, then the classification loss is defined as

$$\mathcal{L}_{\mathrm{class}}(\mathbf{x}, \mathbf{M}_{gt}) = \frac{1}{N} \sum_{i=1}^{N} \gamma_i H(y_i(\mathbf{M}_{gt}), \sigma(o_i(\mathbf{x}))) \quad (2)$$

where $\mathbf{M}_{gt}$ denotes the ground-truth transformation matrix, $H(\cdot)$ denotes the binary cross-entropy, $\sigma(\cdot)$ denotes the sigmoid function, and γ_i is a per-label balancing factor. The label $y_i(\mathbf{M}_{gt})$ for each correspondence is obtained by thresholding its projected distance according to the ground truth transformation matrix

$$y_i(\mathbf{M}_{gt}) = \begin{cases} 1, & \text{if } ||T(\mathbf{p}_i, \mathbf{M}_{gt}) - \mathbf{p}_i'|| \leq 5 \text{ pixels} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $\mathbf{p}_i$ denotes the Euclidean coordinate of source, $\mathbf{p}_i'$ denotes that of target, and $T(\mathbf{p}_i, \mathbf{M}_{gt})$ denotes the transformed Euclidean coordinate of $\mathbf{p}_i$ under transformation $\mathbf{M}_{gt}$.

The matrix regression loss is a mean square error (MSE) loss on the transformation matrix

$$\mathcal{L}_{\mathrm{matrix}}(\mathbf{x}, \mathbf{M}_{gt}) = ||\mathbf{M}_{gt} - \mathbf{M}(\mathbf{x})||^2 \quad (4)$$

where $\mathbf{M}(\cdot)$ is the function to calculate the transformation matrix.

The Dice loss is one minus the Dice coefficient on the aligned image pair. For binary segmentation images, the Dice coefficient is defined as

$$\mathrm{Dice}(\mathcal{I}_1, \mathcal{I}_2) = \frac{2 \times \sum(\mathcal{I}_1 \odot \mathcal{I}_2)}{\sum \mathcal{I}_1 + \sum \mathcal{I}_2} \quad (5)$$

where $\odot$ denotes element-wise product. In our case, the soft Dice coefficient for grayscale segmentation [13] is defined as

$$\mathrm{Dice}_s(\mathcal{I}_1, \mathcal{I}_2) = \frac{2 \times \sum \mathrm{ele\_min}(\mathcal{I}_1, \mathcal{I}_2)}{\sum \mathcal{I}_1 + \sum \mathcal{I}_2} \quad (6)$$

with $\mathrm{ele\_min}$ denoting the element-wise minimum. In our case, $\mathcal{I}_1$ is the warped source segmentation image with the predicted matrix $\mathbf{M}(\mathbf{x})$, and $\mathcal{I}_2$ is the target segmentation. Then the soft Dice loss is

$$\mathcal{L}_{\mathrm{dice}}(\mathbf{x}, \mathcal{I}_{\mathrm{src}}, \mathcal{I}_{\mathrm{tgt}}) = 1 - \mathrm{Dice}_s(\mathrm{warp}(\mathcal{I}_{\mathrm{src}}, \mathbf{M}(\mathbf{x})), \mathcal{I}_{\mathrm{tgt}}) \quad (7)$$

The total loss is defined as

$$\begin{aligned} &\mathcal{L}(\mathbf{x}, \mathcal{I}_{\mathrm{src}}, \mathcal{I}_{\mathrm{tgt}}, \mathbf{M}_{gt}) = \lambda_{\mathrm{class}} \mathcal{L}_{\mathrm{class}}(\mathbf{x}, \mathbf{M}_{gt}) \\ &+ \lambda_{\mathrm{matrix}} \mathcal{L}_{\mathrm{matrix}}(\mathbf{x}, \mathbf{M}_{gt}) + \lambda_{\mathrm{dice}} \mathcal{L}_{\mathrm{dice}}(\mathbf{x}, \mathcal{I}_{\mathrm{src}}, \mathcal{I}_{\mathrm{tgt}}) \end{aligned} \quad (8)$$

where we choose $\lambda_{\mathrm{class}} = 1$, $\lambda_{\mathrm{matrix}} = 0.1$, $\lambda_{\mathrm{dice}} = 0.1$.

4. EXPERIMENTS

We compare our multimodal retinal image registration framework to the state-of-the-art in both traditional method [2] and deep learning [10]. Both methods follow the coarse-to-fine pipeline, but we only compare with their coarse registration part, since our method only includes coarse registration.

4.1. Dataset

We use our own dataset collected from Jacobs Retina Center at Shiley Eye Institute for training, validation, and testing. The dataset consists of pairs of source color fundus images (RGB, 3000×2672) and target infrared reflectance (IR) images (grayscale, 768×768 or 1536×1536), with 530 pairs in the training set, 90 in the validation set, and 253 in the test set.

The ground truth matrices in the training and validation set are first obtained by our vessel segmentation network + SuperPoint + RANSAC in the success cases, and calculated by manually labeled correspondences in the failure cases. The image pairs include noise, blur, over or under exposure, and severe diseases, which makes our dataset challenging for registration.

4.2. Implementation

All image intensities are normalized between $[0, 1]$, and the target grayscale images are converted to 3 channel by stacking the channel 3 times. The images are first padded to square shape and resized to 768×768 before feeding into the vessel segmentation network, and homography is estimated based on the segmentation maps. The transformation matrices of the padding and resizing step are recorded to calculate the transformation matrix for the original size.

The traditional method [2] uses mean phase image + dense HOG feature + RANSAC for affine registration. We implemented their

method in MATLAB since we couldn't find any original implementation. The RANSAC algorithm uses the OpenCV implementation with 5 pixel reprojection threshold and 2000 iterations by default. The CNNGeo uses the pretrained model in PyTorch with ResNet101 as feature extraction provided by the authors, and we resize our images to 240×240 to feed into CNNGeo. We also retrained CNNGeo on our training set for 1000 epochs using learning rate 10^{-3} and batch size 16, which demonstrates better performance than the pretrained model.

All the code for our network is implemented in PyTorch and optimized with Adam. The vessel segmentation network is trained on our dataset using a similar method discussed in our previous paper [13]. The SuperPoint uses the pretrained model by the original authors. The outlier rejection network is trained with saved interest point coordinates by SuperPoint on our dataset, using a learning rate of 10^{-4} and batchsize of 32 for 1000 epochs. All the coordinates are normalized in $[-1, 1]$, and the transformation matrices are modified correspondingly.

4.3. Evaluation

We evaluate the robustness of the proposed method by the success rate in our dataset. Successful alignment can be defined by thresholding the maximum error (MAE) on manually labeled pairs of points [5]. In this paper, we manually labeled 6 correspondences for test, and determine success registration by MAE less than 20 pixels based on the tolerance of our deformable registration network [13].

To evaluate the accuracy of registration, we calculate the Dice coefficient of the aligned binary segmentation images (threshold at 0.5) using eq. (5). The bordering artifacts in the segmentation map are masked when calculating the Dice coefficient. The Dice coefficient before registration is 0.0466 in our test set.

The quantitative result is shown in Table 1, and the qualitative result is shown in Fig. 2. The first part in Table 1 shows the result using different combination of algorithms, which reflects how our framework was designed. Replacing the dense HOG feature in [2] with SuperPoint increases the success rate by 31.62% from 48.22%, and increases the Dice coefficient by 0.18. Introducing the vessel segmentation network leads to higher success rate than using mean phase image, but the Dice coefficient becomes slightly lower. Finally, the outlier rejection network improves the success rate by another 11.86%, and achieved the highest Dice coefficient.

In comparison with other works, our proposed method outperforms the traditional method [2] and CNNGeo [10] with significant margins in both success rate and Dice coefficient. As shown in Fig. 2, our method successfully registered the two example pairs. Method [2] succeeded for pair 1, with slightly lower accuracy than our method, but failed for pair 2. Finally, the retrained CNNGeo failed for both pairs. Notice that the retrained CNNGeo can roughly align the images, but the overall MAE is too large, resulting in the low success rate and low Dice coefficient in Table 1.

Table 1: Registration success rate and Dice coefficient using different method on our test set. ("Phase": calculating mean phase image, "Seg.": vessel segmentation network, "Sup.": SuperPoint network, "Ran.": RANSAC, Our method: vessel segmentation network + SuperPoint network + outlier rejection network)

Method	Success rate	Dice
Phase+HOG+Ran. [2]	48.22% (122/253)	0.2956 ($\pm$0.2570)
Phase+Sup.+Ran.	79.84% (202/253)	0.4730 ($\pm$0.2130)
Seg.+Sup.+Ran.	82.21% (208/253)	0.4590 ($\pm$0.2145)
Our method	**94.07% (238/253)**	**0.5518 ($\pm$0.1654)**
Phase+HOG+Ran. [2]	48.22% (122/253)	0.2956 ($\pm$0.2570)
CNNGeo [10]	0.79% (2/253)	0.0854 ($\pm$0.0269)
Retrained CNNGeo [10]	5.13% (13/253)	0.0961 ($\pm$0.0398)
Our method	**94.07% (238/253)**	**0.5518 ($\pm$0.1654)**

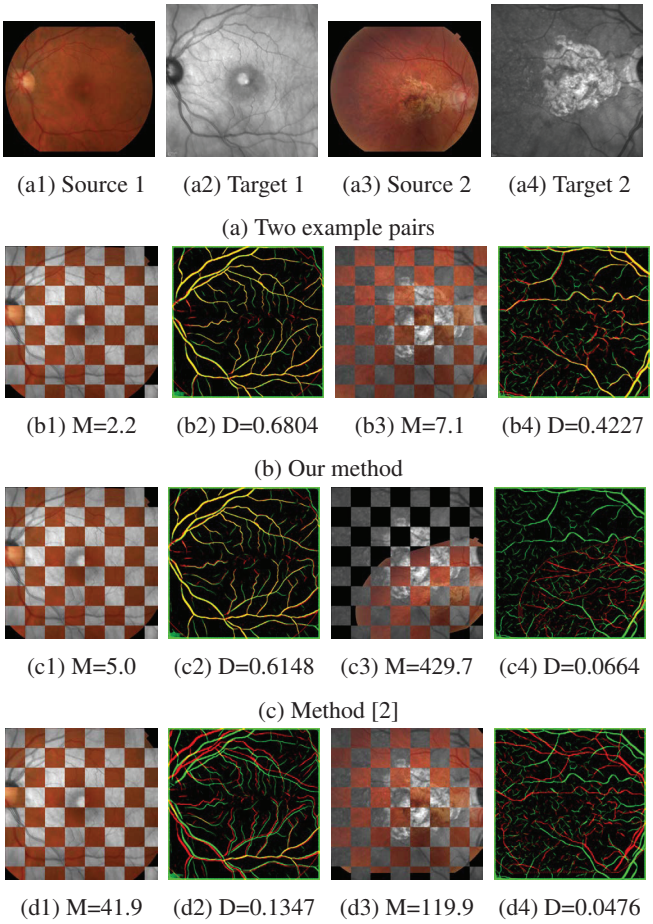

Fig. 2: Registration results of two example pairs using different methods. Please zoom-in to see details. In each method, (1), (2) show the results for pair 1, and (3), (4) show the registration results for pair 2. (1), (3) show the checkerboards of the registered images ("M": MAE in pixels, RGB tiles: warped source image, grayscale tiles: target image), and (2), (4) show the vessel segmentation overlay ("D": Dice coefficient, red: warped source segmentation, green: target segmentation, yellow: overlapping).

5. CONCLUSION

In this paper, we propose a deep learning framework for multimodal retinal image registration that focuses on the globally coarse registration step using perspective transformation. The framework cascades a vessel segmentation network, feature extraction and feature description network, and an outlier rejection network. The proposed framework is evaluated on registration of color fundus and IR images, and it demonstrates a significant improvement in both robustness and accuracy, reflected by much higher success rate and Dice coefficient compared to the other coarse registration methods.

6. REFERENCES

[1] T. J. MacGillivray, E. Trucco, J. R. Cameron, B. Dhillon, J. G. Houston, and E. J. R. Van Beek, "Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions," *The British journal of radiology*, vol. 87, no. 1040, pp. 20130832, 2014.

[2] Z. Li, F. Huang, J. Zhang, B. Dashtbozorg, S. Abbasi-Sureshjani, Y. Sun, X. Long, Q. Yu, B. t. Haar Romeny, and T. Tan, "Multi-modal and multi-vendor retina image registration," *Biomedical optics express*, vol. 9, no. 2, pp. 410–422, 2018.

[3] T. Chanwimaluang, G. Fan, and S. R. Fransen, "Hybrid retinal image registration," *IEEE transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 129–142, 2006.

[4] N. Ritter, R. Owens, J. Cooper, R. H. Eikelboom, and P. P. Van Saarloos, "Registration of stereo and temporal images of the retina," *IEEE Transactions on medical imaging*, vol. 18, no. 5, pp. 404–418, 1999.

[5] J. A. Lee, J. Cheng, B. H. Lee, E. P. Ong, G. Xu, D. W. K. Wong, J. Liu, A. Laude, and T. H. Lim, "A low-dimensional step pattern analysis algorithm with application to multimodal retinal image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1046–1053.

[6] J. Zhang, B. Dashtbozorg, E. Bekkers, J. P. Pluim, R. Duits, and B. M. t. Haar Romeny, "Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores," *IEEE transactions on medical imaging*, vol. 35, no. 12, pp. 2631–2644, 2016.

[7] D. G. Lowe et al., "Object recognition from local scale-invariant features.," in *iccv*, 1999, vol. 99, pp. 1150–1157.

[8] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[9] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[10] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6148–6157.

[11] B. D. d. Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical image analysis*, vol. 52, pp. 128–143, 2019.

[12] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Deep retinal image understanding," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 140–148.

[13] J. Zhang, C. An, J. Dai, M. Amador, D.-U. Bartsch, S. Borooah, W. R. Freeman, and T. Q. Nguyen, "Joint vessel segmentation and deformable registration on multi-modal retinal images based on style transfer," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 839–843.

[14] C. G. Harris, M. Stephens, et al., "A combined corner and edge detector.," in *Alvey vision conference*. Citeseer, 1988, vol. 15, pp. 10–5244.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005.

[16] 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.

[17] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *European Conference on Computer Vision*. Springer, 2016, pp. 467–483.

[18] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," in *Advances in Neural Information Processing Systems*, 2016, pp. 2414–2422.

[19] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[20] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American statistical association*, vol. 79, no. 388, pp. 871–880, 1984.

[21] H. Zhang, X. Liu, G. Wang, Y. Chen, and W. Zhao, "An automated point set registration framework for multimodal retinal image," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2857–2862.

[22] Z. Ghassabi, J. Shanbehzadeh, A. Sedaghat, and E. Fatemizadeh, "An efficient approach for robust multimodal retinal image registration based on ur-sift features and piifd descriptors," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 25, 2013.

[23] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6684–6692.

[24] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2666–2674.

[25] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging*, 2019.

[26] 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)*. IEEE, 2018, pp. 1449–1453.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.