Published in final edited form as:
IEEE Trans Image Process. 2022 ; 31: 5733–5747. doi:10.1109/TIP.2022.3201476.

Self-Supervised Rigid Registration for Multimodal Retinal Images

Cheolhong An,
Yiqian Wang,
Junkang Zhang [Member, IEEE],
Truong Q. Nguyen [Fellow, IEEE]
Department of Electrical and Computer Engineering, University of California at San Diego, San Diego, CA 92093 USA.

Abstract

The ability to accurately overlay one modality retinal image to another is critical in ophthalmology. Our previous framework achieved the state-of-the-art results for multimodal retinal image registration. However, it requires human-annotated labels due to the supervised approach of the previous work. In this paper, we propose a self-supervised multimodal retina registration method to alleviate the burdens of time and expense to prepare for training data, that is, aiming to automatically register multimodal retinal images without any human annotations. Specially, we focus on registering color fundus images with infrared reflectance and fluorescein angiography images, and compare registration results with several conventional and supervised and unsupervised deep learning methods. From the experimental results, the proposed self-supervised framework achieves a comparable accuracy comparing to the state-of-the-art supervised learning method in terms of registration accuracy and Dice coefficient.

Index Terms—

Rigid image registration; self-supervised learning; multimodal retinal image; convolutional neural network

I. Introduction

The retina is the only part of the central nervous system which can be imaged at high resolution in the living patient. It is accomplished by various retinal imaging modalities such as color fundus (CF), infrared reflectance (IR), fluorescein angiography (FA), and optical coherence tomography (OCT). Each imaging modality can provide different levels of pathology and visualize various retinal related diseases [1]. For this reason, the ability to accurately register other types of retinal imaging onto IR images and to align their vessels is critical to analyze multimodal information about the retina. For the multimodal retinal

Corresponding author: Truong Q. Nguyen. tqn001@eng.ucsd.edu.

image registration task, these images are aligned pixel-to-pixel to build a comprehensive visual representation of the eye, and help ophthalmologists confirming their diagnosis with multiple evidences. However, it is challenging to detect and match common patterns across modalities [2] because multimodal images not only significantly differ in appearance but also are captured in various image resolution and field of view from different instruments. Imaging quality and retinal diseases could also distort useful patterns for matching and increase the difficulty of registration. Therefore, it has been a widely studied topic to design a robust registration approach.

A *coarse-to-fine* pipeline is a common framework for registration [2], [3], [4] such that large mismatch is handled by the coarse (rigid) registration and small error is corrected by the fine (deformable) registration. Although the retina itself is close to a spherical shape, the imaging area can be approximated by a plane and registered with the affine, or perspective transformation in the globally coarse alignment step. The remaining error due to the planar approximation can be corrected by a deformable registration method in the locally fine alignment step. Between the two steps, the coarse alignment step is crucial for successful registration because most fine alignment methods cannot correct large errors of the coarse alignment step. Therefore, we focus on the coarse alignment step to increase registration accuracy using the self-supervised learning approach.

Most existing coarse alignment approaches fall into two categories: *Area-based* and *Feature-based*. The first category aims to minimize the mutual information [5] or the entropy correlation coefficient [6] between source and target images. Since the area-based methods are computationally intensive and the performance degrades when substantial texture difference exists, it is not suitable for the multimodal registration. The second category is based on detecting feature points and finding point correspondences among registration images. A feature-based pipeline often includes vessel extraction, feature detection and description, and outlier rejection [2], [7], [8], [9], [10], [11]. In our previous works [12], [13], we overcame existing limitations using the end-to-end deep learning framework. We proposed a content-adaptive multimodal retinal image registration method that focuses on the globally coarse alignment and includes three weakly supervised neural networks for vessel segmentation, feature detection and description, and outlier rejection. Our proposed framework of [13] achieved state-of-the-art performance on two different datasets. Although we minimized annotation efforts to prepare for training datasets using the unsupervised vessel segmentation and the pre-trained model of the feature detection, manual annotations were still required for keypoints between source and target images, ground-truth homography matrices, and inliers of the outlier rejection network. In this paper, we propose a self-supervised registration framework that does not require human-annotated labels to train the network. To the best of our knowledge, our proposed approach is the first self-supervised learning framework for multimodal retinal image registration.

II. Related Work

The pipeline structure of many conventional image registration includes vessel extraction (if any), feature detection and description, keypoint matching and outlier rejection. The structure is applicable for the deep learning methods and even for the self-supervised learning approaches. However, various frameworks can be designed by choosing different combinations with emphasis on different aspects and target applications. In this paper, we focus on the rigid registration framework for the multimodal retinal image.

A. Vessel Segmentation

The retinal vessel is the most common landmark in the multimodal retinal images. Therefore, many approaches first extract vascular information from the source and target images and unify the extracted vessel as a common modality after enhancing edges and corners of the vessel. In conventional (not learning) literature, [11] used an edge map based on strip fitting, [2] proposed a mean phase image using the Reize transform and log-Gabor filters, and [14] explicitly generated the vessel segmentation to achieve more robust matching result. In deep learning literature, DRIU [15], DUNet [16], and IterNet [17] obtained accurate vessel segmentation maps with strong supervision. In other words, pixel-wise segmentation is used for ground truth such that they require intensive manual annotations which is time consuming and costly. Since there is no known dataset with ground-truth vessel segmentation for IR retinal images, it is very challenging to directly apply them to our dataset. In our previous works [12], [18], we applied the style transfer network to segment vessels of CF and IR images without any experts' annotation and it was further improved by the unsupervised vessel segmentation network with a content-adaptive technique to maintain structure of input images [13]. In this paper, we apply our previous vessel segmentation networks to extract the vessels of source and target images and tightly couple them to the self-supervised learning framework as a common modality.

B. Feature Detection and Description

The retinal image registration task has adopt the feature detection and description methods of computer vision including Harris corner detection [19], histogram of oriented gradients (HOG) [20], scale invariant feature transformation (SIFT) [21], and speeded up robust features (SURF) [22]. Particularly, [8] proposed a partial intensity invariant feature descriptor (PIIFD) for the multimodal retinal images and [11] proposed a low-dimensional step pattern analysis algorithm (LoSPA) to improve the robustness for disease image pairs. In DeepSPA [23], a learning-based descriptor was proposed for retinal images, but it was trained based on the detected classes of hand-crafted step patterns using the conventional LoSPA descriptor [11], which could limit the performance of the network. The learning-based approaches for natural images, including learned invariant feature transform (LIFT) [24] and universal correspondence network (UCN) [25], improved matching performance compared to the conventional methods [24], [25]. LIFT can generate descriptors after detecting keypoints and UCN generates only dense descriptors with each pixel considered as a keypoint. Since keypoints and the corresponding descriptors are not derived jointly, the accuracy of both methods are limited. The SuperPoint network [26] overcame the limitation with one encoder and two decoders to obtain keypoints and descriptors jointly in a single forward pass which outperformed LIFT and UCN. However, the original model was trained first on synthetic dataset and then refined with target images (natural images), which is not ideal for retinal images. Therefore, we refined the Superpoint network with manually annotated keypoints for our dataset in our supervised learning framework [12], [13]. In this paper, we apply a self-supervised learning method to eliminate any manual annotations.

C. Outlier Rejection

Random Sample Consensus (RANSAC) [27] is the most commonly used method for outlier detection in computer vision. It applies an iterative approach that randomly selects matching points and votes for models based on the number of inliers. Another popular method, least median of squares (LMEDS) [28] computes the median of squared error in each iteration but it is only robust when the inlier ratio is more than 50%. Other iterative methods such as PROSAC [29], R-RANSAC [30], and USAC [31] achieved only marginal improvement over RANSAC with higher complexity. The iterative approaches such as GDB-ICP [7] and ED-DB-ICP [32] refining bootstrap regions were applied for the multimodal retina image registration. However, they are sensitive to scale. Adaptive outlier rejection based on asymmetric Gaussian mixture model (AGMM) [10] and root mean square error with feature distance (RMSEFD) [33] required longer runtime with intense tuning. To combine RANSAC and deep learning approaches, DSAC [34] introduced a differentiable RANSAC for end-to-end training with marginal improvement. By contrast, [35] trained a network to predict inliers and to reject outlier matchings such that the network outperformed RANSAC by a significant margin. Since the network was designed to estimate the essential matrix for camera pose estimation, some modifications were made to estimate the perspective transformation matrix for image registration [13].

D. Learning-Based Image Registration

Deep learning has been extensively used in the single-modality image registration task. However, most methods such as FlowNet [36], [37], PWC-Net [38], and the latest IRR-PWC [39] led to deformable registration for natural images. Since several large scale synthetic image datasets are publicly available, these models were first trained with the synthetic data and were fine-tuned to natural image pairs. For the medical image registration, AIRNet [40] and DLIR [4] was proposed for the medical images via the rigid or/and deformable registration. Voxelmorph [41], DIR-Net [42], and other approaches [43], [44], [45] were proposed to register single-modality images like magnetic resonance images (MRI) or computed tomography (CT) 3D volume and only for the deformable registration. This paper proposes to register 2D multimodal retina images with the rigid transformation. We compare the performance of DLIR [4] and Voxelmorph [41] in the experimental section IV.

The authors of [18], [46], [47] proposed the networks for multimodal retinal images but mostly focused on the deformable (local) registration step with assumption that input image pairs are coarsely aligned or their field of view are quite similar. If the deformable methods were directly applied to the original input images where large displacement exists or large difference of the field of view (45° color fundus, 30° IR as inputs in Fig 1), they would not be able to correctly align the images, which will be discussed in section IV. An end-to-end network CNNGeo [3], which was proposed for semantic alignment of multimodal natural images, followed the coarse-to-fine procedure via estimating 6 parameters for affine transformation and 12 parameters for spline transformation. It could handle large displacements but the success criterion of [3] is not as strict as that of the retinal image registration [9], [10], [11], [13]. The experimental results of [13] showed that CNNGeo [3] is worse than IRR-PWC [39] for the multimodal retinal image registration. Similarly, the correlation based registration was applied to register between MRI and DXA [48].

E. Self-Supervised Learning

Recently, self-supervised learning has gained popularity since it can significantly reduce or eliminate manually annotated labels. Self-supervised learning sets the learning objectives properly to obtain supervision from the training data itself. Thus, the learning method converts an unsupervised learning problem into a supervised one without annotated labels. Many self-supervised approaches were proposed and a large number of the pretext tasks have been studied for the self-supervised learning. The contrastive learning [49] sets a learning task to predict bottom blocks of the same column, which can enhance the main goal of representation learning. The rotation task [50] identifies a rotated degree among four degrees {0°, 90°, 180°, 270°} between input and output images. The jigsaw puzzle task [51] finds the relative position among patches. Self-supervised learning has also been applied for the retinal images such as denoising optical coherent tomography [52], diabetic retinal image classification [53], which utilizes the fuzzy clustering algorithm as self-supervision. We are not aware of any work on self-supervised learning for multimodal retinal image registration.

III. Proposed Method

In Fig. 1, we propose a framework for multimodal retinal image registration via the self-supervised learning approach. The proposed framework is composed of a vessel segmentation network, a feature detection and descriptor network, and an outlier rejection network, which are the same elements of our previous framework [12], [13]. However, the networks are trained fundamentally in different way: the previous works were trained in the supervised method whereas the proposed method is done in the self-supervised manner such that we do not directly provide any ground truth for the feature detection and descriptor, inlier matches, and the perspective transformation matrix. Even the vessel segmentation network was trained in the unsupervised manner in our previous works [12], [13]. In this work, we fix the model parameters of the vessel segmentation network with the pretrained models of [12], [13].

The proposed method also follows the feature based registration pipeline and entire steps of the pipeline are optimized in an end-to-end manner with self-supervision. Specifically, the proposed framework of Fig. 1 takes a source image I_{src} and its target image I_{tgt} and generates its vessel segmentations S_{src} and S_{tgt}. Next, the self-supervised feature detection and descriptor network finds keypoint pairs $\left\{\left(\mathscr{K}_{I_{src}}, \mathscr{K}_{S_{src}}\right), \left(\mathscr{K}_{I_{tgt}}, \mathscr{K}_{S_{tgt}}\right)\right\}$ between an image and its segmentation for both source and target images. These keypoints are classified into inlier pairs by the intra outlier rejection network and the intra transform matrix $\mathbf{M}_{intra}$ is derived for self-supervision. The learned keypoints of a common modality (here, vessel segmentation) $\mathscr{K}_{S_{src}}$ and $\mathscr{K}_{S_{tgt}}$ are mapped implicitly to grid points. The inter outlier rejection network predicts inliers and they are used to estimate the inter transform matrix $\mathbf{M}_{inter}$ to align a source image to a target image. To the best of our knowledge, the proposed method is the first fully self-supervised learning approach for the rigid registration of multimodal retinal images.

A. Vessel Segmentation Network

The structure of the vessel segmentation network in Fig. 2 is exactly same as that of our previous work [13]. However, we briefly review the vessel segmentation network to show the complete pipeline steps of our proposed framework. The input image is RGB CF for source or grayscale IR for target where the single channel IR image is converted to three channels with repetition. The output of a segmentation network is a single-channel grayscale image indicating vesselness at every pixel position. The source and target segmentation networks have its own encoder but share the same decoder for segmentation output.

We apply two types of the vessel segmentation network of [12] and [13] to evaluate the performance variations. Both networks were trained in an unsupervised or weakly supervised manner with a style loss to eliminate the dependency on manually labeled vessel segmentation as a ground truth. In [12], we proposed the CNN-based UNet for the vessel segmentation. The vessel segmentation network was improved using the pixel-adaptive convolution (PAC) [54] in order to enhance the robustness when aligning retinal images with various quality [13]. CNN benefits from the weight-sharing nature whereas it is also content agnostic in the sense that the same set of kernel is applied to different image contents and pixel locations. On the contrary, PAC [54] weighs the original 2D convolution kernel according to the feature guidance at different locations such that the network can adapt to different content. The mean phase image $\bar{\phi}(\mathscr{I})$ of Fig. 2, which is a guidance image, is obtained by taking the average of phase images at multiple scales,

$$\bar{\phi}(\mathscr{I}) = \frac{1}{N}\sum_{i=1}^{N}\phi_{\sigma_i}(\mathscr{I}) . \tag{1}$$

The phase of image $\mathscr{I}$ at scale σ is first computed by

$$\phi_{\sigma}(\mathscr{I}) = \arctan\left(\frac{G_{\sigma}(|\mathbf{f}_{\mathbf{R}}(\mathscr{I})|)}{G_{\sigma}(f_{e}(\mathscr{I}))}\right) \tag{2}$$

where $\mathbf{f}_{\mathbf{R}}(\mathscr{I})$ is the odd component of $\mathscr{I}$ extracted by the Reize transform [55], $f_e(\mathscr{I})$ is the even component [56], and $G_\sigma()$ represents the log-Gabor filter at scale σ [57]. To train the segmentation network in an unsupervised fashion, the style transfer technique was adopted [18]. Using style transfer, the network only requires one manually labeled segmentation map from any dataset to serve as a style reference. In this way, the network can be trained without any pixel-wise segmentation for all images in the dataset, which would require extensive manual annotation. Furthermore, since expert annotated vessel segmentation is not publicly available for IR dataset, the unsupervised vessels segmentation network is a crucial step for the success of the proposed framework.

Empirically, we observed that the mean phase image was very robust in enhancing edges in the original image even with low contrast. The pixel adaptive convolution could adjust local convolution kernel weights in the segmentation network according to guidance of the mean phase image such that it can improve vessel segmentation for retinal images with various quality. In this paper, we denote the CNN-based segmentation network of [12] and the PAC-based segmentation network of [13] as *ConvSeg* and *PACSeg*, respectively and fix the network parameters during the training. Interested readers should refer to the [12], [13] for more details. We will compare the performance of these two segmentation networks in the proposed self-supervised learning framework and discuss the benefits of content adaptation in section IV.

B. Feature Detection and Description Network: Self-Supervised Keypoint Detection

The feature detection and description network of [58] is adopted for the self-supervised keypoint learning, which completely eliminates fine-tuning procedure on the pretrained SuperPoint model [26] of [13]. It leverages a shared-encoder backbone with three output heads for feature points, scores, and descriptions, as illustrated in Fig. 3. The structure of the self-supervised feature detection and description network aims to regress a function that takes an image as input and outputs feature points, descriptors, and scores.

A CF source image I_{src} or an IR target image I_{tgt} and its corresponding vessel segmentation map S_{src} or S_{tgt} are three-channel $H \times W$ images where W and H denote the width and height of an image and a single channel I_{tgt}, S_{src}, and S_{tgt} are converted to three channels with repetition. The input images I_{src} and I_{tgt} are warped to I^{w}_{src} and I^{w}_{tgt} using a random homography transformation $\mathbf{M}_{ss}$ for self-supervised learning. The network parameters are shared for all input images and vessel segmentation maps. Specifically, we train the network to map an input image $I \in R^{3 \times H \times W}$ to output keypoint scores $sc \in R^{N}$, feature points (keypoints) $p \in R^{2 \times N}$, and descriptor $f \in R^{256 \times 4N}$ for each input: $W_{\theta}: I \rightarrow \{p, sc, f\}$, where $I \in \{I^{w}_{src}, I^{w}_{tgt}, S_{src}, S_{tgt}\}$ and $N = W_{c} \cdot H_{c} = \frac{W}{8} \cdot \frac{H}{8}$. Instead of selecting interest point locations from the heatmap as in [26], the authors of [58] applied a regression network to predict a single keypoint for each 8 × 8 region of an input image [59] where the predicted location p_i is the sum of the network output o_i and the center of an 8 × 8 region c_i with $i \in \{1, 2, 3, \ldots, N\}$ as illustrated in Fig. 4. Moreover, the predicted location can be outside the 8 × 8 border for better matching and aggregation as denoted in the green regions. The score head regresses confidence of keypoints, which is bounded within [0, 1] using a sigmoid function. The score is also constrained such that scores of a matching keypoint pair should be similar. The descriptor head provides 256-dimensional features with a higher resolution grid to capture finer details. The subpixel convolution via pixel shuffle operation [60] is used to upsample descriptors such that each keypoint can interpolate its corresponding descriptor with half-pixel accuracy descriptors. The top-K keypoints $\mathscr{K}$ among N keypoints of the feature point network are selected based on the confidence score where $\mathscr{K} \in \left\{\mathscr{K}_{I^{w}_{src}}, \mathscr{K}_{I^{w}_{tgt}}, \mathscr{K}_{S_{src}}, \mathscr{K}_{S_{src}}\right\}$. K keypoints and their corresponding descriptors are passed to the outlier rejection network for the next step as illustrated in Fig. 1. Specifically, the top-K confident keypoints are determined with their own scores for source and target keypoint pairs $(\mathscr{K}_{I^{w}_{src}}, \mathscr{K}_{S_{src}})$ and $(\mathscr{K}_{I^{w}_{tgt}}, \mathscr{K}_{S_{tgt}})$ as follows:

$$
\begin{aligned}
&i = \arg\,\text{sort}(sc^{S}, \text{descending}) \quad , \quad i \in \{1, 2, 3, \ldots, K\} \\
&j = \arg\,\text{sort}\left(sc^{t}, \text{descending}\right), \quad j \in \{1, 2, 3, \ldots, K\} \\
&d\left(p_i^s, p_j^{t\,*}\right) = \max_j \, f_i^{s\,\top} f_j^t, \ \forall i
\end{aligned}
\tag{3}
$$

where $d\left(p_i^S, p_j^{t\,*}\right)$ is the correlation distance of descriptors between the top-K source and target keypoints. In Fig. 1, the keypoint pairs and the corresponding correlation distance for source and target images are denoted as $\mathbf{x}_1 = \left\{p_i^s, p_j^{t\,*}, d\left(p_i^s, p_j^{t\,*}\right)\right\}$ where $(p_i^s, p_i^t) \in (\mathscr{K}_{I_{src}}^{w}, \mathscr{K}_{S_{src}})$ and $\mathbf{x}_2 = \left\{p_i^s, p_j^{t\,*}, d\left(p_i^s, p_j^{t\,*}\right)\right\}$ where $(p_i^s, p_i^t) \in \left(\mathscr{K}_{I_{tgt}}^{w}, \mathscr{K}_{S_{tgt}}\right)$ The main goal of the proposed feature detection and description network is learning keypoints and their description to map a CF or IR image to its corresponding vessel segmentation via the self-supervised learning.

C. Feature Detection and Description Network: Implicit Keypoint Detection

Although the proposed self-supervised network learns keypoints and their descriptors, the keypoints of a warped source CF image and its vessel segmentation $\mathscr{K}_s, s \in \{(\boldsymbol{I}_{src}^{w}, \boldsymbol{S}_{src})\}$ and the keypoints of a warped target IR image and its vessel segmentation $\mathscr{K}_t, t \in \{(\boldsymbol{I}_{tgt}^{w}, \boldsymbol{S}_{tgt})\}$ are learned independently. There is no direct loss function to constraint that the keypoints of a source image and its vessel segmentation $\mathscr{K}_s$ are consistent to the keypoints of a target image and its vessel segmentation $\mathscr{K}_t$ since we do not use any ground truth of keypoints between $\boldsymbol{I}_{src}$ and $\boldsymbol{I}_{tgt}$ in our proposed framework. As a result, we might not find matching keypoints from a source CF to a target IR vessel segmentation. Therefore, we propose the cross keypoint matching from the keypoints of the self-supervised learning to grid keypoints for the multimodal registration. Since the implicit keypoint detection is to map from a CF source image to an IR target image through a common modality of a vessel segmentation, we search the top-K keypoints of a source segmentation p_i^s, which are learned from the self-supervised learning, to the grid-points of an IR target segmentation $p_j^t \in \mathscr{G}_{S_{tgt}}$, that is, output of feature points $\boldsymbol{P}_{map}$ in Fig. 3. Moreover, the top-K keypoints of an IR vessel segmentation p_i^t are mapped to the grid-points of a source segmentation $p_j^s \in \mathscr{G}_{S_{src}}$ to maximize correlation of the keypoints descriptor as follows:

$$
\begin{aligned}
&i \in \{1, 2, 3, \ldots, K\} = \arg\,\text{sort}\left(sc\left(p_i^l\right),\ \text{descending}\ \right), \\
&j \in \{1, 2, 3, \ldots, N\} = \arg\,\text{sort}\left(sc\left(p_j^l\right),\ \text{descending}\ \right), \\
&d_{kg}\left(p_i^s, p_j^{t\,*}\right) = \max_j f_i^s f_j^t, d_{gk}\left(p_i^t, p_j^{s\,*}\right) \\
&= \max_j f_i^t f_j^s, \ \forall i
\end{aligned}
\tag{4}
$$

where $l \in \{t, s\}$, $sc\left(p_i^l\right)$ and $sc\left(p_j^l\right)$ are the corresponding scores of p_i and p_j for the CF source and IR target vessel segmentations, respectively. Note that (eq. 4) is different from (eq. 3). The top-K confident keypoints between N grid points and top-K keypoints are obtained in (eq. 4) for the inter outlier rejection network while the top-K confident keypoints between image and vessel segmentation are determined as described in (eq. 3) for intra outlier rejection network. Since we do not use any ground-truth keypoints of the supervised method in our proposed method, we cannot directly apply any loss function to pair keypoints during training. Furthermore, we cannot use the self-supervised method directly to map keypoints between a CF source image and an IR target image because CF and IR images are taken by different instrument with different image characteristics such as field of view, resolution, imaging method and so on, that is, they are multimodal images. The mismatch loss of keypoints and descriptors is implicitly affected by the Dice loss function between a CF vessel segmentation and an IR vessel segmentation after applying the inter outlier rejection network and estimating the inter homography matrix as illustrated in Fig. 1. The keypoint pairs and the associated descriptor score $\mathbf{x}_3 = \left\{\left(p_i^s, p_j^{t^*}, d_{kg}\right)\right\}$ from the source keypoints to the target grid-points and $\mathbf{x}_4 = \left\{\left(p_j^{s^*}, p_i^t, d_{gk}\right)\right\}$ from the source grid-points to the target keypoints are stacked to be $\mathbf{x}_{inter} \in R^{2K \times 5}$ and they are fed into the inter outlier rejection network.

D. Outlier Rejection Network

The outlier rejection network is intended to replace the conventional RANSAC with the CNN network to estimate confidence of keypoint pairs, which ultimately guides the sampling of minimal sets. It cascades the residual blocks [13] as shown in Fig. 5 but we expand its structure to take correlation distance in addition to pairs of keypoints' position as in [58] and [61]. In order to register multimodal data with the self-supervised approach, we propose two outlier rejection networks: *intra multimodal outlier rejection network* and *inter multimodal outlier rejection network*. The intra multimodal outlier rejection network estimates inliers for the self-supervised registration from CF and IR images to its corresponding vessel segmentation. On the other hand, the inter multimodal outlier rejection network predicts those for the multimodal registration from a CF vessel segmentation to an IR vessel segmentation as illustrated in Fig. 1. Although two outlier rejection networks share the network parameters, the input of the intra multimodal outlier rejection network $\mathscr{K}_t, t \in \{(I_{src}^w, S_{src}), (I_{tgt}^w, S_{tgt})\}$ is different from one of the inter multimodal outlier rejection network $\mathscr{K}_t, t \in \{(S_{src}, S_{tgt}^g), (S_{src}^g, S_{tgt})\}$.

In the intra outlier rejection network, the top-K pairs of keypoints for a CF source image and a IR target image support $K \times 5$ dimension, which are denoted as $\mathbf{x}_1$ and $\mathbf{x}_2$ in Fig 1. They are stacked to be $\mathbf{x}_{intra} \in \mathbb{R}^{2K \times 5}$ and are fed into the intra multimodal outlier rejection network. The output of the network is a vector $\mathbf{w} \in \mathbb{R}^{2K \times 1}$ containing a probability score of an inlier $w_i \in [0, 1)$ for each correspondence. We reduce the number of residual blocks of the proposed network from 12 to 8 since the performance is better, which will be discussed in subsection IV-C. The major difference of the proposed outlier rejection network to the one in [58] is that the score $\mathbf{w}$ of the outlier rejection network is directly applied to estimate a homography matrix. Consequently, the proposed outlier rejection network is used to register a source to a target but that of [58] is only utilized as a proxy task to improve the keypoint detection and description. Therefore, the outlier rejection network of [58] is only applied for training while the proposed outlier rejection network is applied for testing as well as for training.

Since the perspective transformation has 8 degrees of freedom, at least 4 pairs of correspondences are required. In order to estimate the 3 × 3 perspective transformation matrix $\mathbf{M}$ from pairs of coordinates and weighting scores, we adapt a weighted version of

4-point algorithm, which allows soft-assignment to put more weights on inliers with higher probabilities whereas the conventional 4-point algorithm uses hard-assignment, which is considered as a special case of the weighted 4-point algorithm with binary weighting scores. For the i-th pair of correspondence, let us denote the source and the target coordinates by (x_i, y_i) and $(x_i^{'}, y_i^{'})$ and define matrix $\mathbf{A}_j \in \mathbb{R}^{4K \times 9}$ as

$$\mathbf{A}_j = \begin{bmatrix} 0 & 0 & 0 & x_1 & y_1 & 1 & -x_1 y_1^{'} & -y_1 y_1^{'} & -y_1^{'} \\ x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1 x_1^{'} & -y_1 x_1^{'} & -x_1^{'} \\ \vdots & & & & \vdots & & & & \vdots \\ 0 & 0 & 0 & x_{2K} & y_{2K} & 1 & -x_{2K} y_{2K}^{'} & -y_{2K} y_{2K}^{'} & -y_{2K}^{'} \\ x_{2K} & y_{2K} & 1 & 0 & 0 & 0 & -x_{2K} x_{2K}^{'} & -y_{2K} x_{2K}^{'} & -x_{2K}^{'} \end{bmatrix},$$

then the transformation matrix is obtained by solving the optimization problem:

$$\mathbf{M}_{\mathrm{j}}^{*} = \underset{\mathbf{M}_j}{\arg\min} \left\| \mathbf{W}_j \mathbf{A}_j \mathrm{Vec}(\mathbf{M}_j) \right\|, \tag{5}$$

where $j \in \{intra, inter\}$ and $\mathbf{W}_j \in \mathbb{R}^{4K \times 4K}$ is a diagonal matrix of the output scores $\mathbf{W}_j = \mathrm{diag}([w_1, w_1 \ldots, w_{2K}, w_{2K}])$ where $\mathbf{w}$ is the inliner confidence of the outlier rejection network as shown in Fig. 5. It can be proved that the solution is the corresponding eigenvector of the smallest eigenvalue of $\mathbf{A}_j^T \mathbf{W}_j^2 \mathbf{A}_j$. We derive the homography transformation matrices, $\mathbf{M}_{intra}$ and $\mathbf{M}_{inter}$ using the corresponding output score $\mathbf{W}_j$ of the intra and inter outlier rejection networks as in (eq. 5) and illustrated in Fig. 1.

E. Loss Function

The self-supervised feature detection and descriptor network in Fig 1 is trained with the positional loss function λ_{pos}, the descriptor loss function $\mathcal{L}_{\mathrm{desc}}$, and the score loss function $\mathcal{L}_{\mathrm{score}}$ in (eq. 10). The position distance d_{ij} of λ_{pos} is computed by L2 norm between the warped source keypoints p^w using randomly generated homography transformation matrix $\mathbf{M}_{ss}$ and the target keypoints p^t: $d_{ij} = \|p_i^w - p_j^t\|_2$, where $p_i^w = T(p_i^s, \mathbf{M}_{ss})$ and $i, j \in \{1, 2, 3, \ldots, N\}$. We find matching pairs $\{(p_i^s, p_j^t)\}$ with a threshold Th of a confident distance as in (eq. 6) and the positional loss $\mathcal{L}_{pos}$ is derived as a mean of distance of confident matching pairs $\{(p_i^{s*}, p_j^{t*})\}$ as follows:

$$\left\{ \left(p_i^{s*}, p_j^{t*}\right) = \underset{i, j}{\arg\min}\ d_{ij}, \quad \mathrm{s.t.}\ d_{ij} < Th \right\} \tag{6}$$

$$\mathcal{L}_{\mathrm{pos}} = \frac{1}{N*} \sum d_{ij}^{*}, \tag{7}$$

where $N*$ is the total number of confident matching pairs and $d_{ij}^* = \left\| T\left(p_i^{s\,*}, \mathbf{M}_{ss}\right) - p_j^{t\,*} \right\|_2$.

The output of the descriptor head is a $256 \times \frac{H}{4} \times \frac{W}{4}$ tensor as shown in Fig. 3. The tensor contains 256-dimensional descriptors for a total of $\frac{H}{4} \times \frac{W}{4}$ after upsampling by a factor of two at the center of each 8 × 8 block of the original resolution. The descriptor loss $\mathcal{L}_{\text{desc}}$ in (eq. 8) is a triplet loss to minimize the distance between the anchor descriptor f^S and the positive descriptor f_+^w, while it maximizes the distance between the anchor descriptor f^S and the negative descriptor f_-^w, which is the closest in the descriptor space after excluding a positive descriptor. Although any sample other than the true match can be used as the negative pair, we choose the hardest negative sample to train the network.

$$\mathcal{L}_{\text{desc}} = \frac{1}{N} \sum_i \max(0, d(f_i^s, f_{i,+}^w) - d(f_i^s, f_{i,-}^w) + m), \tag{8}$$

where $d(f_i^s, f_i^w) = \|f_i^s - f_i^w\|_2$, $f_i^s = f^S(p_i^s)$, $f_{i,+}^w = f^t(T(p_i^s, \mathbf{M}_{ss}))$.

The score loss $\mathcal{L}_{\text{score}}$ [59] is composed of the unsupervised point (USP) loss ℓ_{usp} and the similarity loss of a point-pair score ℓ_{sim}, as follows:

$$\begin{aligned} \mathcal{L}_{\text{score}} &= \frac{1}{N} \sum_i \ell_i^{usp} + \ell_i^{sim} \\ \ell_i^{usp} &= \frac{sc^S\left(p_i^{s\,*}\right) + sc^t\left(p_i^{w\,*}\right)}{2}\left(d_{ij}^* - \mathcal{L}_{\text{pos}}\right) \\ \ell_i^{sim} &= \left(sc^S\left(p_i^{s\,*}\right) - sc^t\left(p_i^{w\,*}\right)\right)^2, \end{aligned} \tag{9}$$

where $p_i^{w\,*} = T\left(p_i^{s\,*}, \mathbf{M}_{ss}\right)$ and sc^S, sc^t are scores of source and target, respectively. The overall objective of the score loss is to improve repeatability, i.e., keypoints are consistent regardless of the homography transformation $\mathbf{M}_{ss}$ and multimodal images (here, image and vessel segmentation). The total loss of the self-supervised keypoint network is

$$\mathcal{L}_{\text{ss-key}} = \lambda_{pos}\mathcal{L}_{\text{pos}} + \lambda_{desc}\mathcal{L}_{\text{desc}} + \lambda_{score}\mathcal{L}_{\text{score}}\,. \tag{10}$$

The multimodal intra outlier rejection network in Fig 1 is trained with the loss function $\mathcal{L}_{\text{intra-io}}$ in (eq. 14), which is defined as a weighted sum of two terms. The first term $\mathcal{L}_{\text{class}}$ in (eq. 11) is a classification loss, that is a cross-entropy loss between the predicted and ground truth labels for each correspondence. Let $o_i(\mathbf{x})$ be the last linear layer output for the i-th correspondence, and $y_i(\mathbf{M}_{ss})$ in (eq. 12) be its label as an inlier or an outlier given the ground-truth transformation matrix $\mathbf{M}_{ss}$, which is obtained from the self-supervised keypoint detection in Fig. 3. Then the classification loss is defined as

$$\mathcal{L}_{\text{class}}(\mathbf{x}_{intra};\mathbf{M}_{ss}) = \sum_{i=1}^{2K} -\frac{1}{N_c} y_i(\mathbf{M}_{ss}) \cdot \log \sigma(o_i(\mathbf{x}_{intra})), \tag{11}$$

where $c \in \{p, n\}$ and $\sigma(\cdot)$ denotes the sigmoid function. N_p is the number of inliers and N_n is the number of outliers such that $N_p + N_n = 2K$. The label $y_i(\mathbf{M}_{ss})$ for each correspondence is obtained by thresholding L_2 distance between the target coordinates p_i^t and warped source coordinates $T(p_i^s, \mathbf{M}_{ss})$

$$y_i(\mathbf{M}_{ss}) = \begin{cases} y_i^p = 1, & \text{if } \| T(p_i^s, \mathbf{M}_{ss}) - p_i^t \| \le 3 \text{ pixels} \\ y_i^n = -1, & \text{otherwise} \end{cases} \tag{12}$$

where p_i^s and p_i^t are defined in (eq. 3) and (eq. 4) for the self-supervised keypoints and the implicit keypoints, respectively. The matrix regression loss $\mathcal{L}_{\text{matrix}}$ is a mean squared error (MSE) between the predicted and ground-truth transformation matrices

$$\mathcal{L}_{\text{matrix}}(\mathbf{x}_{intra};\mathbf{M}_{ss}) = \text{MSE}(\mathbf{M}_{ss} - \mathbf{M}_{intra}(\mathbf{x}_{intra})) \tag{13}$$

where $\mathbf{M}_{intra}$ denotes the predicted transformation matrix in (eq. 5) given $\mathbf{x}_{intra}$. The $\mathcal{L}_{\text{matrix}}$ loss makes $\mathbf{M}_{intra}$ to be close to $\mathbf{M}_{ss}$, which is a ground-truth transformation matrix in the self-supervised manner for the intra alignment (alignment between a warped image to its own vessel segmentation). Then the total loss of the multimodal intra outlier rejection network is

$$\mathcal{L}_{intra-\text{io}}(\mathbf{x}_{intra};\mathbf{M}_{ss}) = \lambda_{\text{class}}\mathcal{L}_{\text{class}}(\mathbf{x}_{intra};\mathbf{M}_{ss}) + \lambda_{\text{matrix}}\mathcal{L}_{\text{matrix}}(\mathbf{x}_{intra};\mathbf{M}_{ss}) . \tag{14}$$

In order to learn the implicit keypoint matching and its corresponding multimodal inter outlier rejection network in Fig 1, we apply the Dice loss $\mathcal{L}_{\text{Dice}}(\mathcal{S}_{\text{src}}^{w}, \mathcal{S}_{\text{tgt}})$ between a warped source vessel segmentation $\mathcal{S}_{\text{src}}^{w} = T(\mathcal{S}_{\text{src}}, \mathbf{M}_{inter})$ and a target vessel segmentation $\mathcal{S}_{\text{tgt}}$ in addition to the classification loss $\mathcal{L}_{\text{class}}(\mathbf{x}_{inter};\mathbf{M}_{inter})$ of (eq. 11) where $\mathbf{M}_{inter}$ is the predicted transformation matrix in (eq. 5) given $\mathbf{x}_{inter}$. The Dice coefficient is commonly used to evaluate the registration accuracy where higher Dice coefficient indicates more overlap between two segmentation maps. Furthermore, the Dice coefficient is shown to have the best correlation to subjective score from ophthalmologist [62]. We define the Dice loss as one minus the soft Dice coefficient on the aligned vessel segmentation. The Dice coefficient for binary segmentation is defined as

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

(15)

where ⊙ denotes element-wise product and its value is between [0,1]. Since the binary Dice coefficient is not differentiable, the differentiable soft Dice coefficient [18] is applied for Dice loss for grayscale segmentation maps:

$$\mathscr{L}_{\text{Dice}}\left(\mathscr{S}_{\text{src}}^{w}, \mathscr{S}_{\text{tgt}}\right) = 1 - \text{Dice}_{s}\left(\mathscr{S}_{\text{src}}^{w}, \mathscr{S}_{\text{tgt}}\right),$$
$$\text{where Dice}\left(\mathscr{I}_{1}, \mathscr{I}_{2}\right) = \frac{2 \times \sum \text{ele\_min}\left(\mathscr{I}_{1}, \mathscr{I}_{2}\right)}{\sum \mathscr{I}_{1} + \sum \mathscr{I}_{2}} \tag{16}$$

and ele_min denotes the element-wise minimum. The total loss of the inter outlier rejection network is defined as

$$\mathscr{L}_{\text{inter-io}}\left(\mathbf{x}_{inter}, \mathscr{S}_{\text{src}}, \mathscr{S}_{\text{tgt}}; \mathbf{M}_{inter}\right) = \lambda_{\text{class}}^{*} \mathscr{L}_{\text{class}}\left(\mathbf{x}_{inter}; \mathbf{M}_{inter}\right) + \lambda_{\text{Dice}} \mathscr{L}_{\text{Dice}}\left(\mathscr{S}_{\text{src}}^{w}, \mathscr{S}_{\text{tgt}}\right). \tag{17}$$

IV. Experiments

We compare the overall performance of the proposed self-supervised framework to several existing methods including three conventional, four supervised, and two unsupervised methods for multimodal retinal image registration. The experimental results show that the proposed self-supervised method achieves comparable performance comparing to the state-of-the-art results from the previous supervised learning approaches. We perform experiments on two multimodal retinal datasets: *CF-IR dataset* to register color fundus (CF) source images to infrared reflectance (IR) target images and *CF-FA dataset* to register CF source images to fluorescein angiography (FA) target images.

A. CF-IR Dataset

1) Dataset: The first dataset collected by Jacobs Retina Center (JRC) at Shiley Eye Institute consists of CF images (RGB, 3000 × 2672) for source and IR images (grayscale, 768 × 768 or 1536 × 1536) for target. The image pairs contain a variety of pathologies including diabetes, hemorrhages, and macular degeneration. We partition 873 CF-IR pairs into 530 pairs for the training set, 90 for the validation set, and 253 for the test set. Ophthalmologists classify the quality of each image as good, usable, and bad as listed in Table I.

2) Criteria: The robustness of registration is measured by the success rate, which is determined by the criterion that the maximum error (MAE) in (eq. 18) is less than or equal to 10 pixels [8], [9], [10], [11], [23] on 6 manually labeled correspondences $\mathscr{P}$ and the ground-truth transformation matrix $\mathbf{M}_{\text{gt}}$ [13].

$$\text{MAE} = \max_{\mathbf{p} \in \mathscr{P}} \left\| T\left(T\left(\mathbf{p}, \mathbf{M}_{\text{gt}}^{-1}\right), \mathbf{M}_{inter}\right) - \mathbf{p} \right\| \tag{18}$$

Note that we do not use any manually labeled correspondences and the ground-truth transformation matrix $\mathbf{M}_{gt}$ for training and validation. These are only used to determine the success rate to evaluate performance. The accuracy of registration is also measured by the Dice coefficient in (eq. 15) on the binary segmentation maps of aligned images regardless of success registration or not. The binary segmentation maps are obtained via *ConvSeg or PACSeg* vessel segmentation network with 0.5 threshold. We multiply two segmentation maps with a valid mask to compute the Dice coefficient in their overlapping region.

3) Implementation: All images are first padded to square shape and resized to 768 × 768 before applying for registration test. Since directly downsampling the source images could increase the noise and adversely affect the segmentation, the images are anti-aliased with Gaussian filter before bicubic downsampling. In our method, all pixel intensities are normalized between [0, 1], and the target grayscale images are converted to 3 channels by stacking the input channel 3 times.

4) Training: We use the pretrained models of [12], [13] for the vessel segmentation network denoted as *ConvSeg* and *PACSeg*, respectively and fix the network parameters during the training. Interested readers should refer to the [12], [13] for more details. The self-supervised feature detection and descriptor network, the multimodal intra and inter outlier rejection networks, and intra and inter matrix estimation are trained in an end-to-end fashion. All networks of the proposed framework are trained in PyTorch using Adam optimizer, and the best model is chosen with the lowest Dice loss on the validation set. The learning rate 10^{-4} with 10% decays every 200 epochs up to 10^{-5} and the batch size is set to be 4. We set the maximum epochs to be 3,000 and the best loss weights including λ_{class}, λ_{matrix}, and λ_{Dice} in eqs. (10), (14), and (17) are searched experimentally as listed in Table V. All the coordinates are normalized within [−1, 1], and the transformation matrices are modified accordingly. The margin of the triplet loss in (eq. 8) m and the threshold Th in (eq. 6) are set to be 1 and 4, respectively.

5) Comparison: To compare the performance, we consider three conventional (not learning-based) methods such as SURF-PIIFD-RPM [9], URSIFT-PIIFD-AGMM [10], and Phase-HOG-RANSAC [2] and four supervised deep-learning methods including IRR-PWC [39], our previous works ICASSP [12] and TIP [13], and two unsupervised methods DLIR [4] and Voxelmorph [41]. We use the original authors' MATLAB code for SURF-PIIFD-RPM [9] and URSIFT-PIIFD-AGMM [10] and our own implementation in MATLAB for the affine registration part of phase-HOG-RANSAC [2]. For IRR-PWC [39], we use the pretrained model on FlyingThings3D dataset [63] in PyTorch and fine-tune the model using Adam optimizer with learning rate 10^{-4}, batch size 1, and weight decay 10^{-4} for 100 max epochs. We test IRR-PWC [39] with the pretrained model and the fine-tuned model on our test set, which are denoted as "IRR-PWC (pre-trained)" and "IRR-PWC (fine-tuned)", respectively. We also use publicly available authors' code for DLIR [4] and Voxelmorph [41] to train and test for our dataset. Moreover, we modify Voxelmorph [41] to register 2D image instead of 3D data. Specifically, we compare the performance of the proposed method (*ConvSeg*) with that of our ICASSP [12], which proposed the convolutional segmentation. We also compare the performance of the proposed method (*PACSeg*) with that of our TIP [13], which proposed the content adaptive segmentation. Furthermore, we present the proposed endto-end framework, which consists of the keypoint detection and outlier rejection networks, achieves significantly better performance than the keypoint detection network with the conventional RANSAC method to replace the multimodal inter and intra outlier rejection network, which is denoted as the proposed method (*RANSAC*).

6)Results and Discussion: The proposed methods are compared with three conventional methods, four supervised, and two unsupervised deep-learning methods in Table II for the quantitative result and in Fig. 6 for the qualitative result. The criterion of MAE ≤ 10 pixels in (eq. 18) is used to determine successful registration in this experiment and the average Dice coefficient is expressed as mean and (± standard deviation). In Table II, we first compute the Dice coefficient of the original input images without any registration denoted as *no registration for* baseline.

a) Conventional (no learning) methods: SURF-PIIFD-RPM [9] achieves 27.27% success rate and an average of 0.262 Dice coefficient on the entire test set. The performance improves after excluding bad quality images, that is, in case of both source and target image are “good” or “usable”, but degrades in case of either source or target image is “bad” where its success rate is significantly low at 4.0%. URSIFT-PIIFD-AGMM [10] performs better than SURF-PIIFD-RPM [9] in some cases as in Fig. 6, but the overall result is worse than [9] as shown in Table II. Phase-HOG-RANSAC [2] achieves the highest success rate 40.32% and Dice coefficient 0.331 among three conventional methods whereas 14.00% success rate on the bad quality images indicates very limited robustness.

b) Supervised learning methods: The pretrained IRR-PWC [39] fails on every image pair, while the fine-tuned IRR-PWC reaches 1.19% success rate and 0.096 Dice coefficient on average. Since the original input image pairs show large resolution gaps, a wide field of view difference, and large quality variations, applying only a deformable registration method does not yield accurate results. ICASSP [12], which utilizes the convolutional vessel segmentation, achieves significantly higher performance comparing to three conventional methods and IRR-PWC at 86.56% success rate, 0.592 average Dice coefficient. Our TIP [13] further improves and clearly achieves the highest success rate 97.63% and average Dice coefficient 0.631 with the improvement of 11.07% and 0.039 comparing to the results in ICASSP [12]. TIP [13] ranks the highest and reaches 99.51% success rate after excluding bad quality images (both source and target image are good or usable quality).

c) Unsupervised learning methods: Although we perform network parameters and various hyperparameters including loss functions, both DLIR [4] and Voxelmorph [41] fail to provide any alignment. DLIR [4] achieves slightly higher Dice coefficient of 0.090, comparing to Voxelmorph [41]’s 0.081. Since the appearance of input image pairs is quite different with resolution, wide field of view, and image quality, DLIR [4], which learns directly the transformation parameters from the input images in a unsupervised manner, fails to align multimodal retina images with large difference in field of view. Voxelmorph [41], which proposed only for single modal deformable registration, also fails all the alignment for our datasets because the deformable registration algorithms are mainly proposed to align small misalignment.

d) Proposed self-supervised learning method: Although the supervised learning method [13] achieves the highest performance as in Table II, it requires manually labeled correspondences to prepare the ground truth of the training set such as keypoints and transform matrices, and fine-tunes a pretrained model on a unrelated dataset with a target dataset. Consequently, a lot of time and effort are needed for the supervised method. Moreover, we might need to do all procedure again for a different target dataset. In this paper, we register multi-modal retinal images via a self-supervised method to alleviate all these burdens and aim to automatically register multi-modal images without any human annotations. The proposed method with two types of segmentation networks, which are denoted as *ConvSeg* and *PACSeg*, have better performance comparing to the work in ICASSP [12]. The performance of *PACSeg* is higher than that of *ConvSeg* regardless of image quality. It indicates that the content adaptive vessel segmentation of [13] is helpful even for the self-supervised learning. However, the performance gain (1.98 success rate) is not significant comparing to that of the supervised methods (11.07). For high-quality images, the proposed self-supervised learning *PAC Seg* shows comparable performance to TIP [13]. However, TIP [13] achieves better performance than *PACSeg* for images with bad quality. In practice, this is not a critical issue since the bad quality images would not be used for diagnosis. Thus, the proposed self-supervised learning method achieves comparable performance to the state-of-the-art supervised learning approach [13] as shown in Table II. Furthermore, the proposed self-supervised learning method is significantly better than the keypoint detection network with the conventional RANSAC (*RANSAC*), which only achieves 41.50% success rate. Especially, the conventional RANSAC is much worse at the difficult dataset (bad dataset).

Fig. 6 (1)–(3) show qualitative registration results among conventional, supervised learning, unsupervised learning, and self-supervised learning methods for three image pairs. For each image pair, sub-images (a) and (b) are the input images resized to 768 × 768. Sub-image (c) presents the checkerboard overlay of the aligned images, where the RGB and gray tiles show the warped CF source image and the target IR image, respectively. Note that the checkerboard image can easily visualize the quality of registration since the vessels of source and target images should be continuous across the tiles if two images are well aligned. Sub-image (d) shows the overlay of the aligned vessel segmentation, where the source segmentation and the target segmentation are assigned to the red channel and the green channel, respectively. The vessel channel image is also effective to visualize qualitative registration because the vessels look yellow if segmentation maps overlap accurately. For the first image pair (1), which is classified as good quality for both the source CF image (1a) and the target IR image (1b), two conventional methods [2], [9], two supervised methods [12], [13], and the proposed methods succeed while the other methods including URSIFT-PIIFD-AGMM [10] and fine-tuned IRR-PWC [39] fail since their MAEs in (eq. 18) are larger than 10 pixels as illustrated from (1c) and (1d). For the second example (2), where both images include some disease and the CF image is blur, the supervised methods of [12], [13] and the proposed methods succeed while the other methods fail. For the third image pair (3) where the optic disk in CF image is located near the center of the image, only the supervised method of TIP [13] can successfully align the images. Note that since the number of keypoint correspondences of [9], [10] is insufficient to calculate an affine transformation matrix (fewer than 3), their results are set to be the same as before registration.

B. CF-FA Dataset

We also test the performance of the proposed framework on the public dataset [64], which consists of 59 pairs of color fundus (720 × 576, RGB) and fluorescein angiography (720 × 576, grayscale) images. In this dataset, 29 pairs are normal while the other 30 pairs of images have diabetic retinopathy. The field of view of CF and FA images are similar and there are no images with poor quality. We manually labeled 6 pairs of keypoint correspondences to obtain the ground-truth transformation matrix for all image pairs. We fine-tune the supervised deep learning methods [12], [13], [39] on this dataset using similar hyper parameters as in the JRC CF-IR dataset (reducing batch size to 30 for outlier rejection network) for comparison and apply the same parameters to fine-tune the self-supervised proposed methods. 30 pairs with odd indices are used for training and the other 29 pairs with even indices are applied for testing. Due to the small number of images, we do not use a validation set and simply stop training at the maximum epoch.

Table III shows the experimental results of three conventional, four supervised deep-learning, and the proposed methods for the CF-FA dataset [64] where the success rate and Dice coefficient are evaluated in the same way as in the JRC CF-IR dataset. Most methods achieve higher success rate compared to JRC CF-IR dataset overall due to similar field of view of source and target images and better image quality. The proposed methods as well as Phase-HOG-RANSAC [2] and TIP [13] achieve 100% success, while our proposed method *PACSeg* reaches the 2nd highest Dice coefficient at 0.663. Fig. 7 shows one challenging pair whose the overlapping ratio between source and target images is small. SURF-PIIFD-RPM [9], URSIFT-PIIFD-AGMM [10], fine-tuned IRR-PWC [39], DLIR [4] and Voxelmorph [41], and ICASSP [12] fail to align the pair such that they yield MAE larger than 10 pixels whereas Phase-HOG-RANSAC [2], TIP [13], and our proposed methods succeed. The proposed methods produce similar accurate alignment to the supervised method [13], which is observable from the segmentation overlay. The experimental results on CF-FA dataset demonstrate that the proposed framework can be easily generalized for other modalities of the retinal images via fine-tuning on a small training set without drastically adjusting hyper parameters and requiring large dataset. We evaluate testing runtime and memory usage of the conventional, the supervised, the unsupervised, and the proposed self-supervised methods on CPU or GPU devices in Table IV where CPU has 8 cores Intel(R) Core(TM) i7-5960X with 64 GB memory and GPU is a NVidia Tesla M40 with 24 GB memory. All the learning based methods, which run on the GPU, show much shorter runtime than the conventional methods. Among the learning methods, the execution time of the proposed methods takes longer time from 0.2 to 0.97 second. However, the proposed methods still have much shorter runtime compared to the conventional methods. The memory usage of the learning methods, except for the proposed method, is also lower than that of the conventional methods. The proposed methods use the most GPU memory since our architecture is the most complex comparing to the others.

C. Ablation Study

We perform three ablation studies for: seven hyperparameters λ_{pos}, λ_{desc}, λ_{score}, λ_{class}, λ_{matrix}, λ_{class}^{*}, and λ_{Dice} in eqs (10), (14), and (17) and the number of keypoints K, and the number of residual blocks in the outlier rejection network. As listed in Table V, all the seven hyperparameters are set to be a default value 1.0. We first change one hyperparameter from 0.05 to 10.0 and the best candidate parameters are collected based on the success rate or Dice. Note that we mainly show the hyperparameters to achieve better results than the default configuration although we search various parameters in Table V. Next, we use the combinations of the candidate hyperparameters from two to seven. After all the experiments, $\lambda_{pos} = \lambda_{score} = \lambda_{class}^{*} = \lambda_{Dice} = 1.0$, $\lambda_{desc} = 3.0$, $\lambda_{class} = 0.1$, and $\lambda_{matrix} = 5.0$ are the best hyperparameters achieving the highest success rate 97.23. On the other hand, $\lambda_{pos} = \lambda_{score} = \lambda_{matrix} = \lambda_{Dice} = 1.0$, $\lambda_{desc} = 3.0$, $\lambda_{class} = 0.1$, and $\lambda_{class}^{*} = 5.0$ are the best hyperparameters achieving the highest Dice 0.631. We choose the hyperparameters to achieve the highest Dice coefficient since they are more stable.

We also evaluate the performance of the proposed network with various number of keypoints K, and the number of residual blocks of the outlier rejection network along with *ConvSeg* and *PACSeg* vessel segmentation networks on the CF-IR test set in Table VI. We first fix the number of residual blocks, 4 or 8 and configure different number of keypoints K from 300 to 1200 for each vessel segmentation network. 900 keypoints configuration achieves the highest success rate with higher Dice coefficient with PAC vessel segmentation network and 600 keypoints are the best for the highest success rate with Conv vessel segmentation network. In the third ablation study, we choose the best number of keypoints K (900) and set the number of residual blocks as 4, 8, and 12 to evaluate the performance variations. The best number of residual blocks are 4 and 8 for *ConvSeg* and *PACSeg* vessel segmentation network, respectively. We also observe that the Dice coefficient increases along with the number of residual blocks.

V. Conclusion

In this paper, we propose a self-supervised learning framework for the multimodal retinal image registration. The structure of the proposed framework is similar to that of the supervised learning method, which consists of three neural networks for vessel segmentation, feature detection and description, and outlier rejection. From the experimental results, the proposed method achieves similar performance as that of the state-of-the-art supervised learning method without any human-annotated labels or ground truth. The proposed method shows that the self-supervised learning can even be used for the multimodal registration task. In future work, we will concatenate the proposed framework with a self-supervised locally fine alignment method to form a complete pipeline without any labels. In addition, we briefly discuss the multimodal retina registration for clinical applications. From the multimodal retina images of the diabetic patient in Fig. 8, the CF image reveals clearly white lesions as indicated by a yellow arrow in sub-Figure 8(a) and the IR image shows better red legions denoted by a blue arrow in sub-Figure 8(b). The aligned multimodal images can reveal both legions as overlaid in sub-Figure 8(c). The preliminary example shows the potentials of multimodal retina registration for clinical applications.

We will further research on the clinical applications to help ophthalmologists in improving diagnosis accuracy and speed.

Acknowledgment

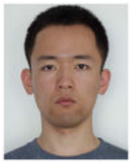
The authors would like to thank Jacobs Retina Center, Shiley Eye Institute for data collection.

This work was supported in part by the National Eye Institute under Grant 1R01EY033847. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou.

Biographies

Cheolhong An received the B.S. and M.S. degrees in electrical engineering from Pusan National University, Busan, South Korea, in 1996 and 1998, respectively, and the Ph.D. degree in electrical and computer engineering in 2008. He is currently an Assistant Adjunct Professor in electrical and computer engineering with the University of California at San Diego. Earlier, he worked at Samsung Electronics, South Korea, and Qualcomm, USA. His current research interests include medical image processing and the real-time bio image processing, 2D and 3D image processing with machine learning, and sensor technology.

Yiqian Wang received the B.S. degree in electrical engineering from the Beijing Institute of Technology, Beijing, China, in 2018. She is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, University of California at San Diego. Her research interests include medical image processing, signal processing, and machine learning.



Junkang Zhang (Member, IEEE) received the B.E. degree in automation from Hohai University, Changzhou, China, in 2014, and the M.E. degree in pattern recognition and intelligent system from Southeast University, Nanjing, China, in 2017. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of California at San Diego, San Diego, CA, USA. His research interests include image processing and computer vision.

Truong Q. Nguyen (Fellow, IEEE) is currently a Distinguished Professor at the ECE Department, University of California at San Diego, San Diego. He is the coauthor (with Prof. Gilbert Strang) of a popular textbook *Wavelets & Filter Banks* (Wellesley-Cambridge Press, 1997). He has over 450 publications, a Google H-index of 66 with over 26K citations. His current research interests are 3D video processing, machine learning with applications in health monitoring/analysis, and 3D modeling.

He received the IEEE Transaction in Signal Processing Paper Award in 1992 and NSF Career Award in 1995. He received the Distinguished Teaching Award at UC San Diego in 2019. He served as an Associate Editor for IEEE Transaction on Signal Processing, IEEE Signal Processing Letters, IEEE transaction on Circuits & Systems, and IEEE Transaction on Image Processing.

References

[1]. MacGillivray TJ, Trucco E, Cameron JR, Dhillon B, Houston JG, and van Beek EJR, “Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions,” Brit. J. Radiol, vol. 87, no. 1040, Aug. 2014, Art. no. 20130832.

[2]. Li Z et al. , “Multi-modal and multi-vendor retina image registration,” Biomed. Opt. Exp, vol. 9, no. 2, pp. 410–422, 2018.

[3]. Rocco I, Arandjelovic R, and Sivic J, “Convolutional neural network architecture for geometric matching,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6148–6157.

[4]. de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, and Išgum I, “A deep learning framework for unsupervised affine and deformable image registration,” Med. Image Anal, vol. 52, pp. 128–143, Feb. 2018 [PubMed: 30579222]

[5]. Ritter N, Owens R, Cooper J, Eikelboom RH, and Van Saarloos PP, “Registration of stereo and temporal images of the retina,” IEEE Trans. Med. Imag, vol. 18, no. 5, pp. 404–418, May 1999.

[6]. Chanwimaluang T, Fan G, and Fransen SR, “Hybrid retinal image registration,” IEEE Trans. Inf. Technol. Biomed, vol. 10, no. 1, pp. 129–142, Jan. 2006. [PubMed: 16445258]

[7]. Yang GH, Stewart CV, Sofka M, and Tsai C-L, “Registration of challenging image pairs: Initialization, estimation, and decision,” IEEE Trans. Pattern Anal. Mach. Intell, vol. 29, no. 11, pp. 1973–1989, Nov. 2007. [PubMed: 17848778]

[8]. Chen J, Tian J, Lee N, Zheng J, Smith RT, and Laine AF, “A partial intensity invariant feature descriptor for multimodal retinal image registration,” IEEE Trans. Biomed. Eng, vol. 57, no. 7, pp. 1707–1718, Jul. 2010. [PubMed: 20176538]

[9]. Wang G, Wang Z, Chen Y, and Zhao W, “Robust point matching method for multimodal retinal image registration,” Biomed. Signal Process. Control, vol. 19, pp. 68–76, May 2015.

[10]. Zhang H, Liu X, Wang G, Chen Y, and Zhao W, “An automated point set registration framework for multimodal retinal image,” in Proc. 24th Int. Conf. Pattern Recognit. (ICPR), Aug. 2018, pp. 2857–2862.

[11]. Lee JA et al. , “A low-dimensional step pattern analysis algorithm with application to multimodal retinal image registration,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1046–1053.

[12]. Wang Y et al. , “A segmentation based robust deep learning framework for multimodal retinal image registration,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 1369–1373.

[13]. Wang Y et al. , “Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework,” IEEE Trans. Image Process, vol. 30, pp. 3167–3178, 2021. [PubMed: 33600314]

[14]. Zhang J, Dashtbozorg B, Bekkers E, Pluim JPW, Duits R, and Romeny BMTH, “Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores,” IEEE Trans. Med. Imag, vol. 35, no. 12, pp. 2631–2644, Dec. 2016.

[15]. Maninis K-K, Pont-Tuset J, Arbeláez P, and Van Gool L, “Deep retinal image understanding,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent New York, NY, USA: Springer, 2016, pp. 140–148.

[16]. Jin Q, Meng Z, Pham TD, Chen Q, Wei L, and Su R, “DUNet: A deformable network for retinal vessel segmentation,” Knowl.-Based Syst, vol. 178, pp. 149–162, Aug. 2019.

[17]. Li L, Verma M, Nakashima Y, Nagahara H, and Kawasaki R, “IterNet: Retinal image segmentation utilizing structural redundancy in vessel networks,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020.

[18]. Zhang J et al. , “Joint vessel segmentation and deformable registration on multi-modal retinal images based on style transfer,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2019, pp. 839–843.

[19]. Harris CG et al. , “A combined corner and edge detector,” in Proc. Alvey Vis. Conf, vol. 15, 1988, p. 5244.

[20]. Dalal N and Triggs B, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2005, pp. 886–893.

[21]. Lowe DG et al. , “Object recognition from local scale-invariant features,” in Proc. ICCV, vol. 99, Sep. 1999, pp. 1150–1157.

[22]. Bay H, Tuytelaars T, and Van Gool L, “SURF: Speeded up robust features,” in Proc. Eur. Conf. Comput. Vis New York, NY, USA: Springer, 2006, pp. 404–417.

[23]. Lee J, Liu P, Cheng J, and Fu H, “A deep step pattern representation for multimodal retinal image registration,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 5077–5086.

[24]. Yi KM, Trulls E, Lepetit V, and Fua P, “LIFT: Learned invariant feature transform,” in Proc. Eur. Conf. Comput. Vis New York, NY, USA: Springer, 2016, pp. 467–483.

[25]. Choy CB, Gwak J, Savarese S, and Chandraker M, “Universal correspondence network,” in Proc. Adv. Neural Inf. Process. Syst, 2016, pp. 2414–2422.

[26]. DeTone D, Malisiewicz T, and Rabinovich A, “SuperPoint: Self-supervised interest point detection and description,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 224–236.

[27]. Fischler MA and Bolles R, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” Commun. ACM, vol. 24, no. 6, pp. 381–395, 1981.

[28]. Rousseeuw PJ, “Least median of squares regression,” J. Amer. Statist. Assoc, vol. 79, no. 388, pp. 871–880, 1984.

[29]. Chum O and Matas J, “Matching with PROSAC—Progressive sample consensus,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Jun. 2005, pp. 220–226.

[30]. Chum OR and Matas J, “Optimal randomized RANSAC,” IEEE Trans. Pattern Anal. Mach. Intell, vol. 30, no. 8, pp. 1472–1482, Aug. 2008. [PubMed: 18566499]

[31]. Raguram R, Chum O, Pollefeys M, Matas J, and Frahm J-M, “USAC: A universal framework for random sample consensus,” IEEE Trans. Pattern Anal. Mach. Intell, vol. 35, no. 8, pp. 2022–2038, Aug. 2013. [PubMed: 23787350]

[32]. Tsai CL, Li C-Y, Yang G, and Lin K-S, "The edge-driven dual-bootstrap iterative closest point algorithm for registration of multimodal fluorescein angiogram sequence," IEEE Trans. Med. Imag, vol. 29, no. 3, pp. 636–649, Mar. 2010.

[33]. Ghassabi Z, Sedaghat A, Shanbehzadeh J, and Fatemizadeh E, "An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors," EURASIP J. Image Video Process, vol. 2013, no. 1, p. 25, 2013.

[34]. Brachmann E et al. , "DSAC-differentiable RANSAC for camera localization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, Jul. 2017, pp. 6684–6692.

[35]. Yi KM, Trulls E, Ono Y, Lepetit V, Salzmann M, and Fua P, "Learning to find good correspondences," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, Jun. 2018, pp. 2666–2674.

[36]. Dosovitskiy A et al. , "FlowNet: Learning optical flow with convolutional networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2758–32766.

[37]. Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, and Brox T, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2462–2470.

[38]. Sun D, Yang X, Liu M-Y, and Kautz J, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, Jun. 2018, pp. 8934–8943.

[39]. Hur J and Roth S, "Iterative residual refinement for joint optical flow and occlusion estimation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5754–5763.

[40]. Chee E and Wu Z, "AIRNet: Self-supervised affine registration for 3D medical images using neural networks," 2018, arXiv:1810.02583.

[41]. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, and Dalca AV, "VoxelMorph: A learning framework for deformable medical image registration," IEEE Trans. Med. Imag, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.

[42]. de Vos BD, Berendsen FF, Viergever MA, Staring M, and Išgum I, "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. New York, NY, USA: Springer, 2017, pp. 204–212.

[43]. Hu Y et al. , "Weakly-supervised convolutional neural networks for multimodal image registration," Med. Image Anal, vol. 49, pp. 1–13, Oct. 2018 [PubMed: 30007253]

[44]. Canalini L, Klein J, Miller D, and Kikinis R, "Segmentation-based registration of ultrasound volumes for glioma resection in image-guided neurosurgery," Int. J. Comput. Assist. Radiol. Surg, vol. 14, pp. 1697–1713, Aug. 2019 [PubMed: 31392670]

[45]. Sui X, Zheng Y, Jiang Y, Jiao W, and Ding Y, "Deep multispectral image registration network," Computerized Med. Imag. Graph, vol. 87, Jan. 2021, Art. no. 101815.

[46]. Mahapatra D, "GAN based medical image registration," 2018, arXiv:1805.02369.

[47]. Mahapatra D, Antony B, Sedai S, and Garnavi R, "Deformable medical image registration using generative adversarial networks," in Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI), Apr. 2018, pp. 1449–1453.

[48]. Windsor R, Jamaludin A, Kadir T, and Zisserman A, "Self-supervised multi-modal alignment for whole body medical imaging," 2021, arXiv:2107.06652.

[49]. van den Oord A, Li Y, and Vinyals O, "Representation learning with contrastive predictive coding," 2018, arXiv:1807.03748.

[50]. Gidaris S, Singh P, and Komodakis N, "Unsupervised representation learning by predicting image rotations," in Proc. 6th Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, Apr./May 2018, pp. 1–16.

[51]. Noroozi M and Favaro P, "Unsupervised learning of visual representations by solving jigsaw puzzles," in Proc. ECCV, vol. 6, 2016, pp. 69–84.

[52]. Gisbert G, Dey N, Ishikawa H, Schuman J, Fishbaugh J, and Gerig G, "Self-supervised denoising via diffeomorphic template estimation: Application to optical coherence tomography," in Ophthalmic Medical Image Analysis (Lecture Notes in Computer Science). New York, NY, USA: Springer, 2020, pp. 72–82.

[53]. Luo Y, Pan J, Fan S, Du Z, and Zhang G, “Retinal image classification by self-supervised fuzzy clustering network,” IEEE Access, vol. 8, pp. 92352–92362,2020.

[54]. Su H, Jampani V, Sun D, Gallo O, Learned-Miller E, and Kautz J, “Pixel-adaptive convolutional neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 11166–11175.

[55]. Felsberg M and Sommer G, “The monogenic signal,” IEEE Trans. Signal Process, vol. 49, no. 12, pp. 3136–3144, Dec. 2001

[56]. Cifor A, Risser L, Chung D, Anderson EM, and Schnabel JA, “Hybrid feature-based diffeomorphic registration for tumor tracking in 2-D liver ultrasound images,” IEEE Trans. Med. Imag, vol. 32, no. 9, pp. 1647–1656, Sep. 2013.

[57]. Wong A, Clausi DA, and Fieguth P, “CPOL: Complex phase order likelihood as a similarity measure for MR–CT registration,” Med. Image Anal, vol. 14, no. 1, pp. 50–57, Feb. 2010. [PubMed: 19892585]

[58]. Tang J, Kim H, Guizilini V, Pillai S, and Ambrus R, “Neural outlier rejection for self-supervised keypoint learning,” in Proc. Int. Conf. Learn. Represent, 2020, pp. 1–14.

[59]. Hviid Christiansen P, Fly Kragh M, Brodskiy Y, and Karstoft H, “UnsuperPoint: End-to-end unsupervised interest point detector and descriptor,” 2019, arXiv:1907.04011.

[60]. Shi W et al. , “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1874–1883.

[61]. Brachmann E and Rother C, “Neural-guided RANSAC: Learning where to sample model hypotheses,” in Proc. ICCV, 2019, pp. 4322–4331.

[62]. Wang Y et al. , “Study on correlation between subjective and objective metrics for multimodal retinal image registration,” IEEE Access, vol. 8, pp. 190897–190905, 2020.

[63]. Mayer N et al. , “A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4040–4048.

[64]. Hajeb Mohammad Alipour S, Rabbani H, and Akhlaghi MR, “Diabetic retinopathy grading by digital curvelet transform,” Comput. Math. Methods Med, vol. 2012, Sep. 2012, Art. no. 761901.

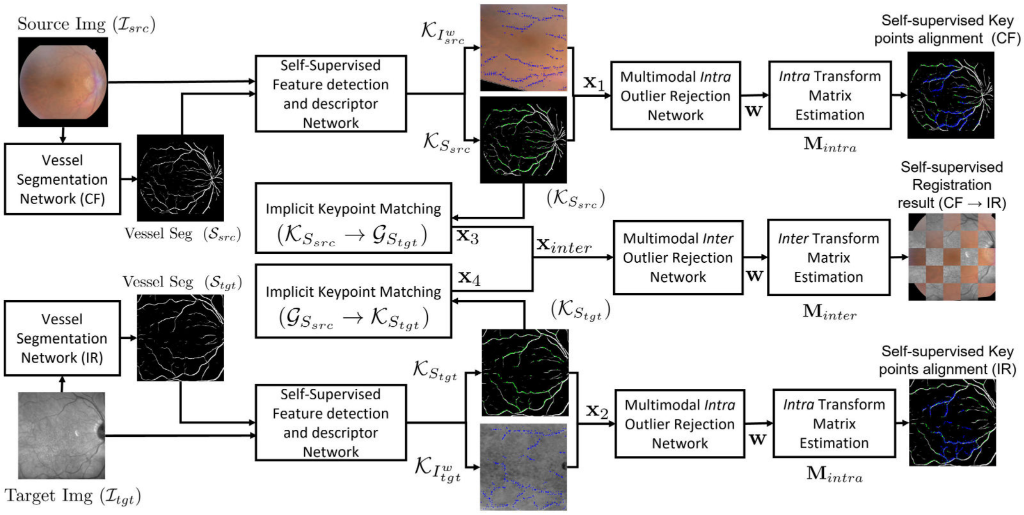

Source Img $(\mathcal{I}_{src})$
Vessel Segmentation Network (CF)
Vessel Seg $(\mathcal{S}_{src})$
Self-Supervised Feature detection and descriptor Network
$\mathcal{K}_{I^w_{src}}$
$\mathcal{K}_{S_{src}}$
$\mathbf{x}_1$
Multimodal *Intra* Outlier Rejection Network
$\mathbf{w}$
Intra Transform Matrix Estimation
$\mathbf{M}_{intra}$
Self-supervised Key points alignment (CF)
Implicit Keypoint Matching $(\mathcal{K}_{S_{src}} \rightarrow \mathcal{G}_{S_{tgt}})$
$\mathbf{x}_3$
$(\mathcal{K}_{S_{src}})$
$\mathbf{x}_{inter}$
Multimodal *Inter* Outlier Rejection Network
$\mathbf{w}$
Inter Transform Matrix Estimation
$\mathbf{M}_{inter}$
Self-supervised Registration result (CF → IR)
Implicit Keypoint Matching $(\mathcal{G}_{S_{src}} \rightarrow \mathcal{K}_{S_{tgt}})$
$\mathbf{x}_4$
$(\mathcal{K}_{S_{tgt}})$
Vessel Seg $(\mathcal{S}_{tgt})$
Vessel Segmentation Network (IR)
$\mathcal{K}_{S_{tgt}}$
Self-Supervised Feature detection and descriptor Network
$\mathcal{K}_{I^w_{tgt}}$
$\mathbf{x}_2$
Multimodal *Intra* Outlier Rejection Network
$\mathbf{w}$
Intra Transform Matrix Estimation
$\mathbf{M}_{intra}$
Self-supervised Key points alignment (IR)
Target Img $(\mathcal{I}_{tgt})$

Fig. 1.

The proposed registration framework takes a source image I_{src} and its target image I_{tgt} and generates their corresponding vessel segmentations S_{src} and S_{tgt}. Next, the self-supervised feature detection and descriptor network find keypoint pairs $\left\{\left(\mathscr{K}_{I^w_{src}}, \mathscr{K}_{S_{src}}\right), \left(\mathscr{K}_{I_{tgt}}, \mathscr{K}_{S_{tgt}}\right)\right\}$ between a warped image and segmentation for both source and target images, where they are denoted as blue and green keypoints and they are overlayed onto an image and a vessel segmentation before and after registration, respectively. The multimodal intra outlier rejection network takes the Top-K pairs of keypoints $\mathbf{x}_1$ and $\mathbf{x}_2$ for source and target to estimate the intra transform matrix $\mathbf{M}_{intra}$. For the multimodal inter registration, the keypoints of the source and target segmentation $\mathscr{K}_{S_{sr}}$ and $\mathscr{K}_{S_{tgt}}$ are implicitly aligned into the grid-points of the source and target segmentation $\mathscr{G}_{S_{src}}$ and $\mathscr{G}_{S_{tgt}}$. The multimodal inter outlier rejection network takes the Top-K pairs of keypoints $\mathbf{x}_3$ and $\mathbf{x}_4$ to estimate the inter transform matrix $\mathbf{M}_{inter}$.

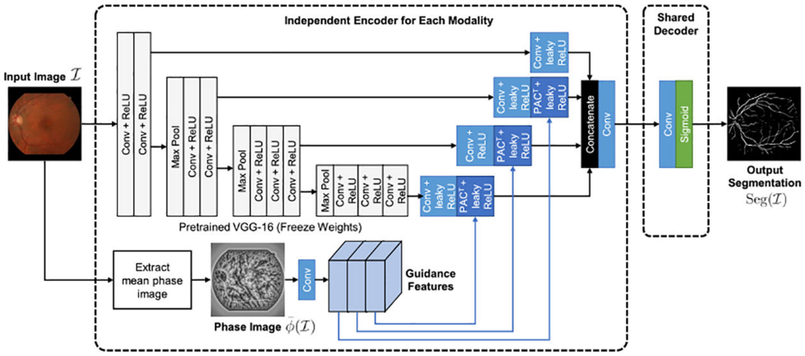


Fig. 2.

Structure of the content-adaptive vessel segmentation network [13] where PAC^T stands for transposed pixel-adaptive convolution [54].

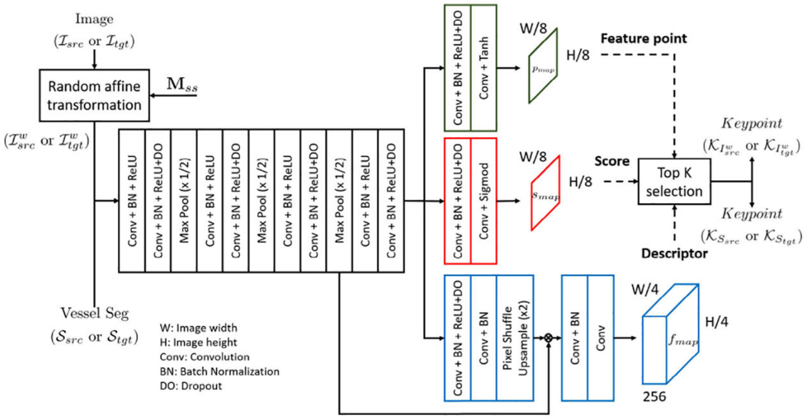


Fig. 3.

Structure of the self-supervised feature detection and description network for the multimodal keypoint learning. The warped source and target images I^w_{src}, I^w_{tgt} are generated after applying the random transformation matrix $\mathbf{M}_{ss}$. The keypoint detection and description network estimates top-K key points of the vessel segmentation and warped images for source and target, which are denoted as $(\mathscr{K}_{I^w_{src}}, \mathscr{K}_{S_{src}})$ and $(\mathscr{K}_{I^w_{tgt}}, \mathscr{K}_{S_{tgt}})$, respectively.

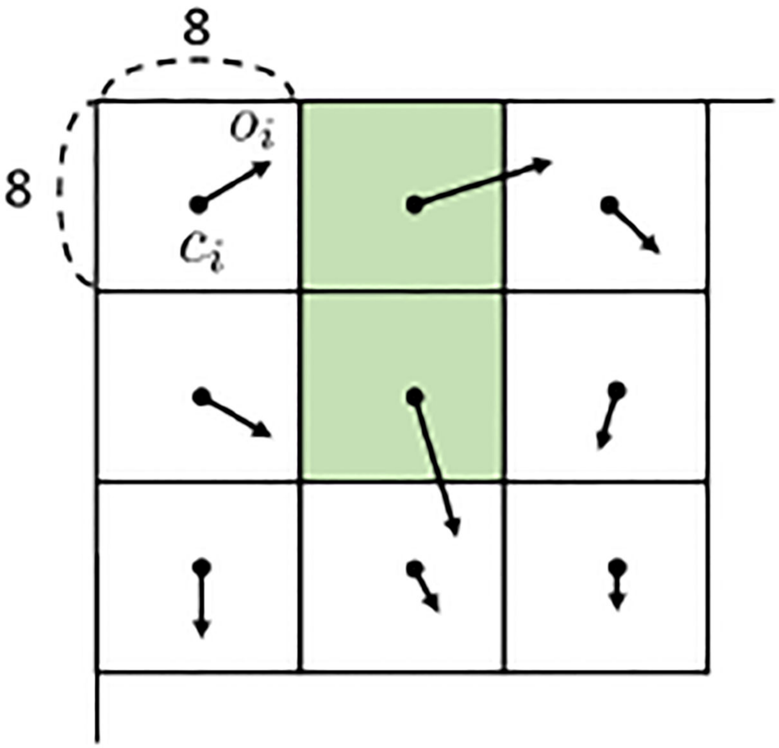


Fig. 4.
8 × 8 grid prediction scheme for the feature positions.

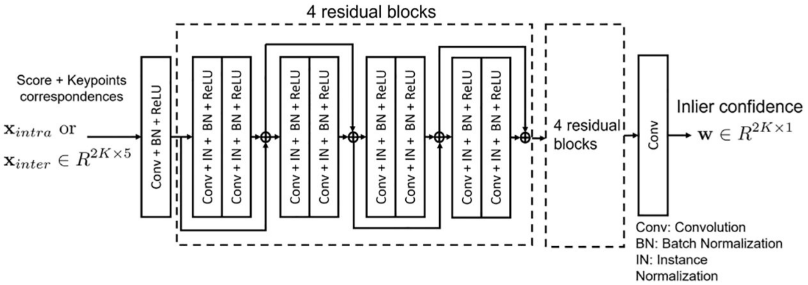


Fig. 5.

Structure of the outlier rejection network for the intra and inter alignment

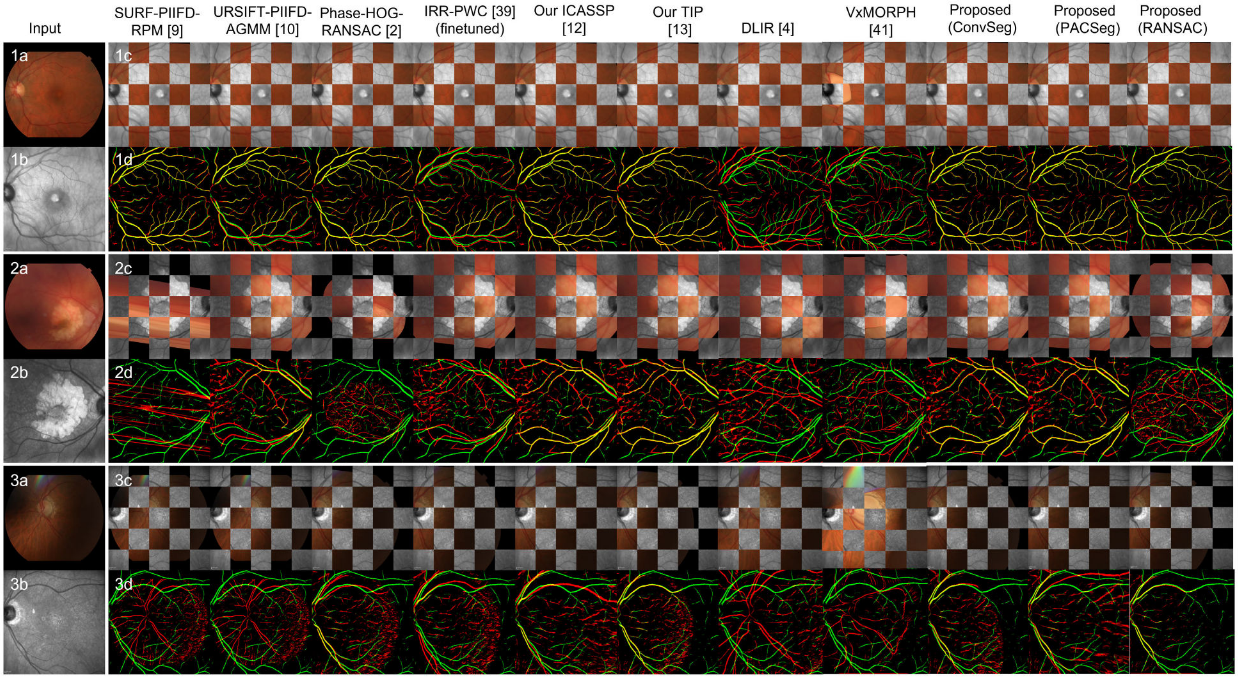


Fig. 6.

Registration results of three example pairs in the CF-IR test set using three conventional, three supervised learning, and two self-supervised learning methods. (a) and (b) show the input image pair resized to 768 × 768. (c) shows the checkerboard of the aligned original images (RGB tiles: warped source image, gray tiles: target image). (d) shows the vessel segmentation overlay (red: warped source segmentation, green: target segmentation, yellow: overlap).

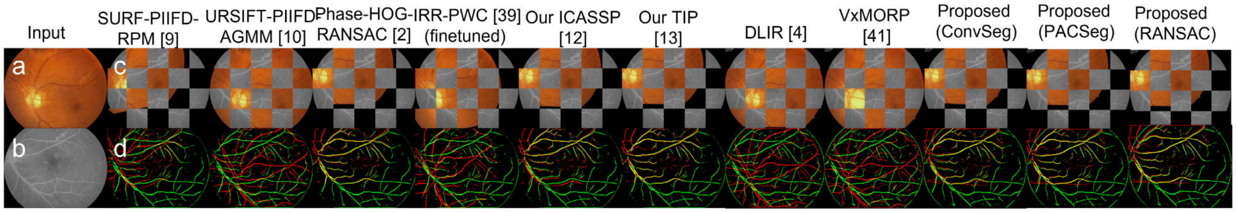


Fig. 7.

Registration results of one challenging pair in CF-FA dataset using different methods. (a) and (b) show the input image pair. (c) shows the checkerboard of the aligned original images (RGB tiles: warped source image, gray tiles: target image). (d) shows the vessel segmentation overlay (red: warped source segmentation, green: target segmentation, yellow: overlap).

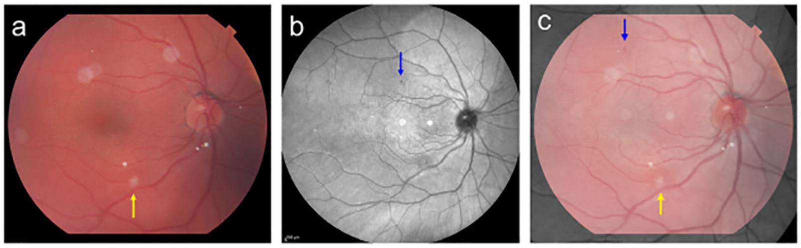


Fig. 8.

Potential clinical application: the white lesion (yellow arrow) is clearly visualized in the CF image (a) and the red lesion (blue arrow) is more visible in the IR image (b). Two lesions are clearly visible in the overlay image (c).

TABLE I

Number of Good, Usable, and Bad Quality Images in the CF-IR Dataset

Dataset	Good	Usable	Bad	Total
Training CF	256	193	81	530
Training IR	327	168	35	530
Validation CF	47	30	13	90
Validation IR	58	23	9	90
Test CF	124	90	40	253
Test IR	181	59	14	253

TABLE II

Result of Various Methods With Three Image Qualities on the CF-IR Test Set Where C, S, US, and SS of M Column Stand for Conventional, Supervised, Unsupervised, And Self-Supervised Method, Respectively. The Best Results Are Colored in Red and the 2nd Best Are Marked In Bold

Method		All images (good + usable + bad)		Exclude bad images (good + usable)		Bad images only	
M	**Name**	**Success rate**	**Dice**	**Success rate**	**Dice**	**Success rate**	**Dice**
–	No registration	–	0.078 (±0.018)	–	0.060 (±0.015)	–	0.072 (±0.019)
C	SURF-PDFD-RPM [9]	27.27% (69/253)	0.262 (±0.245)	33.00% (67/203)	0.293 (±0.256)	4.00% (2/50)	0.134 (±0.133)
C	URSIFT-PIIFD-AGMM [10]	24.90% (63/253)	0.248 (±0.238)	30.05% (61/203)	0.282 (±0.249)	4.00% (2/50)	0.113 (±0.106)
C	Phase-HOG-RANSAC [2]	40.32% (102/253)	0.331 (±0.262)	46.80% (95/203)	0.372 (±0.262)	14.00% (7/50)	0.162 (±0.181)
S	IRR-PWC [39] (pretrained)	0.00% (0/253)	0.059 (±0.018)	0.00% (0/203)	0.060 (±0.018)	0.00% (0/50)	0.055 (±0.021)
S	IRR-PWC [39] (fine-tuned)	1.19% (3/253)	0.096 (±0.060)	1.48% (3/203)	0.102 (±0.064)	0.00% (0/50)	0.073 (±0.033)
S	Our ICASSP [12] (ConvSeg)	86.56% (219/253)	0.592 (±0.168)	95.57% (194/203)	0.643 (±0.108)	50.00% (25/50)	0.386 (±0.204)
S	Our TIP [13] (PACSeg)	97.63% (247/253)	0.631 (±0.126)	99.51% (202/203)	**0.666** (±0.085)	90.00% (45/50)	0.485 (±0.154)
US	DLIR [4]	0.00% (0/253)	0.090 (±0.021)	0.00% (0/203)	0.093 (±0.019)	0.00% (0/50)	0.081 (±0.021)
US	VxMORPH [41]	0.00% (0/253)	0.081 (±0.021)	0.00% (0/203)	0.084 (±0.020)	0.00% (0/50)	0.066 (±0.018)
SS	Proposed (ConvSeg)	94.46% (239/253)	0.592 (±0.115)	98.03% (199/203)	0.645 (±0.094)	80.00% (40/50)	0.376 (±0.211)
SS	Proposed (PACSeg)	**96.44% (244/253)**	0.631 (±0.128)	99.01% (201/203)	0.668 (±0.078)	**86.00% (43/50)**	**0.480** (±0.171)
SS	Proposed (RANSAC)	41.50% (105/253)	0.338 (±0.254)	48.28% (98/203)	0.376 (±0.253)	14.00% (7/50)	0.181 (±0.193)

TABLE III

Result of Different Methods on CF-FA Test Set Where C, S, US, and SS of M Column Stand for Conventional, Supervised, Unsurpervised, and Self-Supervised Method, Respectively. The Best Results Are Colored in Red And The 2nd Best Are Marked In Bold

M	Method	Success rate	Dice
-	No registration	–	0.120 (±0.023)
C	SURF-PIIFD-RPM [9]	82.76% (24/29)	0.553 (±0.218)
C	URSIFT-PIIFD-AGMM [10]	68.97% (20/29)	0.500 (±0.240)
C	Phase-HOG-RANSAC [2]	100.00% (29/29)	0.648 (±0.100)
S	IRR-PWC [39] (pretrained)	0.00% (0/29)	0.098 (±0.023)
S	IRR-PWC [39] (fine-tuned)	3.44% (1/29)	0.200 (±0.069)
S	Our ICASSP [12] (ConvSeg)	96.55% (28/29)	0.645 (±0.102)
S	Our TIP [13] (PACSeg)	100.00% (29/29)	0.679 (±0.087)
US	DLIR [4]	0.00% (0/29)	0.121 (±0.028)
US	VxMORPH [41]	0.00% (0/29)	0.116 (±0.022)
SS	Proposed (ConvSeg)	100.00% (29/29)	0.656 (±0.096)
SS	Proposed (PACSeg)	100.00% (29/29)	**0.663** (±0.085)
SS	Proposed (RANSAC)	89.66% (26/29)	0.582 (±0.158)

TABLE IV

Testing Runtime and Memory Usage at CPU or GPU Devices Where C, S, US, AND SS OF M Column Stand For Conventional, Supervised, Unsupervised, And Self-Supervised Method, Respectively

M	Method	Time (s)	Device (Memory usage)
C	[9]	16.98	CPU (5.7 GB)
C	[10]	64.89	CPU (5.8 GB)
C	[2]	20.85	CPU (5.3 GB)
S	[39] (pretrained)	1.17	GPU (1.3 GB)
S	[39] (fine-tuned)	1.17	GPU (1.3 GB)
S	[12] (ConvSeg)	1.17	GPU (2.3 GB)
S	[13] (PACSeg)	1.18	GPU (4.8 GB)
US	[4]	0.41	GPU (2.1 GB)
US	[41]	0.42	GPU (4.0 GB)
SS	Proposed (ConvSeg)	1.34	GPU (9.3 GB)
SS	Proposed (PACSeg)	1.38	GPU (9.3 GB)
SS	Proposed (RANSAC)	1.30	GPU (8.9 GB)

TABLE V

Ablation Study With Seven Hyperparameters on the Test Set Where the Default Hyperparameters $\lambda_{pos} = \lambda_{desc} = \lambda_{score} = \lambda_{class} = \lambda_{matrix} = \lambda^{*}_{class} = \lambda_{Dice} = 1.0$. the Candidate Hyperparameters Are Marked in Bold and the Best Hyperparameters Are Colored in Red. We Mainly Show the Hyperparameters to Achieve Better Results Than the Default Configuration Although We Search Various Parameters

Hyperparameters							All images	
λ_{pos}	λ_{desc}	λ_{score}	λ_{class}	λ_{matrix}	λ^{*}_{class}	λ_{Dice}	**Success rate**	**Dice**
1.0	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**95.26%**	0.613 (±0.142)
5.0	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	94.86%	**0.625** (±0.137)
1.0	**3.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**95.65%**	**0.622** (±0.134)
1.0	**1.0**	**3.0**	**1.0**	**1.0**	**1.0**	**1.0**	94.86%	**0.625** (±0.136)
1.0	**1.0**	**1.0**	**0.1**	**1.0**	**1.0**	**1.0**	**96.05%**	**0.624** (±0.126)
1.0	**1.0**	**1.0**	**1.0**	**5.0**	**1.0**	**1.0**	94.86%	**0.619** (±0.139)
1.0	**1.0**	**1.0**	**1.0**	**1.0**	**5.0**	**1.0**	**95.26%**	**0.625** (±0.134)
1.0	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**0.1**	**96.05%**	**0.624** (±0.133)
5.0	1.0	1.0	0.1	1.0	1.0	1.0	95.26%	0.622 (±0.131)
1.0	**3.0**	**1.0**	**0.1**	**1.0**	**1.0**	**1.0**	**96.05%**	**0.625** (±0.132)
1.0	3.0	1.0	0.1	1.0	1.0	0.1	95.65%	0.619 (±0.136)
1.0	1.0	3.0	0.1	1.0	1.0	1.0	95.26%	0.620 (±0.143)
1.0	1.0	1.0	0.1	1.0	5.0	1.0	95.26%	0.613 (±0.133)
1.0	1.0	1.0	1.0	1.0	5.0	0.1	95.26%	0.617 (±0.136)
5.0	3.0	1.0	0.1	1.0	1.0	1.0	94.47%	0.609 (±0.144)
1.0	3.0	3.0	0.1	1.0	1.0	1.0	94.47%	0.611 (±0.139)
1.0	**3.0**	**1.0**	**0.1**	**5.0**	**1.0**	**1.0**	**97.23%**	0.619 (±0.134)
1.0	3.0	1.0	0.1	1.0	5.0	1.0	96.44%	**0.631** (±0.128)
1.0	3.0	1.0	0.1	1.0	1.0	0.1	94.47%	0.616 (±0.139)
5.0	3.0	1.0	0.1	1.0	5.0	1.0	95.26%	0.620 (±0.133)
1.0	3.0	3.0	0.1	5.0	1.0	1.0	95.26%	0.623 (±0.127)
1.0	3.0	3.0	0.1	1.0	5.0	1.0	96.05%	0.622 (±0.132)
1.0	**3.0**	**1.0**	**0.1**	**5.0**	**5.0**	**1.0**	**96.05%**	**0.628** (±0.129)
1.0	3.0	1.0	0.1	5.0	1.0	0.1	95.26%	0.623 (±0.133)
1.0	3.0	1.0	0.1	1.0	5.0	0.1	95.26%	0.623 (±0.132)

Hyperparameters							All images	
λ_{pos}	λ_{desc}	λ_{score}	λ_{class}	λ_{matrix}	λ_{class}^{*}	λ_{Dice}	**Success rate**	**Dice**
5.0	3.0	1.0	0.1	5.0	5.0	1.0	95.26%	0.618 (±0.134)
1.0	3.0	3.0	0.1	5.0	5.0	1.0	94.86%	0.625 (±0.131)
1.0	**3.0**	**1.0**	**0.1**	**5.0**	**5.0**	**0.1**	**96.44%**	**0.622** (±0.130)
5.0	3.0	1.0	0.1	5.0	5.0	0.1	96.05%	0.625 (±0.129)
1.0	3.0	3.0	0.1	5.0	5.0	0.1	96.05%	0.626 (±0.132)
5.0	3.0	3.0	0.1	5.0	5.0	0.1	93.28%	0.612 (±0.141)

TABLE VI

Ablation Study With Various Number of Keypoints K and the Number of Residual Blocks of the Outlier Rejection Network Along With *ConvSeg* and *PACSeg* Vessel Segmentation Networks on the CF-IR Test Set. The Best Network Parameters Are Colored in Red and the Best Results Are Marked in Bold

NETWORK CONFIG			All images (good + usable + bad)		Exclude bad images (good + usable)		Bad images only	
SEG	**K**	**#RESBLK**	**Success rate**	**Dice**	**Success rate**	**Dice**	**Success rate**	**Dice**
Conv	300	4	91.70% (232/253)	0.582 (±0.169)	97.54% (198/203)	0.637 (±0.095)	68.00% (34/50)	0.359 (±0.213)
	600	4	**94.46%** (239/253)	0.592 (±0.115)	98.03% (199/203)	0.645 (±0.094)	**80.00%** (40/50)	0.376 (±0.211)
	900	4	94.07% (238/253)	**0.596** (±0.164)	98.52% (200/203)	0.649 (±0.090)	76.00% (38/50)	**0.381** (±0.213)
	1200	4	94.07% (238/253)	**0.596** (±0.168)	**99.51%** (202/203)	**0.650** (±0.093)	72.00% (36/50)	0.375 (±0.219)
PAC	300	8	93.68% (237/253)	0.591 (±0.172)	98.52% (200/203)	0.621 (±0.111)	74.00% (37/50)	0.365 (±0.212)
	600	8	**96.44%** (244/253)	0.618 (±0.128)	**99.01%** (201/203)	0.621 (±0.121)	**86.00%** (43/50)	0.416 (±0.170)
	900	8	**96.44%** (244/253)	**0.631** (±0.128)	**99.01%** (201/203)	0.668 (±0.078)	**86.00%** (43/50)	**0.480** (±0.171)
	1200	8	95.65% (242/253)	0.620 (±0.133)	**99.01%** (201/203)	**0.670** (±0.081)	82.00% (41/50)	0.463 (±0.179)
Conv	600	4	**94.46%** (239/253)	0.592 (±0.115)	98.03% (199/203)	0.645 (±0.094)	**80.00%** (40/50)	0.376 (±0.211)
	600	8	92.09% (233/253)	0.588 (±0.172)	97.54% (198/203)	0.645 (±0.093)	70.00% (35/50)	0.358 (±0.221)
	600	12	94.07% (238/253)	**0.599** (±0.168)	**98.52%** (200/203)	**0.653** (±0.089)	76.00% (38/50)	**0.378** (±0.221)
PAC	900	4	94.07% (238/253)	0.609 (±0.149)	**99.01%** (201/203)	0.653 (±0.090)	74.00% (37/50)	0.433 (±0.196)
	900	8	**96.44%** (244/253)	**0.631** (±0.128)	**99.01%** (201/203)	**0.668** (±0.078)	**86.00%** (43/50)	**0.480** (±0.171)
	900	12	94.86% (240/253)	0.621 (±0.137)	98.52% (200/203)	0.661 (±0.081)	80.00% (40/50)	0.456 (±0.186)