SuperJunction: Learning-Based Junction Detection for Retinal Image Registration

Yu Wang[1], Xiaoye Wang[1,2], Zaiwang Gu[1], Weide Liu[1], Wee Siong Ng[1], Weimin Huang[1], Jun Cheng[1]*

[1]Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore
[2]Department of Mathematics, Harbin Institute of Technology, Weihai, China
{yu_wang, gu_zaiwang, liu_weide, wsng, wmhuang, cheng_jun}@i2r.a-star.edu.sg, wangxiaoye951@gmail.com

Abstract

Keypoints-based approaches have shown to be promising for retinal image registration, which superimpose two or more images from different views based on keypoint detection and description. However, existing approaches suffer from ineffective keypoint detector and descriptor training. Meanwhile, the non-linear mapping from 3D retinal structure to 2D images is often neglected. In this paper, we propose a novel learning-based junction detection approach for retinal image registration, which enhances both the keypoint detector and descriptor training. To improve the keypoint detection, it uses a multi-task vessel detection to regularize the model training, which helps to learn more representative features and reduce the risk of over-fitting. To achieve effective training for keypoints description, a new constrained negative sampling approach is proposed to compute the descriptor loss. Moreover, we also consider the non-linearity between retinal images from different views during matching. Experimental results on FIRE dataset show that our method achieves mean area under curve of 0.850, which is 12.6% higher than 0.755 by the state-of-the-art method. All the codes are available at https://github.com/samjcheng/SuperJunction.

Introduction

Image registration is to spatially align two or more images into the same coordinate system. It has received much attention in the past years as it is often important to align multiple retinal images for ocular disease diagnosis and management (Lee et al. 2019; Liu et al. 2022; Motta, Casaca, and Paiva 2019; Ding et al. 2023). There are many challenges in retinal image registration. Firstly, large spatial displacements may exist and lead to imaging distortion. Secondly, the images could be taken over time to reveal disease progression, which leads to morphological changes. The quality of the retinal images could be adversely affected by various pathologies or noise. Lastly, the various illumination from different devices and settings often result in non-linear intensity changes.

Retinal image registration methods can be grouped into intensity-based and keypoint-based (or feature-based) (Oliveira and Tavares 2014). The first type of approaches is carried out through optimizing similarity metric such as mutual information (Ritter et al. 1999) and correlation (Chanwimaluang, Fan, and Fransen 2006; Wang et al. 2006) to find parameters of transformation function. However, the performances of mutual information based methods drop significantly when there is low overlapping area (Chanwimaluang, Fan, and Fransen 2006; Chen et al. 2010) or a substantial amount of change in texture or scale (Zana and Klein 1999). Correlation coefficient based methods (Chanwimaluang, Fan, and Fransen 2006) rely on the vascular tree, which restricts it from effective registration of low-quality and disease affected retinal images (Ghassabi et al. 2013). Another method, phase correlation (Wang et al. 2006), is robust to lighting variation, but fails to register images with high translation and content changes (Ritter et al. 1999). Although mathematically well-posed, intensity-based approaches are generally computationally intensive due to the use of entire image content (Ghassabi et al. 2013) especially when image resolution is high, and are susceptible to occlusion, background changes caused by pathologies, and camera pose changes (Chen et al. 2010; Stewart, Tsai, and Roysam 2003).

Keypoint-based approaches are often preferable over intensity-based methods. The keypoint-based image registration usually consists of four main steps: 1) keypoint detection, 2) keypoint description, 3) matching, and 4) transformation estimation. An algorithm of keypoint detection is designed to extract the local distinctive region, and an algorithm of keypoint description is designed to represent the detected local region with invariance to geometric deformations, illumination changes, etc. Detection and description of keypoints have been intensively studied (Rosten and Drummond 2005; Leutenegger, Chli, and Siegwart 2011; Alahi, Ortiz, and Vandergheynst 2012; Lowe 2004; Lee et al. 2015). One of the most successful and well-known keypoint-based approaches is the Scale Invariant Feature Transform (SIFT) (Ng and Henikoff 2003), which detects keypoints using scale-space derivatives. Descriptor vectors are constructed on the same scale space to the detector, and the local area of detected keypoint is divided into non-overlapping subareas to support formation of the final descriptor. Since the success of SIFT, many similar approaches have been developed to improve the robustness of keypoint detection and description (Leutenegger, Chli, and Siegwart 2011; Rublee

*Corresponding author

et al. 2011; Salti, Lanza, and Di Stefano 2013; Tombari and Stefano 2014). Most keypoint-based approaches use broad-domain interest point detectors, such as Harris corner detector (Chen et al. 2010), SIFT (Yang et al. 2007; Tsai et al. 2009) or SURF (Wang et al. 2015), to produce a large number of interest points for matching. However, the detected points are not necessarily representative characteristics of the retinal image contents. Many of them may not present across the different images. In contrast, domain-specific methods (Hervella et al. 2018; Laliberté, Gagnon, and Sheng 2003) rely on the extraction of natural interest points, such as vessel intersections. This benefits the registration as the domain-specific interest points are easier to be preserved across different images as compared to those generic points produced by broad-domain methods. Meanwhile, with reduction of non-representative detected points, the computational cost can be dramatically saved. This also helps in the posterior point matching as it reduces the probability of matching wrong correspondences.

With the development of deep learning, many deep learning-based keypoint detection and description methods (Liao et al. 2017; Fu et al. 2020; Silva et al. 2021; Zhang et al. 2022; Bharati et al. 2022) have been introduced over the past few years, which have shown to be more promising than traditional methods such as SIFT. SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018) is a self-supervised framework for training a keypoint detector and its descriptor decoder. This network computes pixel-level feature locations and their associated descriptors on images in a single forward pass. During the training process, a warped image is created by applying a randomly generated homography $\mathcal{H}$. The loss between the labels of the original image and the warped image is computed to train the network. R2D2 (Revaud et al. 2019) is another framework for jointly learning the keypoint detection and description. It avoids ambiguous areas and increases the reliability of keypoint detection and description. GLAMpoint (Truong et al. 2019) proposes a trainable detector for retinal image registration. As a self-supervised keypoint detection method, it exploits known spatial correspondence between original image and its geometric transformed image produced by a controlled homography. However, it detects many points on non-vascular areas, which is unreliable for high-resolution image registration. Moreover, the random sampling of negative points is ineffective. Depart from SuperPoint, Liu et al. (Liu et al. 2022) propose SuperRetina, which is an end-to-end method for retinal image matching with jointly trainable keypoint detector and descriptor. However, additional manually annotated data are used. It tries to maximize the distance between the keypoint and the most similar non-matching point, which faces challenges for keypoints with similar local appearances.

To overcome the limitations of existing methods, we propose a constrained negative sampling (CNS) method to train the descriptor decoder more effectively. We also propose to use vessel detection task as an auxiliary regularization to improve keypoint detection. Moreover, we take the 3D-2D mapping deformation into consideration and propose to use polynomial fitting to compute the mapping between different views of images.

The main contributions of our method include:

1. We introduce vessel detection via multi-task to regulate training of the model. It helps to learn more representative features and reduce the risk of over-fitting for keypoint detection.
2. We propose a constrained negative sampling approach for self-supervised descriptor decoder training. It randomly selects a non-matching point from a group of most similar keypoints to compute the negative loss and helps to obtain better discrimination.
3. We propose hybrid matching to compute a non-linear mapping for image pairs, which improves the results especially for images with large translation.
4. Our experimental results show that the proposed method outperforms other methods.

Methodology

Network Architecture

Inspired by SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018), we propose SuperJunction, a learning-based keypoint detection and description framework for retinal image registration. Our network is illustrated in Figure 1. It consists of a feature encoder module regularized by a multi-task vessel decoder module, an interest point decoder, and a descriptor decoder.

Baseline Approach In the SuperJunction architecture, we follow the network in SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018) except that the pretrained ResNet-34 (He et al. 2016) is used in the feature encoder module as backbone to extract features. We follow SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018) and adopt multi-task interest point decoder and descriptor decoder. The interest point decoder head computes over feature maps with size $H/8\times W/8\times 128$ and outputs a tensor sized $H/8\times W/8\times 65$ through convolutions. Softmax and reshape are carried out to bring the representation back to a tensor sized $H \times W \times 1$. This output tells the probability of each pixel being keypoint given the input. For descriptor decoder, it computes and outputs a tensor sized $H/8 \times W/8 \times 256$ through convolutions from the same feature maps. And then bi-cubic interpolation and l_2 normalization are performed to output a tensor sized $H \times W \times 256$ to describe the pixels.

Inspired by recent development of deep learning, a pure data-driven model may converge to a local optimum. To alleviate this concern, we propose to regularize the feature encoder using an auxiliary vessel detection task.

Multi-task Vessel Regularization Multi-task learning is a subfield of machine learning in which multiple tasks are simultaneously learned by a shared model. Such approaches offer advantages to reduce overfitting by leveraging auxiliary information. Junctions in retinal images are often reliable keypoints for image matching. In retinal images, vessel crossovers and bifurcations are usually identified as junctions. Motivated from this, we propose to use vessel detection task as an auxiliary task for keypoint detection. As

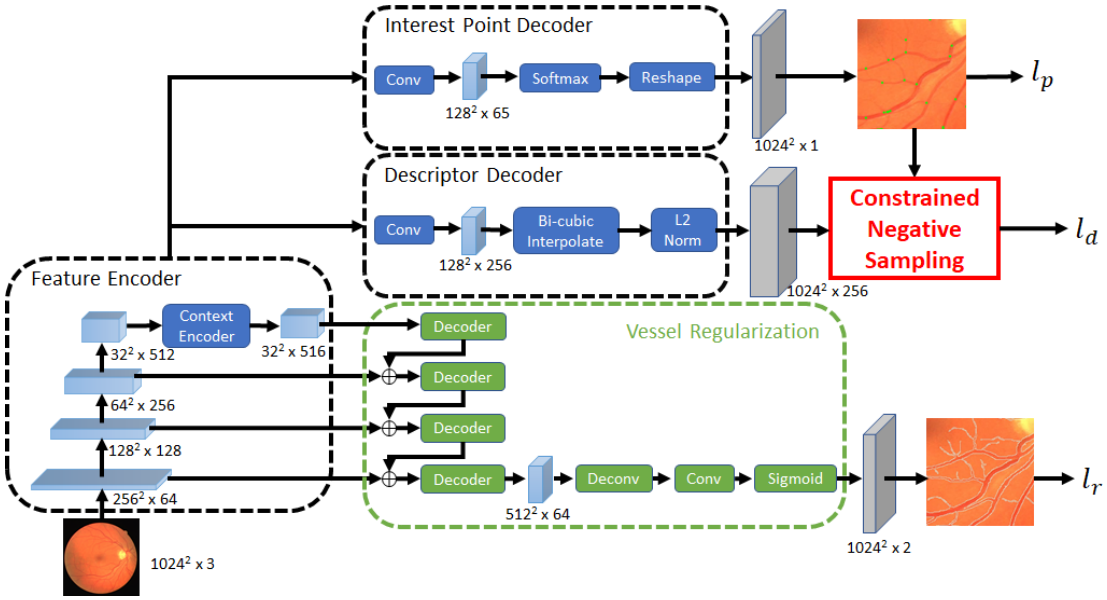


Figure 1: Network architecture: it includes a feature encoder block, a vessel regularization block, and interest point decoder and description blocks. We propose a constrained negative sampling (CNS) approach to compute the loss for descriptor training. We also use vessel detection to regularize the model training.

shown in Figure 1, it consists of four decoder modules (Gu et al. 2019), where each decoder block includes a 1×1 convolution, a 3×3 transposed convolution and a 1×1 convolution consecutively.

Constrained Negative Sampling

The descriptor of keypoint is trained in a self-supervised manner to make it invariant to homography transform. Previously, a triplet loss (Schroff, Kalenichenko, and Philbin 2015; DeTone, Malisiewicz, and Rabinovich 2018) was computed by comparing a baseline point to its paired point and an unpaired point. It is straightforward to use the same keypoint after homography transform as paired keypoint. However, the selection of unpaired keypoint is important as well. When all points are used in SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018), the differentiation among non-keypoints is often challenging. A spontaneous idea is to choose from detected keypoints instead of any random point. However, this is ineffective as the chance of selecting very similar but unpaired keypiont is low.

In this paper, we propose a constrained negative sampling approach to choose a better unpaired keypoint for self-supervised descriptor decoder training. Given a keypoint, we randomly select a point from top K most similar unpaired keypoints ranked by Euclidean distance and a random point from all unpaired keypoints. By choosing the two unpaired keypoints, it helps to obtain a better discrimination among keypoints. In this paper, we use $K = 10$ empirically.

Loss Function

Due to the sparse and imbalance nature of keypoints annotation, the binary labels of keypoints are converted to soft labels through 2D Gaussian blur, then a cross entropy loss L_p is computed:

$$L_p = \frac{1}{HW} \sum_{h=1,w=1}^{H,W} CE(x_{h,w}, y_{h,w}), \tag{1}$$

where $x_{h,w}$ and $y_{h,w}$ denote the Gaussian blurred ground truth map and the predicted map, $CE(\cdot)$ denotes the function to compute cross entropy loss.

A regularization loss L_r is calculated based on the performance of vessel detection. As the regions of vessel and non-vessel areas are unbalanced, we use focal loss (Lin et al. 2017) to compute L_r, as depicted below:

$$L_r = \mathcal{F}_{focal}(v_{gt}, v_{est}). \tag{2}$$

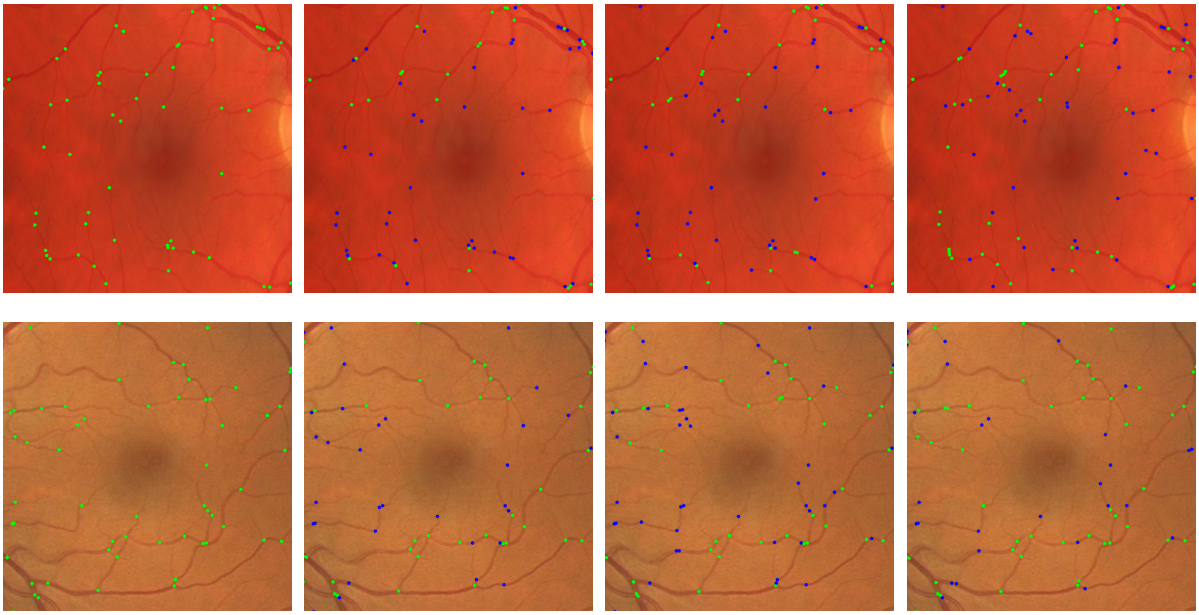
where v_{gt} and v_{est} denote the ground truth vessel and the estimated vessel. $\mathcal{F}_{focal}$ is the function used to compute the focal loss (Lin et al. 2017).

As the detector and descriptor should be invariant to homography transform, a self-supervised inconsistent loss is computed. Denote P as a keypoint in the original image I, its corresponding keypoint in the transformed image is represented by $\mathcal{H}(P)$. Since P and $\mathcal{H}(P)$ are paired, we expect their descriptors to be similar. Therefore, we shall minimize the distance $D(P, \mathcal{H}(P))$ between the two descriptors. Let R and R_c represent a random keypoint and a constrained negative sampled keypoint respectively, the descriptor loss is computed as

$$L_d = D(P, \mathcal{H}(P)) - \frac{1}{2} D(P, R) - \frac{1}{2} D(P, R_c) \tag{3}$$

We combine the three types of loss above to get the overall loss as:

$$L = L_p + L_{\mathcal{H}(p)} + \lambda_d L_d + \lambda_r L_r \tag{4}$$



(a) Ground truth (b) Baseline (c) +Regularization (d) +CNS

Figure 2: Visualization of keypoint detection performance. The green points show the ground truth points and the points that are correctly detected. The blue points show both the wrongly detected and the missed points.

where λ_d and λ_r control the balance between different losses and are set to 0.1 and 100 empirically in the experiment. The performance of the algorithm is not sensitive to small changes of the two values. Our interest point decoder, descriptor decoder and vessel regularization are jointly trained by minimizing this combined loss.

Retinal Image Matching

Retinal image matching is a critical step to register the moving image with the fixed reference image. Given a fixed reference image I_f and a moving image I_m, SuperJunction computes their keypoints as well as the corresponding descriptors. To match the keypoints based on descriptors, we first train a SuperGlue (Sarlin et al. 2020) model based on optimal transport. Then a mapping between the two images is computed using homography given matched points, following many earlier methods (DeTone, Malisiewicz, and Rabinovich 2018; Sarlin et al. 2020; Liu et al. 2022). Homography estimation is a good solution for linear registration tasks as it reduces errors in keypoint detection and matching. However, as the retina is not a plane surface, the mapping between two different views of retinal images is non-linear when the translation between two images is large. Ignoring such non-linearity in the matching would often lead to misalignment in the mosaic visualization. In consideration of this issue, we propose a hybrid matching algorithm. We first compute a global homography given the matching points from the two images. Then we evaluate the translation based on the transformation matrix. When the translation is below a pre-defined threshold, we adopt the global homography to accomplish the matching. In this paper, we use a threshold of 300 pixels. Otherwise, a global high-order polynomial-fitting method is introduced as non-linear mapping (Harris, Stephens et al. 1988; Ryan, Heneghan, and de Chazal 2004; Sharma et al. 2018).

In the polynomial fitting, we aim to establish a non-linear transformation between the matched points obtained by our keypoint detection and SuperGlue matching from the moving-fixed image pair. Given a point (x_f, y_f) in the fixed reference image, it shall be matched to point (x_m, y_m) in the moving image. We apply the polynomial fitting as below:

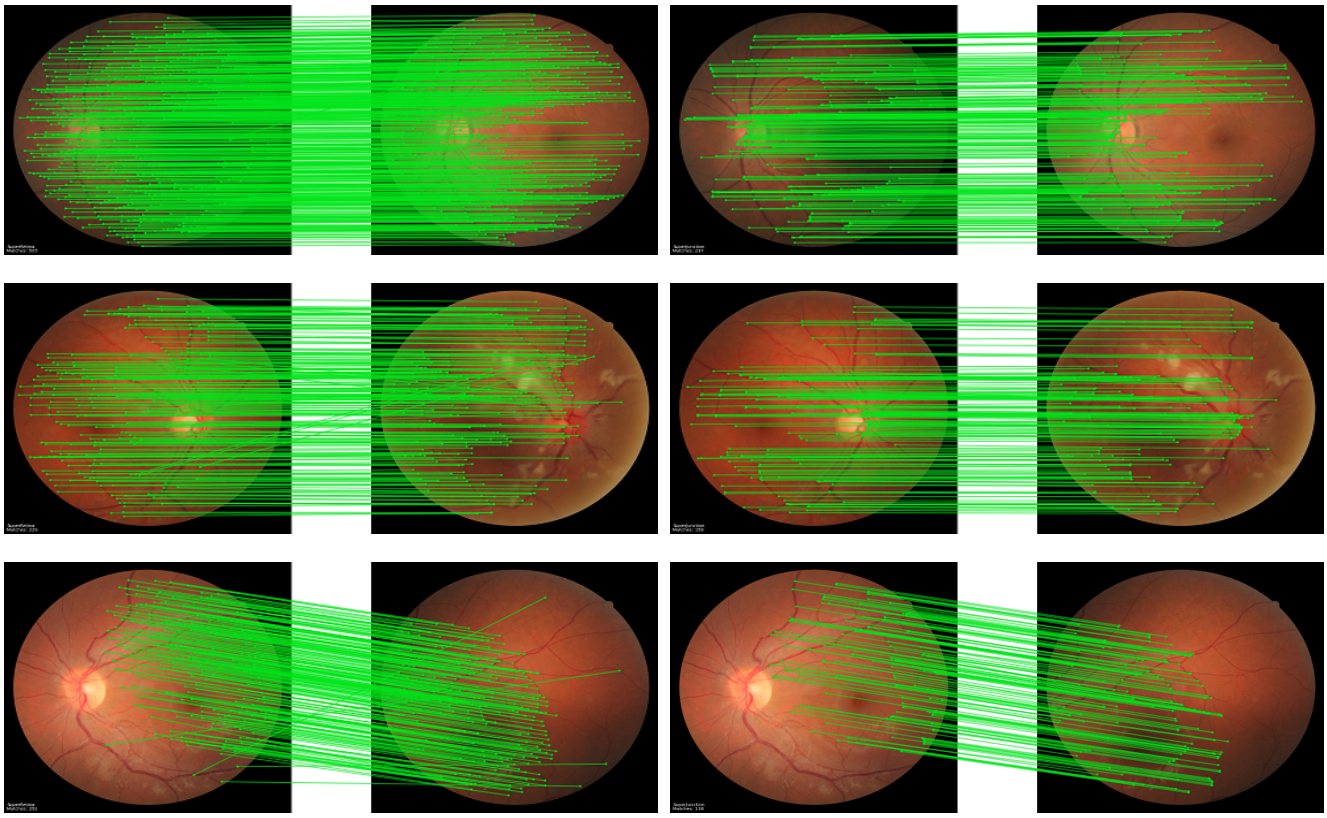
$$x_m = \sum_{k=0}^{N} \sum_{\substack{i+j=k,\\ i,j\geq 0}} a(i,j,k) x_f^i y_f^j, \tag{5}$$

$$y_m = \sum_{k=0}^{N} \sum_{\substack{i+j=k,\\ i,j\geq 0}} b(i,j,k) x_f^i y_f^j, \tag{6}$$

Here, N represents the order of the polynomial function, while $a(i,j,k)$ and $b(i,j,k)$ refer to the coefficients of these polynomial functions. In our experiments, we conduct a study to compare the performance of polynomials with N varying from 2 to 5, and we find that the second-order polynomial functions lead to best performance. Detailed results can be found in Section .

Experimental Results

In this section, we compare our approach with several state-of-the-art methods, including three traditional methods and six deep learning-based methods.



(a) SuperRetina (b) SuperJunction

Figure 3: Comparison on keypoints matching under SuperRetina (left) and the proposed SuperJunction (right). From first row to last row are results from subset $\mathcal{S}$, $\mathcal{A}$, $\mathcal{P}$.

Datasets and Evaluation Metrics

We use four commonly used datasets containing vessel masks for training, including DRIVE (Staal et al. 2004), STARE (Hoover, Kouznetsova, and Goldbaum 2000), HRF (Budai et al. 2013) and CHASE_DB (Fraz et al. 2012). There are altogether 133 images, where 121 randomly selected images are used for training while the rest images are used for evaluation of keypoint detection. FIRE dataset (Hernandez-Matas et al. 2017) [1] is used for registration performance evaluation. FIRE contains 134 pairs of images divided into three categories: $\mathcal{A}$ (14 image pairs) with significant anatomical change and large overlapped region, $\mathcal{P}$ (49 image pairs) with low anatomical changes but small overlap, and $\mathcal{S}$ (71 image pairs) with low anatomical changes and large overlap. FIRE dataset also provides 10 pairs of manually labeled points as ground truths for registration evaluation.

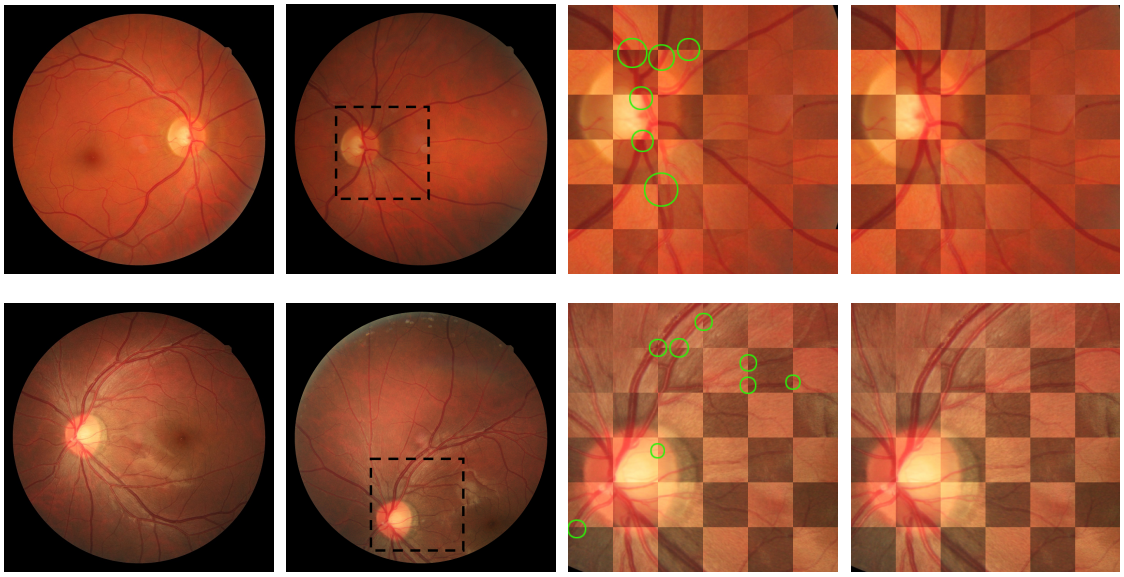
We compute precision and recall to evaluate the performance of keypoint detection. A keypoint is considered as detected successfully if any pixel within 5 pixels is detected as keypoint (Lee et al. 2019). The performance of image registration is evaluated by the root mean square error (RMSE) between 10 manually annotated points and their aligned locations. We follow (Hernandez-Matas et al. 2017) to compute the area under curve (AUC) of the success rate with threshold varying from 1 to 25. Higher score leads to better result. It should be noted that all the performance evaluation is carried out on the original image size of 2912×2912.

[1] https://www.kaggle.com/datasets/phylake1337/fire-dataset

Implementation Details

To train our model, we obtain a set of retinal images with annotated ground truth keypoints. However, manual annotation of the landmark keypoints can be tedious and subjective. Fortunately, the vessel masks of the four datasets have been annotated in previous studies. In this paper, we make use of these existing annotations to compute ground truth junction points using a set of well-established image processing techniques (Haralick and Shapiro 1992). Firstly, we adopt thinning algorithms (Lee, Kashyap, and Chu 1994) to compute the skeleton of the vessels from the vessel masks. Then we find the branching points which are defined as skeleton points with 3 or more neighboring skeleton points in the vessel skeleton. We use these points as ground truth junction points for model training. Common augmentations are applied online, including brightening, contrasting, cropping, rotation and translation.

We adopt ResNet34 pretrained on ImageNet as backbone of our network and implement the proposed SuperJunction on PyTorch platform. The training and testing bed is Ubuntu



(a) Moving Image (b) Fixed Image (c) SuperRetina (d) Proposed

Figure 4: Mosaic visualization of registration results. (a) Moving image (b) Fixed image (c) The checkerboard mosaic of results from the region enclosed by the dash line in black in (b) by SuperRetina (Liu et al. 2022) (d) The checkerboard mosaic of results by the proposed method.

18.04 system with two NVidia Geforce RTX 2080Ti graphics cards. We use image size of 1024×1024 when training the models. The network is trained with batch size 4. The optimizer is Adam with an initial learning rate of 0.0001, which is reduced to 0.00001 after 50 epochs. The maximum number of training epoch is set to 150.

Comparison with Other Methods

To demonstrate the registration performance of our work, we compare it with both traditional methods (SIFT (Lowe 2004), PBO (Oinonen et al. 2010), REMPE (Hernandez-Matas, Zabulis, and Argyros 2020)) and deep learning-based methods (SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018), GLAMpoints (Truong et al. 2019), R2D2 (Revaud et al. 2019), SuperGlue (Sarlin et al. 2020), NCNet (Rocco et al. 2020), SuperRetina (Liu et al. 2022)). Table 1 summarizes the comparison among these methods. It can be found that SuperJunction achieves the best results in general among all the methods. Take note that the AUC score (0.958) under REMPE (Hernandez-Matas, Zabulis, and Argyros 2020) is slightly higher than our method (0.954) for category $\mathcal{S}$ where the registration for image pairs with large overlap but small anatomical changes is easier. However, REMPE does not work well for $\mathcal{A}$ with large anatomical changes and $\mathcal{P}$ with small overlap. More importantly, SuperJunction requires 1 second to perform one registration, which is much faster than REMPE (nearly 3 minutes). This makes our method more suitable for applications with higher requirement on computational efficiency. Compared with SuperRetina, the state-of-the-art learning-based method, we improve the performance of all three subsets of the data and achieves mAUC of 0.850, which is 12.6% higher than 0.755 by SuperRetina. It is worth to mention that the improvement in AUC is more than 47.9% for subset $\mathcal{P}$, where non-linearity is more obvious due to low overlap between the fixed and moving images. We also conduct a statistical test and the results show that the improvement is significant with $p < 0.05$.

To visualize the benefit of our proposed SuperJunction method, we illustrate the matching performance between extracted keypoints from image pairs in Figure 3. Comparing with the state-of-the-art SuperRetina method, SuperJunction relies on less number of keypoints. However, this helps to avoid unreliable keypoints that are prone to matching error. In Figure 4, we also show the mosaic results of two samples from subset $\mathcal{P}$. It visually compares the results by our method with those by SuperRetina in a checkerboard mosaic with alternating patches from the fixed and the moving images. As we can see, SuperRetina suffers from some misalignment (see circles in green) between the alternative patches while our method performs better.

Ablation Study

To justify the effectiveness of the vessel regularization, the CNS module and the polynomial fitting, we conduct a set of ablation studies.

Effectiveness of Regularization and CNS Firstly we examine the keypoint detection performance in terms of precision and recall using SuperJunction on the reserved test images. Baseline is a basic approach when adopting ResNet-34 as backbone of the network. We first include vessel detection as a regularization to improve keypoint detection, denoted as '+Regularization'. Then we further incorporate

	Method	RMSE ↓	AUC-$\mathcal{S}$↑	AUC-$\mathcal{A}$↑	AUC-$\mathcal{P}$↑	mAUC↑
Traditional	SIFT (Lowe 2004)	-	0.903	0.474	0.341	0.573
	PBO (Oinonen et al. 2010)	-	0.844	0.691	0.122	0.552
	REMPE (Hernandez-Matas, Zabulis, and Argyros 2020)	-	**0.958**	0.660	<u>0.542</u>	0.720
Deep Learning	SuperPoint(DeTone, Malisiewicz, and Rabinovich 2018)	-	0.909	0.609	0.465	0.661
	GLAMpoints (Truong et al. 2019)	-	0.850	0.543	0.474	0.622
	R2D2 (Revaud et al. 2019)	-	0.928	0.666	0.540	0.711
	SuperGlue (Sarlin et al. 2020)	-	0.885	0.689	0.488	0.687
	NCNet (Rocco et al. 2020)	-	0.817	0.609	0.410	0.612
	SuperRetina (Liu et al. 2022)	5.475	0.940	<u>0.783</u>	<u>0.542</u>	<u>0.755</u>
	SuperJunction (Proposed)	**2.883**	<u>0.954</u>	**0.794**	**0.802**	**0.850**

Table 1: Comparison of SuperJunction with other methods. Bold font shows the best result and underline shows the second.

Method	Precsion↑	Recall↑	AUC-$\mathcal{S}$↑	AUC-$\mathcal{A}$↑	AUC-$\mathcal{P}$↑	mAUC↑
Baseline	63.9%	47.7%	0.946	0.780	0.700	0.809
+Regularization	63.9%	55.2%	0.948	0.805	0.765	0.839
+CNS	63.9%	63.5%	0.954	0.794	0.802	0.850

Table 2: Ablation study for the regularization and CNS modules of SuperJunction. Both the regularization and CNS improve the detection of keypoints and the subsequent registration. Hybrid matching is used in the final registration step.

constrained negative sampling to improve training of the descriptor, which is denoted as '+CNS'. As there is a trade-off between precision and recall, we choose thresholds to achieve same level of precision for easier comparison on the recall. The performance comparison is presented in Table 2. As we can see, the introduction of the regularization and CNS modules improves the recall by 7.5% and 8.3% respectively.

Meanwhile, we assess the registration performance of SuperJunction on FIRE dataset. The AUCs for all three categories (AUC-$\mathcal{S}$, AUC-$\mathcal{A}$, AUC-$\mathcal{P}$ for subset $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{P}$ resepectively) of the FIRE dataset together with the mean AUCs (mAUC) are computed. As shown in Table 2, both the regularization and the CNS help to improve the overall performance measured by mAUC. We also visualize the keypoint detection under different approaches in Figure 2, where green dots represent the keypoints correctly detected while blue dots denote wrongly detected and missed keypoints. It can be observed that with introducing of regularization and CNS, more keypoints can be correctly detected.

Effectiveness of Hybrid Matching To justify the effectiveness of the hybrid matching for low-overlapping image pairs, we also conduct an ablation study by using polynomial functions with different orders, denoted by N, for subset $\mathcal{P}$ with large translation. Table 3 compares the AUCs of hybrid matching when changing the order N of the polynomial fitting functions from two to five. As we can see from the results, second order polynomial fitting leads to largest improvement while higher order polynomials do not bring further improvement. We adopt $N = 2$ for our algorithms in this paper.

It is worth to note that the polynomial fitting in the hybrid matching relies on accurately matched points from the previous step. We also integrate the polynomial fitting with the state-of-the-art SuperRetina approach. However, we obtain AUC of 0.539 for subset $\mathcal{P}$. The result is not superior than 0.542 by SuperRetina due to the existence of mismatched keypoints (see Figure 3) as well as other less accurate points. This observation shows that accurate matching between points in the proposed method is critical when using polynomial fitting. It indirectly validates the efficacy of the introduced regularization and CNS modules.

N	2	3	4	5
AUC-$\mathcal{P}$	0.802	0.798	0.801	0.782

Table 3: Performances under different order of polynomial-fitting functions for subset $\mathcal{P}$.

Conclusions

In this paper, we propose a novel learning-based junction detection method for retinal image registration. To improve model training, we incorporate vessel detection as a regularization term, which helps to learn more representative features and reduce the risk of over-fitting for keypoint detection. A constrained negative sampling approach is proposed for self-supervised descriptor decoder training to obtain better discrimination among keypoints. In addition, we adopt hybrid matching to overcome the challenge of non-linear mapping in retinal image registration, which improves the performance especially for image pairs with large translation in subset $\mathcal{P}$. Experimental results show that the registration performance has been improved when comparing with the state-of-the-art methods. It should be noted that our method may not work well if the disease affects the vessel pattern dramatically. In the future, we will further extend the application to multi-modality retinal and other image registration to evaluate the effectiveness and generality of our method.

Acknowledgments

This work is supported by the Agency for Science, Technology and Research under its AI^3 Horizontal Technology Coordinating Office Grant C231118001 and C211118006.

References

Alahi, A.; Ortiz, R.; and Vandergheynst, P. 2012. Freak: Fast retina keypoint. In *2012 IEEE conference on computer vision and pattern recognition*, 510–517. Ieee.

Bharati, S.; Mondal, M.; Podder, P.; and Prasath, V. 2022. Deep learning for medical image registration: A comprehensive review. *arXiv preprint arXiv:2204.11341*.

Budai, A.; Bock, R.; Maier, A.; Hornegger, J.; and Michelson, G. 2013. Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013.

Chanwimaluang, T.; Fan, G.; and Fransen, S. R. 2006. Hybrid retinal image registration. *IEEE transactions on information technology in biomedicine*, 10(1): 129–142.

Chen, J.; Tian, J.; Lee, N.; Zheng, J.; Smith, R. T.; and Laine, A. F. 2010. A partial intensity invariant feature descriptor for multimodal retinal image registration. *IEEE Transactions on Biomedical Engineering*, 57(7): 1707–1718.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236.

Ding, L.; Kang, T.; Kuriyan, A.; Ramchandran, R.; Wykoff, C.; and Sharma, G. 2023. Combining Feature Correspondence with Parametric Chamfer Alignment: Hybrid Two-Stage Registration for Ultra-Widefield Retinal Images. *IEEE Transactions on Biomedical Engineering*, 70(2): 523–532.

Fraz, M. M.; Remagnino, P.; Hoppe, A.; Uyyanonvara, B.; Rudnicka, A. R.; Owen, C. G.; and Barman, S. A. 2012. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9): 2538–2548.

Fu, Y.; Lei, Y.; Wang, T.; Curran, W. J.; Liu, T.; and Yang, X. 2020. Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20): 20TR01.

Ghassabi, Z.; Shanbehzadeh, J.; Sedaghat, A.; and Fatemizadeh, E. 2013. An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors. *EURASIP Journal on Image and Video Processing*, 2013(1): 1–16.

Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; and Liu, J. 2019. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10): 2281–2292.

Haralick, R. M.; and Shapiro, L. G. 1992. *Computer and Robot Vision*. USA: Addison-Wesley Longman Publishing Co., Inc., 1st edition. ISBN 0201569434.

Harris, C.; Stephens, M.; et al. 1988. A combined corner and edge detector. In *Alvey vision conference*, volume 15, 10–5244. Citeseer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hernandez-Matas, C.; Zabulis, X.; and Argyros, A. A. 2020. REMPE: Registration of retinal images through eye modelling and pose estimation. *IEEE journal of biomedical and health informatics*, 24(12): 3362–3373.

Hernandez-Matas, C.; Zabulis, X.; Triantafyllou, A.; Anyfanti, P.; Douma, S.; and Argyros, A. A. 2017. FIRE: fundus image registration dataset. *Modeling and Artificial Intelligence in Ophthalmology*, 1(4): 16–28.

Hervella, Á. S.; Rouco, J.; Novo, J.; and Ortega, M. 2018. Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. *Procedia Computer Science*, 126: 97–104.

Hoover, A.; Kouznetsova, V.; and Goldbaum, M. 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3): 203–210.

Laliberté, F.; Gagnon, L.; and Sheng, Y. 2003. Registration and fusion of retinal images-an evaluation study. *IEEE Transactions on Medical Imaging*, 22(5): 661–673.

Lee, J. A.; Cheng, J.; Lee, B. H.; Ong, E. P.; Xu, G.; Wong, D. W. K.; Liu, J.; Laude, A.; and Lim, T. H. 2015. A low-dimensional step pattern analysis algorithm with application to multimodal retinal image registration. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1046–1053.

Lee, J. A.; Liu, P.; Cheng, J.; and Fu, H. 2019. A deep step pattern representation for multimodal retinal image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5077–5086.

Lee, T.-C.; Kashyap, R. L.; and Chu, C.-N. 1994. Building skeleton models via 3-D medial surface axis thinning algorithms. *CVGIP: Graphical Models and Image Processing*, 56(6): 462–478.

Leutenegger, S.; Chli, M.; and Siegwart, R. Y. 2011. BRISK: Binary robust invariant scalable keypoints. In *2011 International conference on computer vision*, 2548–2555. Ieee.

Liao, R.; Miao, S.; de Tournemire, P.; Grbic, S.; Kamen, A.; Mansi, T.; and Comaniciu, D. 2017. An Artificial Agent for Robust Image Registration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 4168–4175. AAAI Press.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*.

Liu, J.; Li, X.; Wei, Q.; Xu, J.; and Ding, D. 2022. Semi-supervised Keypoint Detector and Descriptor for Retinal Image Matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, 593–609. Springer.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2): 91–110.

Motta, D.; Casaca, W.; and Paiva, A. 2019. Vessel Optimal Transport for Automated Alignment of Retinal Fundus Images. *IEEE Transactions on Image Processing*, 28(12): 6154–6168.

Ng, P. C.; and Henikoff, S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13): 3812–3814.

Oinonen, H.; Forsvik, H.; Ruusuvuori, P.; Yli-Harja, O.; Voipio, V.; and Huttunen, H. 2010. Identity verification

based on vessel matching from fundus images. In *2010 IEEE International Conference on Image Processing*, 4089–4092. IEEE.

Oliveira, F. P.; and Tavares, J. M. R. 2014. Medical image registration: a review. *Computer methods in biomechanics and biomedical engineering*, 17(2): 73–93.

Revaud, J.; Weinzaepfel, P.; de Souza, C. R.; and Humenberger, M. 2019. R2D2: Repeatable and Reliable Detector and Descriptor. *NeurIPS*.

Ritter, N.; Owens, R.; Cooper, J.; Eikelboom, R. H.; and Van Saarloos, P. P. 1999. Registration of stereo and temporal images of the retina. *IEEE Transactions on medical imaging*, 18(5): 404–418.

Rocco, I.; Cimpoi, M.; Arandjelović, R.; Torii, A.; Pajdla, T.; and Sivic, J. 2020. Ncnet: Neighbourhood consensus networks for estimating image correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2): 1020–1034.

Rosten, E.; and Drummond, T. 2005. Fusing points and lines for high performance tracking. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, 1508–1515. Ieee.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, 2564–2571. Ieee.

Ryan, N.; Heneghan, C.; and de Chazal, P. 2004. Registration of digital retinal images using landmark correspondence by expectation maximization. *Image and Vision Computing*, 22(11): 883–898.

Salti, S.; Lanza, A.; and Di Stefano, L. 2013. Keypoints from symmetries by wave propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2898–2905.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938–4947.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Sharma, L.; Sharma, J. K.; Anand, D.; and Sharma, S. 2018. An Adaptive Window Based Polynomial Fitting Approach for Pixel Matching in Stereo Images. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 657–661. IEEE.

Silva, T. D.; Chew, E. Y.; Hotaling, N.; and Cukras, C. A. 2021. Deep-learning based multi-modal retinal image registration for the longitudinal analysis of patients with age-related macular degeneration. *Biomed. Opt. Express*, 12(1): 619–636.

Staal, J.; Abràmoff, M. D.; Niemeijer, M.; Viergever, M. A.; and Van Ginneken, B. 2004. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4): 501–509.

Stewart, C. V.; Tsai, C.-L.; and Roysam, B. 2003. The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *IEEE transactions on medical imaging*, 22(11): 1379–1394.

Tombari, F.; and Stefano, L. D. 2014. Interest points via maximal self-dissimilarities. In *Asian Conference on Computer Vision*, 586–600. Springer.

Truong, P.; Apostolopoulos, S.; Mosinska, A.; Stucky, S.; Ciller, C.; and Zanet, S. D. 2019. Glampoints: Greedily learned accurate match points. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10732–10741.

Tsai, C.-L.; Li, C.-Y.; Yang, G.; and Lin, K.-S. 2009. The edge-driven dual-bootstrap iterative closest point algorithm for registration of multimodal fluorescein angiogram sequence. *IEEE transactions on medical imaging*, 29(3): 636–649.

Wang, G.; Wang, Z.; Chen, Y.; and Zhao, W. 2015. Robust point matching method for multimodal retinal image registration. *Biomedical Signal Processing and Control*, 19: 68–76.

Wang, W.; Chen, H.; Li, J.; and Yu, J. 2006. A registration method of fundus images based on edge detection and phase-correlation. In *First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06)*, volume 3, 572–576. IEEE.

Yang, G.; Stewart, C. V.; Sofka, M.; and Tsai, C.-L. 2007. Registration of challenging image pairs: Initialization, estimation, and decision. *IEEE transactions on pattern analysis and machine intelligence*, 29(11): 1973–1989.

Zana; and Klein. 1999. A registration algorithm of eye fundus images using a Bayesian Hough transform. *International Conference on Image Processing And Its Applications*, 2: 479–483 vol.2.

Zhang, J.; Wang, Y.; Dai, J.; Cavichini, M.; Bartsch, D.-U. G.; Freeman, W. R.; Nguyen, T. Q.; and An, C. 2022. Two-Step Registration on Multi-Modal Retinal Images via Deep Neural Networks. *IEEE Transactions on Image Processing*, 31: 823–838.