RetinaRegNet: A Versatile Approach for Retinal Image Registration

Vishal Balaji Sivaraman, Muhammad Imran, Qingyue Wei, Preethika Muralidharan, Michelle R. Tamplin,
Isabella M . Grumbach, Randy H. Kardon, Jui-Kai Wang, Yuyin Zhou, and Wei Shao

***Abstract*— We introduce the RetinaRegNet model, which can achieve state-of-the-art performance across various retinal image registration tasks. RetinaRegNet does not require training on any retinal images. It begins by establishing point correspondences between two retinal images using image features derived from diffusion models. This process involves the selection of feature points from the moving image using the SIFT algorithm alongside random point sampling. For each selected feature point, a 2D correlation map is computed by assessing the similarity between the feature vector at that point and the feature vectors of all pixels in the fixed image. The pixel with the highest similarity score in the correlation map corresponds to the feature point in the moving image. To remove outliers in the estimated point correspondences, we first applied an inverse consistency constraint, followed by a transformation-based outlier detector. This method proved to outperform the widely used random sample consensus (RANSAC) outlier detector by a significant margin. To handle large deformations, we utilized a two-stage image registration framework. A homography transformation was used in the first stage and a more accurate third-order polynomial transformation was used in the second stage. The model's effectiveness was demonstrated across three retinal image datasets: color fundus images, fluorescein angiography images, and laser speckle flowgraphy images. RetinaRegNet outperformed current state-of-the-art methods in all three datasets. It was especially effective for registering image pairs with large displacement and scaling deformations. This innovation holds promise for various applications in retinal image analysis. Our code is publicly available at `https://github.com/mirthAI/RetinaRegNet`.**

***Index Terms*— Retinal image registration, diffusion features, point correspondences, inverse consistency, outlier detector**

I. Introduction

IMAGE registration is a fundamental technique in medical imaging. It has a wide range of applications, such as image-guided surgery, radiation therapy, disease progression monitoring, image fusion, brain mapping, and computer-aided diagnosis [1], [2]. The goal of image registration is to align a moving (source) image with a fixed (target) image by estimating a geometric transformation that matches corresponding features or structures in the two images. Retinal image registration emerges as a crucial tool in ophthalmology, enabling accurate tracking of changes in retinal structures over time and alignment of different modalities, essential for assessing disease progression and treatment efficacy [3]. Color fundus, fluorescein angiography, and laser speckle flowgraphy are different sensors that can capture retinal images for applications in diagnosing retinal conditions, assessing treatment outcomes, and monitoring retinal blood flow dynamics.

The main challenges in retinal image registration include estimating large deformations, registering images with minimal overlap, and learning from relatively small datasets [4]. Large deformations involve the alignment of images with notable differences in position or scale, which often result in overfitting in transformation estimation or complicating point correspondence estimation. Small overlaps present challenges due to limited correspondences and ambiguity in feature detection. In the case of small datasets, a deep learning method that does not require training on retinal images is more preferable compared to those that require a large task-specific training dataset.

This work was supported by the Department of Medicine and the Intelligent Clinical Care Center at the University of Florida College of Medicine.

Vishal Balaji Sivaraman is with the Department of Electrical and Computer Engineering at the University of Florida, Gainesville, FL, USA (e-mail: vi.sivaraman@ufl.edu).

Muhammad Imran is with the Department of Medicine at the University of Florida, Gainesville, FL, USA (e-mail: muhammad.imran@ufl.edu).

Qingyue Wei is with the Department of Computational and Mathematical Engineering at Stanford University, Stanford, CA, USA (e-mail: qywei@stanford.edu).

Preethika Muralidharan is with the Department of Health Outcomes and Biomedical Informatics at the University of Florida, Gainesville, FL, USA (e-mail: p.muralidharan@ufl.edu).

Michelle R. Tamplin is with the Department of Internal Medicine and the Iowa City VA Center for the Prevention and Treatment of Visual Loss, University of Iowa, IA, USA (e-mail: michelle-tamplin@uiowa.edu).

Isabella M. Grumbach is with the Department of Internal Medicine, the Iowa City VA Center for the Prevention and Treatment of Visual Loss, and the Department of Radiation Oncology, University of Iowa, IA, USA (e-mail: isabella-grumbach@uiowa.edu).

Randy H. Kardon is with the Iowa City VA Center for the Prevention and Treatment of Visual Loss and the Department of Ophthalmology and Visual Sciences, University of Iowa, IA, USA (e-mail: randy-kardon@uiowa.edu).

Jui-Kai Wang is with the Iowa City VA Center for the Prevention and Treatment of Visual Loss, the Department of Ophthalmology and Visual Sciences, and the Department of Electrical and Computer Engineering, University of Iowa, IA, USA (e-mail: jui-kai-wang@uiowa.edu).

Yuyin Zhou is with the Department of Computer Science and Engineering, University of California, Santa Cruz, CA, USA (e-mail: yzhou284@ucsc.edu).

Wei Shao is with the Department of Medicine, Department of Electrical and Computer Engineering, and the Intelligent Clinical Care Center, University of Florida, Gainesville, FL, USA (e-mail: weishao@ufl.edu).

To address the aforementioned challenges in retinal image registration, we advocated for feature-based registration methods, which begin with image feature extraction, followed by transformation estimation. The most widely used feature points detector is the Scale-Invariant Feature Transform (SIFT) algorithm [5], which can detect and describe rotation- and scaling-invariant local image features like corners and blobs. However, SIFT has limitations when applied to retinal image registration. First, most detected feature points tend to cluster in areas with rich textural features, leading to sparse or nonexistent point correspondences in homogeneous regions without distinctive features. Consequently, these methods are more effective for estimating global deformations, such as affine or homography transformations, but less so for detecting nonrigid transformations. Second, the image features generated by SIFT lack a comprehensive perspective of the entire image, resulting in outliers in estimated point correspondences. Efforts to address these challenges have been twofold. Firstly, there has been a focus on developing more efficient feature detectors. These include general-purpose detectors [6], [7] as well as those tailored for specific registration applications [8]–[10]. The second approach is detector-free, involving the extraction of two-dimensional (2D) feature maps from both fixed and moving images to compute a four-dimensional (4D) correlation map [11]. This map facilitates the identification of more feature point correspondences in regions with less distinct features.

In this paper, we introduce RetinaRegNet as a versatile model that achieves state-of-the-art performance on various retinal image registration tasks through several innovations. First, we sampled points from both regions with rich features (e.g., edges, corners) and without distinct features to ensure that feature points are distributed throughout the images. This significantly improved registration accuracy in homogeneous regions. Second, our model reliably identified point correspondences using the concept of inverse consistency constraint [12], which was originally used to ensure that if two images are registered to each other and then reversely registered, the estimated transformations are inverses of each other. In the context of estimating point correspondences between two images, inverse consistency is a necessary condition for a good estimator. Third, our model excluded incorrect point correspondences with a transformation-based outlier detector. The most widely used RANSAC (Random Sample Consensus) outlier-detection algorithm [13] relies on random sampling, which can lead to inconsistent results and increased computational time, especially when dealing with complex data with many outliers. Finally, our model utilized a two-stage registration framework. The first stage involved estimating a global deformation to achieve a global, yet accurate, alignment of the two images. This enabled step two, which was the estimation of a more accurate non-rigid local transformation. This two-stage design enhanced both the robustness and accuracy of our model.

We benchmarked RetinaRegNet using three different retinal image registration datasets and compared it with leading retinal registration methods, highlighting its versatility and effectiveness. In the public FIRE dataset [14] of color fundus images, when compared to the best existing method, RetinaRegNet improved the AUC from 0.783 to 0.901 and reduced the mean landmark error from 5.99 to 2.97. In the public FLoRI21 dataset [15] of fluorescein angiography images, when compared to the best existing method, RetinaRegNet improved the AUC of the second-best method from 0.640 to 0.868 and reduced the mean landmark error from 41.47 to 13.83. In the private LSFG dataset of laser-speckle flowgraphy images, when compared to the best existing methods, RetinaRegNet improved the AUC of the second-best method from 0.853 to 0.861 and reduced the mean landmark error from 4.23 to 4.00. This state-of-the-art performance across various retinal image datasets affirmed RetinaRegNet's significant potential in revolutionizing retinal image registration.

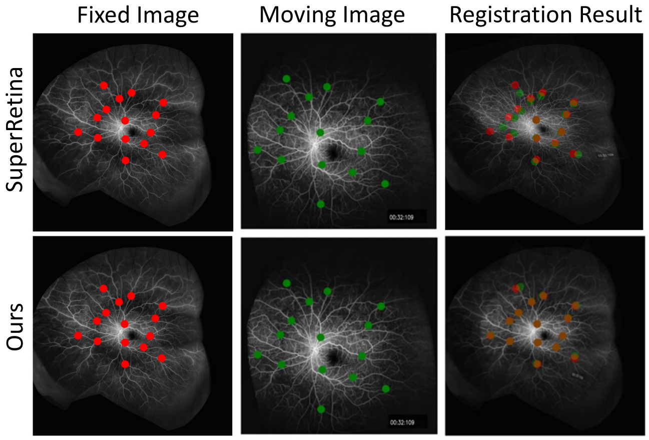

Fig. 1. Visual comparison of the performance of RetinaRegNet and SuperRetina [9] when registering an image pair with large displacement and scaling deformations. Red points represent landmarks on the fixed image, and green points represent the corresponding landmarks on the moving image. In the third column, the deformed images are overlaid with the fixed image, and we showed landmarks on both the fixed and deformed images in this overlay. The results showed that our registration model significantly improves the alignment of the landmarks throughout the image. The improved registration results can also be seen in the much sharper overlaid image produced by our model.

II. Related Work

A. Retinal Image Registration

Retinal image registration methods can be divided into two categories: intensity-based methods and feature-based methods, or a combination of the two [16]. Intensity-based image registration methods typically utilize an intensity similarity function (e.g., mutual information, cross-correlation, sum of squared differences) to align the intensity differences between two images [17]–[19]. While intensity-based registration methods may achieve promising registration results, they are susceptible to overfitting when dealing with images exhibiting large deformations, minimal overlaps, and variations in illumination and textures [16]. Additionally, the choice of the similarity function is task-specific due to variations in imaging modality. In contrast, feature-based registration methods rely on image features such as vasculature bifurcations, fovea, and optic disc for transformation estimation. Feature-based methods are generally more robust and accurate compared to their intensity-based counterparts. Due to these advantages, numerous feature-based retinal image registration techniques, using conventional [20]–[22] and learning-based approaches [23]–[29], have been developed. Learning-based techniques outperform conventional methods by achieving higher accuracy, robustness, and efficiency, but they often require extensive preprocessing [23]–[25] or training on large datasets [26]–[29].

B. Semantic Correspondences

The goal of semantic correspondence is to identify and match similar semantic features across different images. This process often involves finding correspondences between objects (e.g., points, lines, or regions) based on their semantic meaning rather than visual similarities. Semantic correspondence typically relies on advanced deep learning methods, including convolutional neural networks and vision transformers, to extract and link these semantically similar elements across varied visual representations [30]–[32]. There are two major differences between semantic correspondence and image registration. First, semantic correspondence identifies point correspondences by matching semantic features instead of visual features. Consequently, the images in semantic correspondence can depict two different objects, for example, birds of different species, which are not of interest in image registration. Second, while image registration aims to identify one-to-one point correspondences between two images, semantic correspondence seeks to establish sparse point correspondences (not necessarily one-to-one) between parts of images that share the same underlying features.

C. Diffusion Models

Denoising diffusion models [33]–[35] have emerged as a leading class of generative models renowned for their capacity to generate high-quality images. These models function by iteratively refining an initial Gaussian noise image through a series of reverse diffusion steps, which gradually diminishes noise and enhances image quality. Diffusion models have demonstrated promising performance across various applications, including image denoising [36], super-resolution [37], image inpainting [38], and image segmentation [39]. Training and sampling diffusion models in the original image space can be computationally expensive. A more efficient approach is to conduct the forward and reverse diffusion processes in a smaller latent image space. A prime example of this is stable diffusion [40], which has subsequently been repurposed for numerous generation tasks such as text-to-3D conversion [41] and image editing [42]. Stable diffusion features have recently found applications in representation learning [43] and semantic correspondences [44], [45].

III. Method

Figure 2 provides an overview of the proposed retinal image registration model, titled 'RetinaRegNet'. Inspired by the application of diffusion features (DIFT) in semantic correspondence, as detailed in [44], our model began by extracting image features using a pre-trained stable diffusion model [40]. Subsequently, we identified key feature points in the moving image using the SIFT algorithm alongside random point sampling. This approach allowed for the extraction of feature points in areas with both rich and poor textures. For each identified feature point in the moving image, we computed a 2D correlation map to locate its corresponding point in the fixed image. To mitigate potential outliers in the estimated point correspondences, we employed an inverse consistency constraint coupled with a transformation-based outlier detector. These strategies were crucial for significantly improving the accuracy and robustness of our registration method.

A. Diffusion Feature Extraction

Here we will briefly describe the extraction of diffusion image features proposed by Tang et al. [44]. The forward diffusion process is a Markov chain that starts with a noise-free image and gradually adds noise to it over T steps, transforming the image into pure Gaussian noise. The model then learns to reverse this process, starting with a pure noise image and gradually removing noise from the image to generate new data samples. In latent diffusion models, the diffusion process primarily occurs in the latent space. At each time step t, noise is added to the latent image z_{t-1} to generate z_t with the following equation:

$$q(z_t|z_{t-1}) := \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

where $0 < \beta_0 < \cdots < \beta_T << 1$ is a predefined variance scheduler.

The reverse process involves training a model $\epsilon_\theta(z_t, t)$, typically implemented as a U-Net, to predict the Gaussian noise ϵ added at each step of the forward process. The reverse process is represented as:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}t}}\epsilon_\theta(z_t, t)\right) + \sigma_t\mathbf{z} \tag{2}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To generate diffusion features of a high-quality image x_0, we initially encoded the image into the latent representation z_0 using the encoder of a pre-trained autoencoder. Then, we introduced random noise to z_0 by selecting a time step t and calculating the noisy image using: $z_t = \sqrt{\bar{\alpha}_t}z_0 + (1 - \bar{\alpha}_t)\epsilon_t$, where ϵ_t represents a random noise sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Subsequently, z_t was fed into the U-Net denoiser, and we extracted features from intermediate layers to serve as diffusion features for identifying dense point correspondences between the fixed and moving images.

B. Feature Point Extraction

The goal of this step was to identify candidate feature points in the moving image from which we identified corresponding points in the fixed image. The SIFT algorithm [5] is a widely used feature extraction algorithm in computer vision for detecting and describing local features in images that are invariant to scaling and rotation. We first employed the SIFT

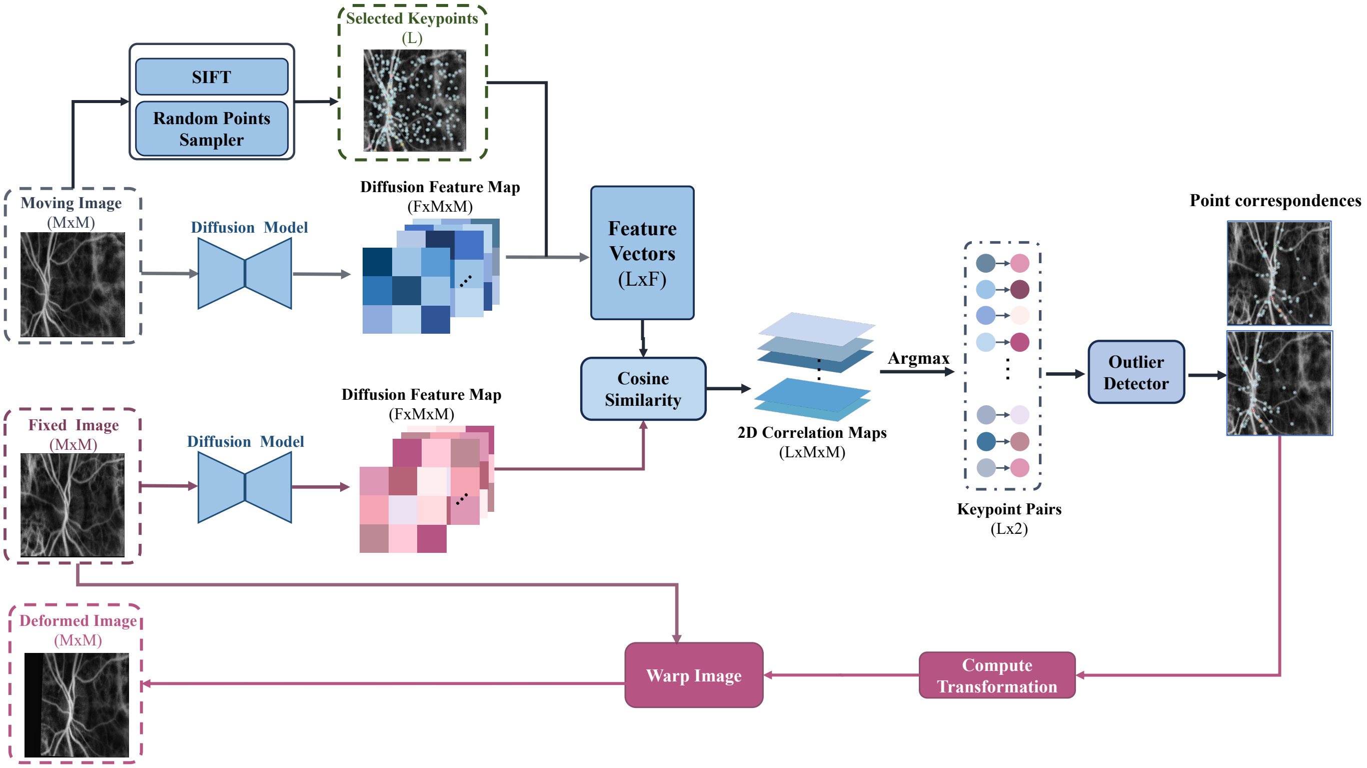

Fig. 2. Overview of the proposed RetinaRegNet model for retinal image registration.

algorithm to detect K feature points in the moving image, ensuring that the distance between any two feature points was greater than a threshold value T_{sift}. For certain applications, such as retinal images, feature points detected by the SIFT algorithm were concentrated in the sparsely distributed vessel trees, which may have caused the estimated transformation to be less accurate in regions without densely distributed vessels. To address this matter, alongside the feature points detected by the SIFT algorithm, we randomly selected additional K feature points throughout the moving image to improve the estimation of point correspondences in homogeneous image regions. We demonstrated that utilizing a combination of SIFT feature points and randomly selected feature points improves the performance of our registration model.

C. Estimation of Point Correspondences

Given a high-resolution fixed image I_f of size $H_f \times W_f$ and a moving image I_m of size $H_m \times W_m$, we first resampled both images to the size of $M \times M$. Diffusion image features were extracted using an intermediate layer of the denoising U-Net in a pre-trained stable diffusion model [40]. We upsampled the feature maps to the size of $M \times M$. The resulting diffusion features were denoted as D_f and D_m, with dimensions $F \times M \times M$, where F represents the size of each feature vector.

We sampled a list of $L = 2K$ feature points $p_1, \cdots, p_L$ in the moving image to find the point in the fixed image corresponding to each p_i. We computed the cosine similarity score between $D_m(p_i)$ and the feature vector of every pixel in the fixed image. The point with the highest similarity score was considered to correspond to p_i. For computational efficiency, we combined the diffusion feature vectors at all feature points into a matrix $[D_m(p_1), \cdots, D_m(p_L]$ of size $L \times F$. We then computed a 2D correlation map for each feature point, resulting in a correlation map $C = [C_1, \cdots, C_L]$ which consisted of L 2D correlation maps:

$$C = \textbf{norm}([D_m(p_1), \cdots, D_m(p_L)]) * \textbf{norm}(D_f^i) \tag{3}$$

where the **norm** operation was applied to each feature vector to convert it into a vector of unit norm, and $*$ is matrix multiplication.

The correlation map C is of dimension $L \times M \times M$, such that each feature point corresponds to a 2D correlation map. We then applied the Argmax operation to each of the L correlation maps to obtain the corresponding points in the fixed image, i.e.,

$$[p'_1, \cdots, p'_L] = [\text{Argmax}(C_1), \cdots, \text{Argmax}(C_L)] \tag{4}$$

D. Outlier Detection

Some of the point correspondences that were estimated via the 2D correlation maps did not align with the ground truth and were deemed as outliers if their Euclidean distance surpasses a certain threshold. We designed two algorithms to effectively eliminate these outliers from the estimated correspondences, which enhanced the robustness of the transformation estimation between image pairs.

a) Inverse Consistency.: To refine the estimated point correspondences, we adopted the concept of inverse consistency [12], which was originally developed for pairwise image registration. Inverse consistency in image registration involves jointly estimating forward and reverse transformations between two images, ensuring that these transformations are inverses of each other. In this paper, we applied this concept to the estimation of point correspondences between two images. For each feature point p_i in the moving image, we determined its corresponding point p'_i in the fixed image using the aforementioned method. Subsequently, we reversed the roles of the fixed and moving images, employing the same method to locate the point p''_i in the moving image that corresponds to p'_i. A necessary criterion for the accuracy of the corresponding point detector was the proximity of p_i to p''_i. We considered the point correspondence (p_i, p'_i) accurate if it met the following condition; otherwise, we excluded this correspondence from the estimation of the geometric transformation between the two images:

$$||p_i - p''_i||_2 \leq T_{IC}, \tag{5}$$

where T_{IC} represents the inverse consistency threshold.

b) Transformation Based Outlier Detector.: Given a set of estimated point correspondences between two images, our goal was to accurately estimate a smooth geometric transformation. However, this estimation is sensitive to outliers in the point correspondence. The RANSAC algorithm [13] is widely used to handle outlier data. It iteratively selects a random subset of the original data, fits a model, and then evaluates how many of the remaining data points conform to this model within a predefined tolerance. While conceptually straightforward and easy to implement, RANSAC can be computationally intensive for large datasets or complex models. Its performance heavily relies on the chosen threshold for distinguishing inliers from outliers and the number of iterations.

We observed that the transformation estimation was most sensitive to outliers when the distance between the true corresponding point and the estimated corresponding point is large. Taking this into account, we proposed a simple yet effective transformation-based outlier detector for refining point correspondences. Formally, given a set of estimated point correspondences (p_i, p'_i) between two images, where p_i and p'_i are corresponding points, our method first estimated a global transformation (e.g., affine) ϕ using all the point correspondences. The global transformation was then applied to each point p_i, resulting in transformed points $\phi(p_i)$. The fidelity of each correspondence was evaluated by computing the Euclidean distance between $\phi(p_i)$ and p'_i. Correspondences for which this distance exceeds a predefined threshold T_{trans} were considered outliers and subsequently removed from the correspondence set. This process can be summarized by the formula:

$$\text{Outlier} = (p_i, p'_i), |, ||\phi(p_i) - p'_i||_2 \geq T_{trans} \tag{6}$$

E. Multi-Stage Image Registration Framework

We designed our registration model as a multi-stage framework to handle large variations in the underlying transformation between images through a coarse-to-fine strategy. In the first stage, we estimated a global transformation ψ_{global} between the two images, where estimating such a global transformation is less sensitive to outliers than a more complex local transformation. In the second stage, we estimated a local deformation ψ_{local} between the globally aligned image pairs. The composition of these two transformations gave the final transformation between the two images, i.e., $\psi = \psi_{local} \circ \psi_{global}$. We have implemented four distinct types of transformation models: affine transformation (6 degrees of freedom), homography transformation (8 degrees of freedom), quadratic transformation (12 degrees of freedom), and third-order polynomial transformation (20 degrees of freedom). For all experiments in this study, we chose the homogeneous transformation model in the first stage for global alignment, followed by the third-order polynomial transformation model in the second stage to refine local feature alignment.

IV. Experiments and Results

A. Datasets

We used three distinct retinal image registration datasets for model evaluation.

a) ***FIRE Dataset:*** The first dataset we used is the public FIRE (Fundus Image Registration Dataset) dataset [14] that contains 129 color fundus images that form 134 image pairs. Acquired using a Nidek AFC-210 fundus camera at the Papageorgiou Hospital from 39 patients, the dataset offered a resolution of 2912×2912 pixels with a field of view (FOV) of 45°. The FIRE dataset provided ground truth correspondences for 10 landmark points for each image pair. These landmark points were manually selected by an annotator, primarily focusing on vessels and crossings, to ensure broad coverage of the overlapping image areas.

Image pairs in the FIRE dataset were divided into three categories. Class S consists of 71 image pairs characterized by high spatial overlap (greater than 75%) and no significant visual anatomical differences, making them easy to register. Class P consists of 49 pairs with a small overlap (less than 75%), making these pairs hard to register. Finally, Class A includes 14 image pairs with considerably large overlap. This exhibits visual anatomical differences which make them moderately hard to register.

b) ***FLoRI21 Dataset:*** The second dataset we used was the public FLoRI21 (Fluorescein-angiography Longitudinal Retinal Image 2021) dataset [15] which consists of 15 pairs of ultra-wide-field fluorescein angiography (FA) images taken 24 weeks apart for each subject. These images were captured using Optos California (Nikon Co. Ltd, Japan) and 200Tx cameras. Each image pair had one montage FA image (4000 × 4000) used as the fixed image, and several raw FA images (3900 × 3072) used as the moving images. Ten pairs of landmark points were chosen at retinal vessel bifurcations, ensuring coverage of the entire overlapping image area.

c) ***LSFG Dataset:*** The third dataset we used was an Institutional Review Board-approved laser speckle flowgraphy (LSFG) dataset that contains 15 pairs of longitudinal LSFG images from patients with uveal melanoma. These patients were treated at the University of Iowa with ^{125}I-plaque brachytherapy. LSFG (Softcare Co., Japan) is non-invasive and can measure mean blur rate, which is linearly proportional to blood flow velocity at the retina and optic nerve head in a video clip of 118 time frames. We averaged each pixel value along the aligned time-sequence frames of the LSFG video to create a 2D blood flow map (751 × 420). Each LSFG image pair has 6-10 ground truth landmark correspondences which were selected by a human annotator.

B. Methods for Comparison

We compared our RetinaRegNet model with six recent image registration methods. Among these, the GFEMR algorithm [20] tackles the registration problem through a probabilistic approach which integrates manifold regularization to preserve the retina's intrinsic geometry. Another method, ASpanFormer [46], uses an adaptive span transformer to find point correspondences. This employs CNN encoders and iterative global-local attention blocks which utilize auxiliary flow maps to adapt local attention span based on matching uncertainty. Another method we considered was SuperGlue [47], a neural network that matches two sets of local features by jointly finding correspondences and rejecting non-matchable points. Additionally, SuperRetina [9], employs a trainable keypoint detector and descriptor. It is trained using semi-supervised learning with a combination of labeled and unlabeled images; however, it requires manual or automatic labeling and is subject to keypoint detection errors and mismatches. Moreover, we also considered the LoFTR algorithm [27], which uses a transformer network to create feature descriptors from both images, establishing dense correspondences at a coarse level before refining them. Lastly, GeoFormer [26], another general-purpose method based on LoFTR, leverages RANSAC geometry to identify attentive regions and uses the transformer's cross-attention mechanism for feature matching during transformation estimation.

C. Evaluation Metrics

a) Mean Landmark Error: Given a pair of fixed and moving images, we have a set of N manually annotated landmark pairs. Each landmark point in one image is denoted as (x_i, y_i), and its corresponding ground truth landmark point in the other image is denoted as (x_i', y_i'). Let ψ represent a geometric transformation that we have estimated to align the fixed image with the moving image. The mean landmark error (MLE) for this pair of images is defined as the mean of the Euclidean distance between each transformed point $(x_i'', y_i'') := \psi(x_i, y_i)$ and its corresponding ground truth point (x_i', y_i'). Mathematically, the MLE can be expressed as:

$$MLE = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(x_i'' - x_i')^2 + (y_i'' - y_i')^2} \tag{7}$$

b) Area Under the Curve: The normalized area under the curve (AUC) evaluates an image registration method by assessing its performance at various MLE thresholds ($[0, T_{AUC}]$). For each threshold value, a registration is considered accurate if its MLE is below the threshold. A registration accuracy, indicating the proportion of accurately registered images, is computed for the entire image set at each threshold. The normalized AUC metric integrates these success rates across all MLE thresholds which provides a comprehensive assessment of registration accuracy. Mathematically, the normalized AUC can be expressed as

$$AUC = \frac{1}{T_{AUC}} \int_0^{T_{AUC}} RA(T) dT \tag{8}$$

where $RA(T)$ denote the registration accuracy at the threshold value T.

c) Registration Success Rate: We categorize the image registration of an image pair as unsuccessful under either of two conditions. Firstly, the registration is considered failed if there are insufficient point correspondences to estimate the required geometric transformation. For example, a minimum of 3 pairs of points is necessary to estimate an affine transformation, and at least 6 pairs are needed for a quadratic transformation. Secondly, the registration is considered unsuccessful if the mean landmark error exceeds a predetermined threshold, T_{SR}, which is determined by the specific characteristics of the registration dataset. By applying these criteria, we can calculate the success rate of each registration method across different datasets. In this paper, we chose $T_{SR} = \frac{1}{2} T_{AUC}$.

D. Implementation Details

The proposed model was implemented using PyTorch and Diffusers on a computing node equipped with two CPU cores, 25 GB of RAM, and a NVIDIA A100 GPU. For image registration, images from the FIRE, FLoRI21, and LSFG datasets were resampled to resolutions of 920×920, 1024×1024, and 740×740 pixels, respectively. After registration, the coordinates of corresponding points were upsampled to their original image sizes for transformation estimation and landmark error evaluation.We extracted $K = 1000$ feature keypoints from the moving image using the SIFT algorithm, ensuring that the distance between any two keypoints was greater than 10 pixels. Additionally, we randomly selected another set of $K = 1000$ feature points from the moving image.Diffusion image features were extracted using the third layer of the denoising U-Net at time step $t = 1$ for each dataset. The inverse consistency threshold was set at $T_{IC} = 3$ pixels for all datasets. For the computation of the AUC, threshold values (T_{AUC}) were set at 25, 100, and 25 for the FIRE, FLoRI21, and LSFG datasets, respectively. For the transformation-based outlier detector, we selected the affine transformation model for both registration stages and set threshold values (T_{trans}) to 25 and 15 for the FIRE dataset, 40 and 30 for FLoRI21, and 25 and 25 for the LSFG datasets, respectively. A higher threshold was selected for the FLoRI21 dataset to accommodate its predominantly non-affine deformation in image pairs.

E. Quantitative Results

a) Results on the FIRE Dataset.: Table I demonstrates that our model significantly outperformed all other methods in the challenging P class of hard-to-register cases, delivering the highest AUC of 0.856, compared to the 0.697 AUC of the second-best method, GFEMR. These results underscore our model's superior ability to handle large displacement deformations and small overlaps between images. Furthermore, in the other two classes, our model achieved state-of-the-art performance, significantly boosting the overall AUC across the FIRE dataset from 0.783 (achieved by SuperRetina) to 0.901, and reducing the mean landmark error from 5.99 to 2.97. Additionally, our model proved to be the most robust, achieving a registration success rate of 99.25%, significantly surpassing the 91.04% success rate of GFEMR. In conclusion, these compelling findings not only demonstrate the exceptional performance and reliability of our model across various challenging scenarios within the FIRE dataset but also establish it as a groundbreaking solution in the field of retinal image registration, setting a new benchmark for future innovations and research.

TABLE I
Registration results on the FIRE Dataset. MLE: mean landmark error. Success rate threshold: $T_{SR} = 12.5$.

Method	Easy	Moderate	Hard	FIRE	MLE	Success Rate
GFEMR [20]	0.849	0.534	0.697	0.761	6.11	91.04%
ASpanFormer [46]	0.878	0.728	0.057	0.562	18.05	64.92%
SuperGlue [47]	0.810	0.602	0.417	0.645	9.59	70.89%
LoFTR [27]	0.941	0.746	0.338	0.703	8.29	74.06%
SuperRetina [9]	0.949	0.774	0.538	0.783	5.99	84.58%
GeoFormer [26]	0.923	0.757	0.577	0.781	5.99	88.72%
Ours	**0.951**	**0.805**	**0.856**	**0.901**	**2.97**	**99.25%**

b) Results on the FLoRI21 Dataset: Challenges in registering image pairs in the FLoRI21 dataset are primarily from the large displacement and scaling deformations between the montage FA images (sized at 4000×4000 pixels) and the raw FA images (sized at 3900×3072 pixels). The results presented in Table II show that our model significantly outperformed the second-best performing model, GeoFormer, in several key metrics. Specifically, our model achieved an AUC of 0.868 compared to GeoFormer's 0.640, a mean landmark error of 13.83 versus 36.49, and a 100% registration success rate, substantially higher than GeoFormer's 93.33%. It was notable that AspanFormer, GFEMR, LoFTR, SuperGlue, and SuperRetina all struggled to achieve accurate registration in at least 20% of the cases. In particular, GFEMR, which is customized for registering color fundus images, was not well-suited for addressing the large scaling and displacement deformations in another retinal dataset. This resulted in a high mean landmark error of 71.58 and a low success rate of 6.67%. These results underscore the achievements of our proposed model, which not only tackled the specific challenges of the FLoRI21 dataset but also set a new benchmark in this domain, achieving state-of-the-art results with a remarkable 100% success rate.

c) Results on the LSFG Dataset.: Image pairs in the LSFG dataset represent longitudinal measurements of blood flow in the eyes. The primary challenges in registering these LSFG image pairs arise from two factors: (1) positional shifts across images, attributable to the different times at which they were acquired, and (2) intensity variations across images, due to changes in blood flow over time. The results, as detailed in Table III, demonstrate that our method outperformed all other methods in several key metrics. Specifically, our model achieved an AUC of 0.861, compared to the second-best method, LoFTR's 0.853, and a mean landmark error of 4.00, slightly better than LoFTR's 4.23. These impressive results not only highlight the robustness of our model in handling the unique challenges presented by LSFG image pairs but also highlight its potential as a valuable tool in advancing the accuracy and efficiency of blood flow analysis in ophthalmic research.

TABLE II
Registration results on the FLoRI21 Dataset. MLE: mean landmark error. Success rate threshold: $T_{SR} = 50$.

Method	AUC	MLE	Success Rate
GFEMR [20]	0.299	71.58	6.67%
ASpanFormer [46]	0.548	45.59	60.00%
SuperGlue [47]	0.604	40.09	80.00%
LoFTR [27]	0.486	51.99	60.00%
SuperRetina [9]	0.590	41.47	80.00%
GeoFormer [26]	0.640	36.49	93.33%
Ours	**0.868**	**13.83**	**100%**

TABLE III
Registration results on the LSFG Dataset. MLE: mean landmark error. Success rate threshold: $T_{SR} = 12.5$.

Method	AUC	MLE	Success Rate
GFEMR [20]	0.587	15.29	73.33%
ASpanFormer [46]	0.813	5.07	93.33%
SuperGlue [47]	0.752	6.66	93.33%
LoFTR [27]	0.853	4.23	100%
SuperRetina [9]	0.843	4.32	100%
GeoFormer [26]	0.845	4.40	100%
Ours	**0.861**	**4.00**	**100%**

d) Computational Complexity: For each retinal registration dataset, we computed the average per-case running time of each image registration method. Results in Table IV suggest that although our RetinaRegNet has achieved state-of-the-art performance on all three retinal registration datasets, it is the slowest method among all, taking up to 20 seconds to register each image pair.

TABLE IV
Per-case computational complexity (seconds).

Method	FIRE	FLoRI21	LSFG
GFEMR [20]	3.032	5.313	0.9211
ASpanFormer [46]	1.00	1.00	1.00
SuperGlue [47]	1.00	4.00	5.00
LoFTR [27]	0.363	**0.508**	0.324
SuperRetina [9]	1.172	2.157	1.148
GeoFormer [26]	**0.36**	0.53	**0.18**
Ours	13.00	19.00	18.00

F. Qualitative Results

Figure 3 shows the registration results of one representative case from each of the three retinal image registration datasets. The corresponding landmarks in the fixed, moving, and deformation images illustrate the high accuracy of our registration model for all three cases. The first row demonstrates that our model has successfully estimated the large displacement deformation between an image pair from the hard-to-register (P) category in the FIRE dataset. The second row shows how our model has accurately aligned an image pair with complex global and local deformations from the FLoRI21 dataset. The last row demonstrates the superior performance of our model on a pair of images in our LSFG dataset with large shifting. The consistently high performance of our model across all three datasets demonstrates its accuracy, robustness, as well as its potential to be utilized in other retinal image registration tasks.

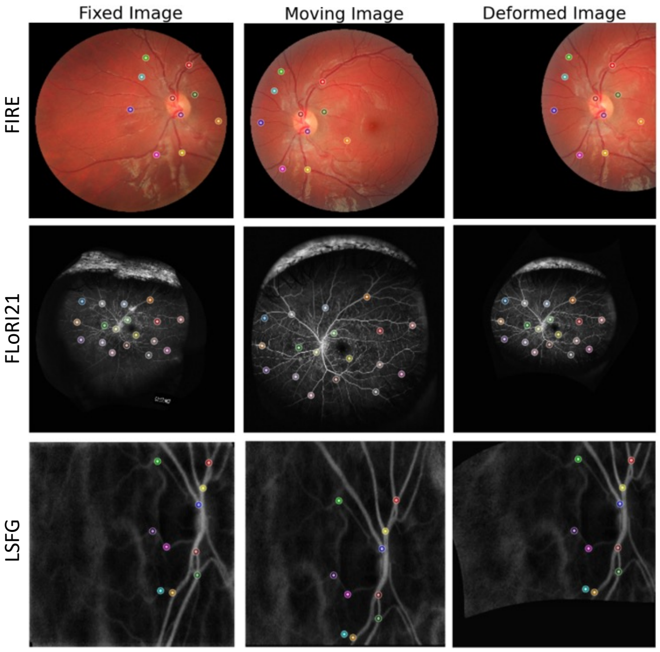

Fig. 3. Registration results of our model are shown for the FIRE (first row), FLoRI21 (second row), and LSFG (third row) datasets. For each landmark, we assigned a unique color, consistent across fixed, moving, and deformed images, enabling clear observation of the alignment of each landmark pair.

V. Ablation Study

A. Impact of Image Feature Extractor

We conducted an in-depth evaluation of the effectiveness of using different feature types in RetinaRegNet, including CNN features [48], vision transformers features [49], and diffusion features [44]. Specifically, we utilized a pretrained VGG19 model for extracting CNN features, a pretrained 8-patch vision transformer network (ViT-S/8) for extracting DINO-ViT features, and a pretrained latent stable diffusion model (SDv2-1) for extracting diffusion features. We applied RetinaRegNet to re-register all pairs of images by replacing the diffusion features with the CNN features and DINO-ViT features. Registration results in Table V suggested that diffusion features achieved the best performance across all three datasets. RetinaRegNet achieved similar registration results on the FIRE and FLoRI21 datasets when using the CNN features and diffusion features. However, the AUC of RetinaRegNet decreased significantly from 0.861 to 0.791 on the LSFG dataset when using CNN features instead of diffusion features. Notably, the DINO-ViT features extracted using only the vision transformers did not exhibit the same level of accuracy compared to the CNN features and diffusion features extracted with convolutional networks involved. Therefore, we chose to use diffusion features (DIFT) in the proposed RetinaRegNet registration framework.

TABLE V
AUC Values of RetinaRegNet Using Different Image Feature Extractors.

Features	FIRE	FLoRI21	LSFG
CNN Features	0.894	0.860	0.791
DINO-ViT Features	0.410	0.537	0.199
Diffusion Features	**0.901**	**0.868**	**0.861**

B. Impact of Diffusion Feature Size

Image features from various layers of the decoder in the stable diffusion model exhibit differing sizes, each smaller than the input image by a downsampling factor. We assessed how variations in diffusion feature size affect RetinaRegNet's performance. Ideally, image features should capture high-level features without being either too shallow or too deep. Shallowness corresponds to a small downsampling factor with limited learned representations and deepness, corresponds to a large downsampling factor unsuitable for dense predictions due to excessively small image features.

Results in Table VI indicated that the diffusion features corresponding to a downsampling factor of 32 exhibited the best registration performance across all three retinal datasets; these features provide a balance between detailed representation and abstraction. Diffusion features corresponding to a downsampling factor of 16 also achieved promising registration results, ranking second best. Notably, the features extracted from the first block and the last block both resulted in poor performance. These blocks represent extremely low and high-level diffusion features, respectively.

TABLE VI
AUC Values of RetinaRegNet Using Diffusion Features Extracted by Different Downsampling Factor.

Feature Selection	Downsampling Factor	FIRE	FLoRI21	LSFG
Block-1 Features	8	0.086	0.000	0.000
Block-2 Features	16	0.742	0.796	0.834
Block-3 Features	**32**	**0.901**	**0.868**	**0.861**
Block-4 Features	64	0.554	0.301	0.051

C. Impact of Key Model Components

We investigated the impact of each of the four key components in RetinaRegNet, including the inverse consistency constraint, the random feature point sampling strategy, the transformation-based outlier detector, and the multi-stage registration framework. We reran RetinaRegNet on all image pairs using one of the following: (1) removing the inverse consistency constraint; (2) omitting the random sampling strategy; (3) replacing the affine transformation-based outlier detector with the RANSAC outlier detector; and (4) replacing the multi-stage framework with a single-stage design.

Results in Table VII indicate that all four components significantly contribute to the high performance of the RetinaRegNet model, although their impact varies across different retinal datasets. For instance, in both the FIRE dataset and the FLoRI21 dataset, the multi-stage image registration design notably improved performance by a significant margin, likely due to the complexity of deformations between image pairs in those two datasets. Conversely, in the LSFG dataset, substituting our affine transformation-based outlier detector with the RANSAC detector resulted in a significantly decreased AUC from 0.861 to 0.416, possibly due to the simpler nature of deformations in these images.

TABLE VII
Impact of Each Model Component on AUC Values

Method	FIRE	FLoRI21	LSFG
RetinaRegNet (proposed)	**0.901**	**0.868**	**0.861**
RetinaRegNet w/o inverse consistency	0.884	0.847	0.850
RetinaRegNet w/o random point sampling	0.890	0.866	0.842
RetinaRegNet w/o affine outlier detector	0.805	0.835	0.416
RetinaRegNet w/o multi-stage design	0.760	0.644	0.845

VI. Discussion

A. Clinical Implications

The retina, a light-sensitive layer at the back of the eye, plays a vital role in our vision [50]. However, retinal diseases like diabetic retinopathy, glaucoma, and age-related macular degeneration [51]–[54] can lead to vision impairment. Early detection and treatment are crucial, emphasizing the need for monitoring retinal blood vessels. Retinal image registration, a technique used to align retinal images, facilitates accurate diagnosis and disease monitoring. By precisely tracking changes over time and assessing treatment response, clinicians can effectively manage retinal pathologies. Our RetinaRegNet, achieving state-of-the-art registration accuracy, holds immense potential to revolutionize clinical practice by empowering clinicians with enhanced diagnostic tools and treatment monitoring capabilities, ultimately leading to improved patient outcomes.

B. Limitations and Future Directions

Our study has several limitations. First, RetinaRegNet is computationally expensive, requiring between 10 to 20 seconds per image pair due to the computation of a 2D correlation map for each selected feature point in the moving images. In this paper, we opted for 2000 feature points (1000 for SIFT; 1000 for random point sampling), however users have the flexibility to choose a lower number of points, particularly when the underlying transformation between the two images is an affine or homography transformation. We will assess the trade-off between computational efficiency and registration accuracy in our future work. Second, our focus was solely on mono-modal image registration, which overlooked the generalization of our method to multi-modal retinal image registration challenges. We plan to evaluate the effectiveness of RetinaRegNet in multi-modal registration by initially training and employing an image-to-image translation network to unify different modalities into a single modality, followed by applying RetinaRegNet. Third, theoretically, RetinaRegNet can serve as a general image registration approach that can be extended to various other registration tasks. In the future, we aim to evaluate RetinaRegNet's performance in other medical image registration contexts (e.g., registering 2D histopathology and MRI images) and nature image registration contexts (e.g., aligning images captured by unmanned aerial vehicles from different angles or times).

VII. Conclusion

We have developed RetinaRegNet, a versatile retinal image registration model that does not require training on any retinal images. RetinaRegNet's innovative use of image features from the stable diffusion model ensures reliable estimation of point correspondences. The combination of the inverse consistency constraint and the transformation-based outlier detector further improves point correspondence estimation by effectively removing outliers. Additionally, the two-stage registration framework allows our model to handle more complex transformations than affine or homography transformations. The effectiveness of RetinaRegNet is evidenced by its state-of-the-art performance across all three retinal image registration datasets, highlighting its potential as a powerful tool for a wide range of applications in the field of image registration.

References

[1] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Physics in Medicine & Biology*, vol. 46, no. 3, p. R1, 2001.

[2] F. P. Oliveira and J. M. R. Tavares, "Medical image registration: a review," *Computer methods in biomechanics and biomedical engineering*, vol. 17, no. 2, pp. 73–93, 2014.

[3] S. K. Saha, D. Xiao, A. Bhuiyan, T. Y. Wong, and Y. Kanagasingam, "Color fundus image registration techniques and applications for automated analysis of diabetic retinopathy progression: A review," *Biomedical Signal Processing and Control*, vol. 47, pp. 288–302, 2019.

[4] A. Hering *et al.*, "Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 697–712, 2022.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[6] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer, 2006, pp. 404–417.

[7] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "Kaze features," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer, 2012, pp. 214–227.

[8] S. Wang, H. You, and K. Fu, "Bfsift: A novel method to find feature matches for sar image registration," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 4, pp. 649–653, 2011.

[9] J. Liu, X. Li, Q. Wei, J. Xu, and D. Ding, "Semi-supervised keypoint detector and descriptor for retinal image matching," in *European Conference on Computer Vision*. Springer, 2022, pp. 593–609.

[10] S. A. Nasser, N. Gupte, and A. Sethi, "Reverse knowledge distillation: Training a large model using a small one for retinal image matching on limited data," *arXiv preprint arXiv:2307.10698*, 2023.

[11] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," *Advances in neural information processing systems*, vol. 31, 2018.

[12] G. E. Christensen and H. J. Johnson, "Consistent image registration," *IEEE transactions on medical imaging*, vol. 20, no. 7, pp. 568–582, 2001.

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

[14] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A. A. Argyros, “Fire: fundus image registration dataset,” *Modeling and Artificial Intelligence in Ophthalmology*, vol. 1, no. 4, pp. 16–28, 2017.

[15] L. Ding *et al.*, “Flori21: Fluorescein angiography longitudinal retinal image registration dataset,” *IEEE Dataport*, 2021.

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

[17] Y.-M. Zhu, “Mutual information-based registration of temporal and stereo retinal images using constrained optimization,” *Computer methods and programs in biomedicine*, vol. 86, no. 3, pp. 210–215, 2007.

[18] P. A. Legg, P. L. Rosin, D. Marshall, and J. E. Morgan, “Improving accuracy and efficiency of mutual information for multi-modal retinal image registration using adaptive probability density estimation,” *Computerized Medical Imaging and Graphics*, vol. 37, no. 7-8, pp. 597–606, 2013.

[19] G. Molodij, E. N. Ribak, M. Glanc, and G. Chenegros, “Enhancing retinal images by nonlinear registration,” *Optics Communications*, vol. 342, pp. 157–166, 2015.

[20] J. Wang *et al.*, “Gaussian field estimator with manifold regularization for retinal image registration,” *Signal Processing*, vol. 157, pp. 225–235, 2019.

[21] C. Hernandez-Matas, X. Zabulis, and A. A. Argyros, “Rempe: Registration of retinal images through eye modelling and pose estimation,” *IEEE journal of biomedical and health informatics*, vol. 24, no. 12, pp. 3362–3373, 2020.

[22] D. Motta, W. Casaca, and A. Paiva, “Vessel optimal transport for automated alignment of retinal fundus images,” *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6154–6168, 2019.

[23] B. Zou, Z. He, R. Zhao, C. Zhu, W. Liao, and S. Li, “Non-rigid retinal image registration using an unsupervised structure-driven regression network,” *Neurocomputing*, vol. 404, pp. 14–25, 2020.

[24] L. Chen, X. Huang, and J. Tian, “Retinal image registration using topological vascular tree segmentation and bifurcation structures,” *Biomedical Signal Processing and Control*, vol. 16, pp. 22–31, 2015.

[25] Y. Wang *et al.*, “A segmentation based robust deep learning framework for multimodal retinal image registration,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1369–1373.

[26] J. Liu and X. Li, “Geometrized transformer for self-supervised homography estimation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9556–9565.

[27] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, “Loftr: Detector-free local feature matching with transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

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

[29] M. Santarossa *et al.*, “Medregnet: Unsupervised multimodal retinal-image registration with gans and ranking loss,” in *Medical Imaging 2022: Image Processing*, vol. 12032. SPIE, 2022, pp. 321–333.

[30] K. Han *et al.*, “Scnet: Learning semantic correspondence,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1831–1840.

[31] Y. Liu, L. Zhu, M. Yamada, and Y. Yang, “Semantic correspondence as an optimal transport problem,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4463–4472.

[32] D. Zhao, Z. Song, Z. Ji, G. Zhao, W. Ge, and Y. Yu, “Multi-scale matching networks for semantic correspondence,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3354–3364.

[33] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[34] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” *arXiv preprint arXiv:2011.13456*, 2020.

[35] D. Kingma, T. Salimans, B. Poole, and J. Ho, “Variational diffusion models,” *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[36] K. Gong, K. Johnson, G. El Fakhri, Q. Li, and T. Pan, “Pet image denoising based on denoising diffusion probabilistic model,” *European Journal of Nuclear Medicine and Molecular Imaging*, pp. 1–11, 2023.

[37] H. Li *et al.*, “Srdiff: Single image super-resolution with diffusion probabilistic models,” *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[38] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, “Repaint: Inpainting using denoising diffusion probabilistic models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 461–11 471.

[39] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, “Segdiff: Image segmentation with diffusion probabilistic models,” *arXiv preprint arXiv:2112.00390*, 2021.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[41] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, “Dreamfusion: Text-to-3d using 2d diffusion,” *arXiv preprint arXiv:2209.14988*, 2022.

[42] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, “Null-text inversion for editing real images using guided diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6038–6047.

[43] X. Yang and X. Wang, “Diffusion model as representation learner,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 938–18 949.

[44] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan, “Emergent correspondence from image diffusion,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[45] E. Hedlin *et al.*, “Unsupervised semantic correspondence using stable diffusion,” *arXiv preprint arXiv:2305.15581*, 2023.

[46] H. Chen *et al.*, “Aspanformer: Detector-free image matching with adaptive span transformer,” 2022.

[47] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superglue: Learning feature matching with graph neural networks,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4937–4946.

[48] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[49] M. Caron *et al.*, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[50] S. Kawamura and S. Tachibanaki, “Explaining the functional differences of rods versus cones,” *Wiley Interdisciplinary Reviews: Membrane Transport and Signaling*, vol. 1, no. 5, pp. 675–683, 2012.

[51] E. J. Duh, J. K. Sun, and A. W. Stitt, “Diabetic retinopathy: current understanding, mechanisms, and treatment strategies,” *JCI insight*, vol. 2, no. 14, 2017.

[52] M. Imran, A. Ullah, M. Arif, R. Noor *et al.*, “A unified technique for entropy enhancement based diabetic retinopathy detection using hybrid neural network,” *Computers in Biology and Medicine*, vol. 145, p. 105424, 2022.

[53] K. K. Chan, F. Tang, C. C. Tham, A. L. Young, and C. Y. Cheung, “Retinal vasculature in glaucoma: a review,” *BMJ open ophthalmology*, vol. 1, no. 1, p. e000032, 2017.

[54] L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, and T. Y. Wong, “Age-related macular degeneration,” *The Lancet*, vol. 379, no. 9827, pp. 1728–1738, 2012.