A Retinal Image Sequence Registration Method Based on Longitudinal 3D Fundoscopy Scene Modeling

Zengshuo Wang, *Graduate Student Member, IEEE*, Haohan Zou, Xin Zhao, *Member, IEEE*, Yan Wang, and Mingzhu Sun, *Member, IEEE*

***Abstract*—Accurate and reliable global registration is the basis and prerequisite for longitudinal studies of related diseases that are based on retinal image sequences. However, most existing retinal image registration methods do not consider the implicit inherent spatiotemporal transformation process between the retinal images to be registered, resulting in low accuracy or poor reliability. In this paper, we propose a 3D spatiotemporal model that simulates the fundoscopy scene which is on a longitudinal time-axis. There are two main components: 1) a dynamic eyeball model that simulates the changes in eyeball shape owing to natural growth or disease progression (e.g., myopia); and 2) a camera array model that simulates the changes in eyeball pose and fundus camera parameters between each fundoscopy. Based on the 3D spatiotemporal model, we implement a retinal image sequence registration framework using a frame-to-reference and frame-to-frame joint registration strategy. The framework uses the matched keypoints as a medium, and adopts pose estimation and optimization algorithm to align all images in the sequence into the same longitudinal fundoscopy scene. Benefiting from the spatiotemporal simulation of the longitudinal fundoscopy scene, the proposed method can generate a large amount of realistic synthetic data given only one retinal image. We conduct comprehensive experiments on two image pair registration datasets and three image sequence registration datasets. The results show that the proposed method achieves state-of-the-art registration accuracy, reliability, and applicability. To promote the study in related fields, we make all codes and datasets publicly available.**

***Index Terms*—Retinal image sequence registration, longitudinal fundoscopy scene, 3D spatiotemporal model, dynamic eyeball model, highly realistic synthetic data generation.**

Received 28 March 2025; revised 9 December 2025, 12 April 2026, and 24 June 2026; accepted 15 July 2026. Date of publication 24 July 2026; date of current version 30 July 2026. This work was supported in part by Shenzhen Science and Technology Program under Grant JCYJ20240813165501003 and in part by the Open Fund of Nankai University Optometry & Vision Science Institute under Grant NKSGY202404. The associate editor coordinating the review of this article and approving it for publication was Dr. Xulei Yang. *(Corresponding authors: Yan Wang; Mingzhu Sun.)*

Zengshuo Wang, Xin Zhao, and Mingzhu Sun are with the National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, the Engineering Research Center of Trusted Behavior Intelligence, Ministry of Education, Tianjin Key Laboratory of Intelligent Robotics, and the Institute of Robotics and Automatic Information System, Nankai University, Tianjin 300350, China, and also with the Institute of Intelligence Technology and Robotic Systems, Shenzhen Research Institute of Nankai University, Shenzhen 518083, China (e-mail: sunmz@nankai.edu.cn).

Haohan Zou and Yan Wang are with Tianjin Eye Hospital, Tianjin Eye Institute, Tianjin Key Laboratory of Ophthalmology and Visual Science, Nankai University Affiliated Eye Hospital, Tianjin 300020, China, and also with the Nankai University Eye Institute, Nankai University, Tianjin 300350, China (e-mail: wangyan7143@vip.sina.com).

Digital Object Identifier 10.1109/TIP.2026.3715100

I. Introduction

THE longitudinal study based on retinal image sequences [1] is an important way to analyze the retinal structural changes that are associated with the progression of related diseases, and mainly includes two key steps: (1) global registration [2] of all the retinal images in each sequence; (2) extraction and analysis of retinal structural changes [1]. For example, as shown in Fig. 1, with the onset and progression of myopia, the main anatomical changes in retina include: (1) the shape of the optic disc changes from approximately circular to vertically oval [3], [4]; (2) parapapillary atrophy occurs and gradually expands [5], [6]; (3) the retinal blood vessels gradually straighten [7]. Obviously, accurate and reliable global registration is the basis and prerequisite for the subsequent accurate extraction and effective analysis of retinal structural changes, which is the focus of this paper.

We believe that, ideally, the transformations that result from the global registration of retinal image sequences for longitudinal study need to satisfy two requirements: (1) completely compensate for the pixel position deviation of the corresponding points in images due to the differences in parameters of camera imaging, eyeball pose, eyeball shape, optical refraction, etc., between each fundoscopy on the longitudinal time axis; (2) only retain the pixel position movement of the corresponding points in images caused by the anatomical structural changes in retina itself. Therefore, the transformation model for image registration needs to have the ability to align images as much as possible without destroying the original anatomical structure differences. Therefore, the design of transformation model for retinal image registration in longitudinal study faces great challenges.

Applying a transformation model with lower complexity can easily lead to under-registration [8], that is, the registration error between images is large. At this time, the pixel movement corresponding to the retinal structural changes is mixed with the residual registration error, thereby resulting in a large amount of noise in the subsequently extracted retinal structural changes. Affine transformation is a common registration model with low complexity and is suitable only for the registration of retinal images with low resolution and narrow field of view (FOV) [9], [10]. This is because the retinal image with a large FOV is strongly affected by factors such as eyeball shape and optical distortion [10], [11]. In comparison, the quadratic

1941-0042 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

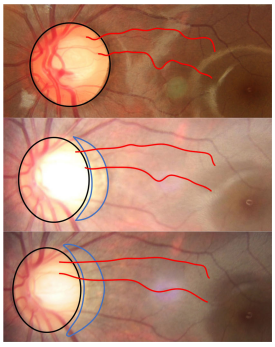

Fig. 1. The anatomical changes in retina with the onset and progression of myopia. From top to bottom, the degree of myopia gradually increases, and the corresponding spherical equivalent (SE) is −0.5, −2.75, and −3.875 respectively. The regions of interest are highlighted using colored lines. The black color represents the region of optic disc, the blue color represents the region of parapapillary atrophy, and the red color represents retinal blood vessels.

transformation [12], [13], [14], [15] has higher complexity, and the results of retinal image registration based on this model usually have higher accuracy. The main reason is that the quadratic transformation has better potential for approximating the 3D curved surface shape of the fundus [11], [16]. However, the quadratic transformation is still somewhat powerless for the registration of retinal images that are further superimposed with optical imaging distortion. In recent years, the more commonly used model in retinal image registration is homography transformation [17], [18], [19], which involves a 33 matrix that describes the mapping of corresponding points in two images of the same plane taken from different viewpoints [18]. In other words, homography transformation essentially includes the perspective projection process of a plane in space to the image planes of two cameras with different poses. This enables the homography transformation to correct the pixel position offset of corresponding points caused by the changes of eyeball pose and camera parameters in a relatively interpretable way. However, the homography transformation assumes that the scene is planar [18], and approximating the retina, which has a curved surface shape, as a plane has significant limitations in many applications.

In contrast to the above transformation model, applying a transformation model with higher complexity can easily lead to over-registration [20], [21], that is, the original anatomical structure differences between longitudinal images are destroyed. In this case, it will be impossible to extract the correct and complete retinal structural changes associated with the disease progression. Deformable registration [22], [23] does not involve the specific transformation formula, but directly generates the pixel-wise deformation field as the transformation model, which means that this transformation process has a high degree of freedom and is highly nonlinear. However, such a transformation process does not have any interpretability and just only blindly pursues pixel-level alignment between images. Therefore, this type of transformation model is usually applicable only to the registration of retinal images acquired during the same examination period [2], [24]. Or after longitudinal retinal images are globally registered, it is used for local registration to obtain the fundus deformation field associated with disease progression [25].

Designing a transformation model that simulates the actual physical processes in fundoscopy enables registering retinal images in a fully interpretable way. This type of transformation model can ensure to the greatest extent that the original retinal anatomical structure will not be destroyed. The registration accuracy depends mainly on the completeness of the transformation model in simulating the fundoscopy scene. Hernandez-Matas et al. [11] converted the change of eyeball pose during fundoscopy into the change of camera pose, and achieved registration by jointly estimating the relative pose of cameras and the parameters of ellipsoidal model of eye. This method, called REMPE [11], achieved competitive results with the best performing method at that time in a novel way. Afterward, we considered optical distortion during the imaging process of fundus camera and introduced a novel myopia development model that simulates changes in ocular axial length and fundus curvature during myopia development [26]. The registration framework constructed based on these models, RIRMD [26], achieved the best registration performance for retinal images with or without myopia development. In this paper, we aim to further develop a more comprehensive and general transformation model for the global registration of retinal images in longitudinal study. Specifically, we construct a 3D spatiotemporal model that simulates the longitudinal fundoscopy scene, according to the process of fundus camera imaging retina and the changes in eyeball shape during natural growth or disease progression. Finally, we implement a framework for **R**etinal **I**mage **Se**quence **R**egistration (**RISeR**) based on the 3D spatiotemporal model using the frame-to-reference and frame-to-frame joint registration strategy. The framework uses the matched keypoints between retinal images as a medium, and adopts pose estimation and optimization algorithm to align all images in the sequence into the same longitudinal 3D fundoscopy scene. Moreover, benefiting from the spatiotemporal simulation of the longitudinal fundoscopy scene, RISeR can generate many realistic synthetic image sequences given only one retinal image, which could be potentially applicable for method evaluation, deep learning-based model training and data augmentation. To demonstrate the effectiveness of RISeR, we conduct a comprehensive experimental evaluation on two publicly available retinal image pair registration datasets, two newly constructed retinal image sequence registration datasets (a real and a synthetic), and a publicly available retinal image sequence registration dataset. The results show that RISeR achieves state-of-the-art registration accuracy, reliability, and applicability.

In order to further promote the study in related fields, we make all the codes and datasets publicly available to the scientific community.

II. Related Work

A. Retinal Image Registration

Corresponding to the key steps of retinal image registration, the focus of existing work includes two main aspects: (1) characterization of the association between images; (2) construction of the transformation model between images.

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

1) Characterization of the Association Between Images: In existing retinal image registration methods, the ways of characterizing the association between images can be categorized into two types: keypoint-based (or feature-based) and intensity-based [19], [27]. **For keypoint-based ways,** there are typically two steps: first, detect the keypoints in images and compute the corresponding feature descriptors, commonly used algorithms include traditional methods such as SIFT [28] and SURF [29], and learning-based methods such as SuperPoint [30] and R2D2 [31]; then, match the keypoints between images [26], [32], commonly used algorithms include traditional methods such as the nearest-neighbor algorithm [28] combined with RANSAC [33], and learning-based methods such as SuperGlue [34]. Many studies focus on developing algorithms that are more suitable for keypoint extraction and matching in retinal images to improve the accuracy of retinal image registration. Chen et al. [12] proposed a Partial Intensity Invariant Feature Descriptor (PIIFD) and combined it with the Harris corner detection algorithm [35] to detect and describe keypoints in multimodal retinal images. Tsai et al. [15] improved keypoint extraction and matching in the GDB-ICP [36] algorithm and achieved better registration accuracy. Wang et al. [37] combined SURF and PIIFD to detect and describe keypoints, and designed a new robust point matching algorithm embedded in the proposed SURF-PIIFD-RPM framework. Inspired by SuperPoint, Liu et al. [17] proposed a semi-supervised deep learning-based keypoint detection and description algorithm specifically for retinal image matching, which was named SuperRetina. **For intensity-based ways,** they carefully designed the similarity measurement function based on the indicator such as Mutual Information (MI) [38], [39], [40], Phase Correlation (PC) [41], Mean Squared Difference (MSD) [42], Cross-Correlation (CC) [43], and then estimated the registration transformation by maximizing it [26], [44].

2) Construction of the Transformation Model Between Images: In the existing retinal image registration studies, traditional transformation models have been selected without any specialized design. In early years, Choe et al. [45] adopted affine transform to register multimodal retinal images. Later, Yang et al. [36] selected a more complex quadratic transform to align retinal images. Lin and Medioni [46] considered the local region of the retinal surface shown in the image as near planar, and applied the homography model to describe the transformation. In recent years, there are still many studies [17], [18], [19] selecting the homography transformation for retinal image registration. However, an increasing number of studies have suggested that the above models cannot meet the requirements of retinal image registration for transformation complexity, and they designed new registration strategies and transformation models. Ding et al. [44] proposed a hybrid two-stage strategy. The first stage estimates the third-order polynomial transformation model to coarsely align the images. The second stage estimates the second-order polynomial transformation model at the local patch level, and interpolates to generate the displacement vector field by integrating the transformations of all patches to finely align the images. Similarly, Liu et al. [24] first calculated the homography matrix for coarse alignment, and then designed a deformable registration network to generate pixel-wise deformation field for fine alignment. These methods have achieved more accurate registration through the highly nonlinear transformation process, but this transformation process severely lacks interpretability and is usually detrimental to maintaining inherent retinal anatomical differences. To this end, we reviewed relevant papers and found that Hernandez-Matas et al. [11] designed a novel 3D transformation model to register retinal images by analyzing the fundoscopy scene. In this model, an ellipsoid is used to simulate the eyeball, and the change in eyeball pose is converted to the change in camera pose. Afterward, based on the outstanding work of Hernandez-Matas et al. [11], we established a new 3D transformation model for the pair of retinal images with myopia development in our previous study [26]. This model further considers optical distortion and introduces a novel myopia development model to simulate changes in ocular axial length and fundus curvature during myopia development. This type of transformation model, which is close to the physical process in the actual scene, is fully interpretable and has great potential to align images as much as possible while maintaining the original anatomical differences.

In general, the characterization of the association between images determines the lower limit of registration performance, while the construction of the transformation model between images determines the upper limit. To date, many studies have focused on the former, but studies on the design of specialized transformation models are lacking, which greatly limits the further improvement of retinal image registration performance. Therefore, this paper focuses on the registration of longitudinal retinal images, which has the highest requirements on transformation model, and aims to design a more comprehensive and general transformation model.

B. Retinal Image Sequence Registration

The existing retinal image sequence registration methods usually further design a certain strategy on the basis of pair-wise registration to align all images in the sequence. Choe and Cohen [45] considered images as nodes, pair-wise registration as edges, and the registration error as the cost of each edge, forming a complete undirected graph. Then, the optimal path for registering the sequence was determined by finding the minimum spanning tree. Later, Perez-Rovira et al. [47] adopted a more direct strategy. This strategy automatically selects the frame with the most vessel segments from the sequence as the reference frame, and then uses the specifically designed pair-wise registration algorithm to align all remaining frames to the reference frame independently. However, the spatiotemporal change property of image sequence also requires good registration results between adjacent frames to ensure continuity and smoothness when switching frames. This strategy lacks alignment constraints between adjacent frames, which will result in a larger registration error in the common area between adjacent frames (especially the part of the common area that does not appear in the reference frame). To this end, Tsai et al. [48] used the correspondences between all available pairs of images in the sequence as alignment constraints to jointly optimize the pair-wise transformations of the remaining frames

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

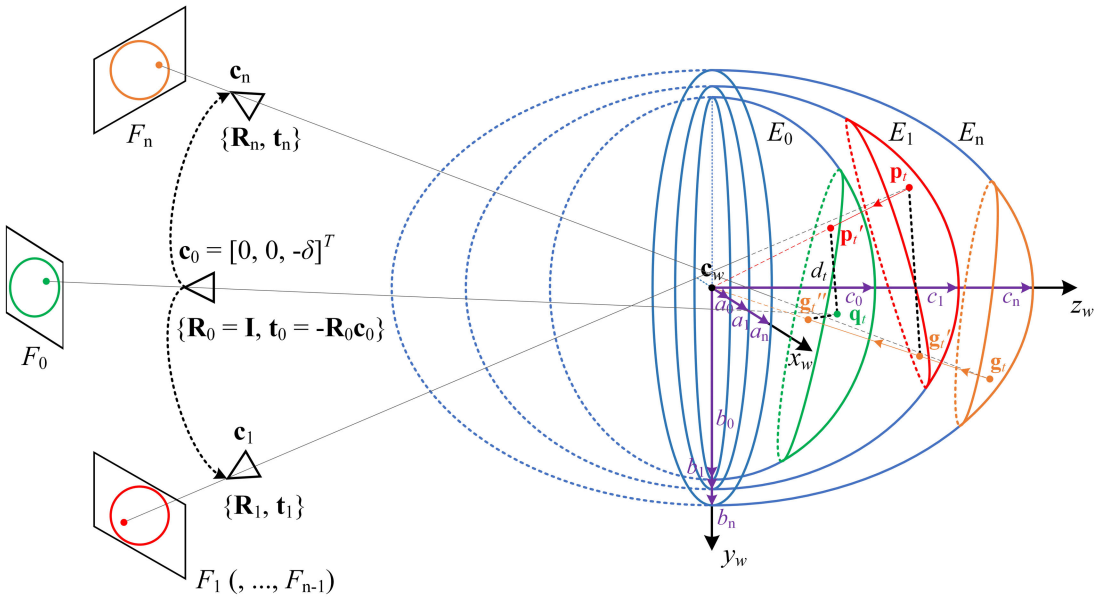

Fig. 2. 3D spatiotemporal model constructed in RISeR.

to the selected reference frame. This strategy considers a very comprehensive set of alignment constraints, but when the size of sequence is large or the complexity of transformation model is high, too many alignment constraints and transform parameters will make the optimization process difficult, making it hard to obtain a satisfactory registration result. Therefore, in this paper, we comprehensively consider the basic requirements of sequence registration, the feasibility of strategy, and the characteristics of our newly designed transformation model, and use frame-to-reference and frame-to-frame (the previous frame) as joint alignment constraints to register all retinal images in a longitudinal sequence into the common 3D space one by one.

III. Longitudinal 3D Fundoscopy Scene Modeling in RISeR

The proposed method, RISeR, constructs a 3D spatiotemporal model for the sequence of longitudinal retinal images $\{F_0, F_1, \ldots, F_n\}$, according to the process of fundus camera imaging retina and the changes in eyeball shape during natural growth or disease progression, as shown in Fig. 2. The 3D spatiotemporal model simulates a longitudinal fundoscopy scene, mainly including the following two important components: (1) Multiple ellipsoids $\{E_0, E_1, \ldots, E_n\}$ with different shapes but the same pose centered at the origin ($\mathbf{c}_w$) of the world coordinate system (x_w-y_w-z_w) form a dynamic eyeball model, which is used to simulate the shape changes that may occur in eye between each examination; (2) Multiple cameras $\{\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_n\}$ with different poses form a camera array model, in which the pose $\{\mathbf{R}_0, \mathbf{t}_0\}$ of the reference camera ($\mathbf{c}_0$) corresponding to the reference image (F_0) is fixed and known, and the changes in camera pose are used to simulate the inevitable changes in eye pose between each examination.

A. Dynamic Eyeball Model

The progression of some diseases usually leads to changes in eyeball shape. Taking myopia as an example, the results of relevant studies [49], [50] have shown that as myopia progresses, the ocular axial length will increase significantly, while the horizontal and vertical diameters of eye will also increase slightly. In addition, the eyeball shape will also change due to natural growth during childhood and adolescence, which is also manifested as an increase in the ocular axial length and the horizontal and vertical diameters of eye. Therefore, we approximately construct the dynamic eyeball model by simulating the length changes of the eye's three axes with natural growth or disease progression.

Specifically, as shown in Fig. 2, we set up n+ 1 eyeballs, $\{E_0, E_1, \ldots, E_n\}$, corresponding to the reference image F_0 and the remaining images $\{F_1, \ldots, F_n\}$. In terms of shape, all eyeballs are approximated using ellipsoids, and the lengths of the three orthogonal semi-axes $[a, b, c]$ between each ellipsoid are different and can be optimized respectively. In terms of pose, the pose parameters between all ellipsoids are completely consistent. The centers of these ellipsoids are placed at the origin of the world coordinate system, $\mathbf{c}_w = [0, 0, 0]^T$. And the postures of these ellipsoids are represented by the angles $[r_a, r_b, r_c]$ of three orthogonal semi-axes *a-b-c* rotating around the world coordinate axes x_w-y_w-z_w. Finally, we obtain the equations of the dynamic eyeball model as follows:

$$\begin{cases} \mathbf{x}_i^T \mathbf{Q}^T \mathbf{A}_i \mathbf{Q} \mathbf{x}_i = 1 \\ \mathbf{A}_i = \begin{bmatrix} a_i^{-2} & 0 & 0 \\ 0 & b_i^{-2} & 0 \\ 0 & 0 & c_i^{-2} \end{bmatrix} \\ \mathbf{Q} = \mathbf{R}_a(r_a) \cdot \mathbf{R}_b(r_b) \cdot \mathbf{R}_c(r_c), \end{cases} \tag{1}$$

where $\mathbf{x}_i$ is a 3D point on the surface of eyeball E_i, $i = 0, 1, \ldots, n$. As we can see, the parameters in the dynamic eyeball model consist of the following two parts: (1) the shapes of all ellipsoids $\{[a_0, b_0, c_0], [a_1, b_1, c_1], \ldots, [a_n, b_n, c_n]\}$; (2) the pose of each ellipsoid $[r_a, r_b, r_c]$.

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

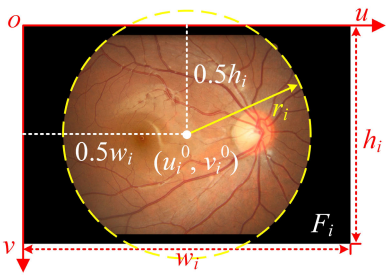

Fig. 3. According to the given retinal image, solve some parameters required to calculate the intrinsic parameter matrix of the corresponding camera.

Furthermore, we set up a new mapping scheme for mapping the corresponding points between different fundus curved surfaces onto the same fundus curved surface, in order to further perform the alignment constraint of the matched keypoints in 3D space. Specifically, we adopt the idea of limit and assume that the size of the ellipsoid decreases continuously as the three orthogonal semi-axes shorten, eventually becoming an infinitely small ellipsoid located at the center of the original ellipsoid. The line connecting the corresponding points between these two ellipsoids can then be approximated as the line connecting the point on the surface of the original ellipsoid and its center point. Since all ellipsoids in the dynamic eyeball model share the same center point, we take this center point as the starting point, and the intersection points of the ray emitted from here on several fundus curved surfaces are approximated as a set of corresponding points which can be mapped to each other. As shown in Fig. 2, the specific mapping process can be described as follows: take the center point $\mathbf{c}_w = [0, 0, 0]^T$ of the dynamic eyeball model as the starting point, connect a point $\mathbf{g}_t$ on the fundus curved surface of eyeball E_n to form a ray (Eq. (2)), and the intersection points $\{\mathbf{g}_t', \mathbf{g}_t''\}$ of the ray with the fundus curved surfaces of eyeballs $\{E_1, E_0\}$ (Eq. (1)) are the corresponding points after mapping.

$$\overrightarrow{\mathbf{c}_w\mathbf{g}_t} = \mu_0\overrightarrow{\mathbf{c}_w\mathbf{g}_t'} = \mu_1\overrightarrow{\mathbf{c}_w\mathbf{g}_t''}, \tag{2}$$

where μ_0 and μ_1 are proportional coefficients, and $\mu_0, \mu_1 > 0$.

B. Camera Array Model

The determination of the camera in a 3D scene usually involves two elements: (1) the intrinsic parameter matrix that is related only to the camera itself; (2) the extrinsic parameter matrix that reflects the pose of the camera. Considering that we convert the changes in eyeball pose into the changes in camera pose, and the camera used for each fundoscopy may also be different. Therefore, we approximately construct the camera array model in the longitudinal 3D fundoscopy scene by setting the intrinsic parameter matrix and extrinsic parameter matrix of the camera used in each fundoscopy.

Specifically, as shown in Fig. 2, we set up n+ 1 cameras, $\{\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_\mathrm{n}\}$, corresponding to the reference image F_0 and the remaining images $\{F_1, \ldots, F_\mathrm{n}\}$. In terms of the intrinsic parameter matrix, given a retinal image sequence, the intrinsic parameter matrix $\mathbf{K}_i$ of each corresponding camera can be directly calculated using Eq. (3) and Eq. (4).

$$\mathbf{K}_i = \begin{bmatrix} \alpha_i^x & 0 & u_i^0 \\ 0 & \alpha_i^y & v_i^0 \\ 0 & 0 & 1 \end{bmatrix}, \quad i = 0, 1, \ldots, \mathrm{n}, \tag{3}$$

where, α_i^x and α_i^y are the focal lengths of camera $\mathbf{c}_i$ measured in pixels, which are approximately calculated using the method (Eq. (4)) proposed by Hernandez-Matas et al. [11], (u_i^0, v_i^0) is the coordinate of the center point of image F_i in the pixel coordinate system, which can be calculated as half of the width and height of the image F_i (as shown in Fig. 3).

$$\alpha_i^x = \alpha_i^y = r_i \frac{l_i + \rho + \rho\cos\left(\frac{k_i}{2}\right)}{\rho\sin\left(\frac{k_i}{2}\right)}, \quad i = 0, 1, \ldots, \mathrm{n}, \tag{4}$$

where, r_i is the radius of the area that shows the retina in image F_i (as shown in Fig. 3), l_i and k_i are the lens-to-cornea distance and the field of view, respectively, which are known from the device specifications of the fundus camera $\mathbf{c}_i$ (if they are unknown, it is acceptable to use the device specifications of a commonly used fundus camera.), ρ is the radius of the roughly estimated spherical eye model [51], and $\rho = 12\ mm$.

In terms of the extrinsic parameter matrix, it consists of two parts: the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$, through which the pose of the camera in a scene can be determined. By using Eq. (5), we set the pose $\{\mathbf{R}_0, \mathbf{t}_0\}$ of the reference camera $\mathbf{c}_0$ corresponding to the reference image F_0 in the longitudinal 3D fundoscopy scene to be fixed and known. The poses of the remaining cameras $\{\mathbf{c}_1, \ldots, \mathbf{c}_\mathrm{n}\}$ in the array are adjusted and optimized relative to the reference camera $\mathbf{c}_0$, and the specific form is shown in Eq. (6).

$$\begin{cases} \mathbf{R}_0 = \mathbf{R}_x(0) \cdot \mathbf{R}_y(0) \cdot \mathbf{R}_z(0) = \mathbf{I} \\ \mathbf{t}_0 = -\mathbf{R}_0\mathbf{c}_0 \\ \mathbf{c}_0 = \begin{bmatrix} 0 & 0 & -\delta \end{bmatrix}^T \\ \delta = l_0 + \rho. \end{cases} \tag{5}$$

$$\begin{cases} \mathbf{R}_j = \mathbf{R}_x(r_j^\theta) \cdot \mathbf{R}_y(r_j^\phi) \cdot \mathbf{R}_z(r_j^\omega) \\ \mathbf{t}_j = \begin{bmatrix} t_j^x & t_j^y & t_j^z \end{bmatrix}^T \end{cases}, j = 1, 2, \ldots, \mathrm{n}, \tag{6}$$

where, $[r_j^\theta, r_j^\phi, r_j^\omega]$ are the rotation angles of camera $\mathbf{c}_j$ relative to the world coordinate system, $[t_j^x, t_j^y, t_j^z]$ is the position of the origin of world coordinate system in the camera coordinate system of camera $\mathbf{c}_j$.

In addition, there is usually some distortion in the process of imaging the retina with a fundus camera, which is actually a part of the inherent parameters in a camera used in practice. In our previous study, we have determined that the basic form of fundus camera distortion is fourth-order radial distortion [26]. Therefore, we add a fourth-order radial distortion model for each camera in the array, as shown in Eq. (7).

$$\begin{cases} x_d = x(1 + k_{i\_1}d^2 + k_{i\_2}d^4) \\ y_d = y(1 + k_{i\_1}d^2 + k_{i\_2}d^4) \\ d^2 = x^2 + y^2 \end{cases}, i = 0, 1, \ldots, \mathrm{n}, \tag{7}$$

where, (x, y) and (x_d, y_d) are the points on the normalized image plane before and after distortion is applied, respectively,

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

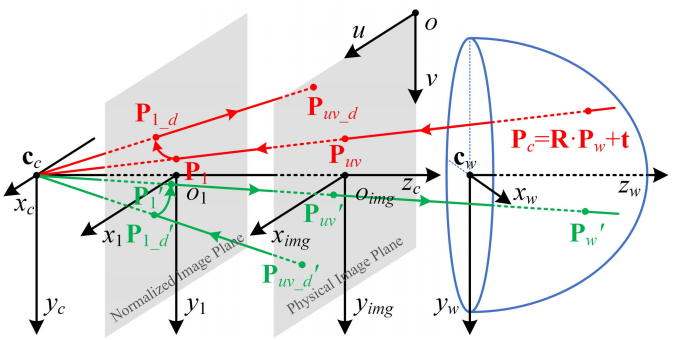

Fig. 4. The processes of mapping between 3D retinal points and 2D image points.

$[k_{i\_1}, k_{i\_2}]$ are the coefficients of fourth-order radial distortion of camera $\mathbf{c}_i$, which reflect the severity of the distortion.

Finally, given a retinal image sequence, the parameters of the camera array model consist of the following two parts: (1) the poses of all remaining cameras except for the reference camera, $\{[r_1^\theta, r_1^\phi, r_1^\omega], [r_2^\theta, r_2^\phi, r_2^\omega],..., [r_n^\theta, r_n^\phi, r_n^\omega]\}$, $\{[t_1^x, t_1^y, t_1^z], [t_2^x, t_2^y, t_2^z],..., [t_n^x, t_n^y, t_n^z]\}$; (2) the fourth-order radial distortion coefficients of all cameras, $\{[k_{0\_1}, k_{0\_2}], [k_{1\_1}, k_{1\_2}],..., [k_{n\_1}, k_{n\_2}]\}$.

C. Mapping Between 3D Retinal Points and 2D Image Points

The mapping between 3D retinal points and 2D image points serves as a bridge connecting the dynamic eyeball model and the camera array model, and it includes two opposite processes: (1) mapping from 3D retinal points to 2D image points; (2) mapping from 2D image points to 3D retinal points. The former is essentially the process of imaging the fundus retina with a camera, while the latter is equivalent to the process of the ray starting from the camera intersecting with the fundus. The specific processes are shown in Fig. 4.

The process of mapping from 3D retinal points to 2D image points (the part displayed in red in Fig. 4) is as follows: (1) Convert the coordinate of the 3D retinal point from the world coordinate system x_w-y_w-z_w to the camera coordinate system x_c-y_c-z_c using the extrinsic parameter matrix $\{\mathbf{R}, \mathbf{t}\}$ of fundus camera, that is, $\mathbf{P}_c = \mathbf{R} \cdot \mathbf{P}_w + \mathbf{t} = [x_c, y_c, z_c]^T$; (2) Transform $\mathbf{P}_c$ into the normalized image plane in the normalized coordinate system x_1-y_1 to obtain $\mathbf{P}_1$, that is, $\mathbf{P}_1 = [x_c, y_c, z_c]^T / z_c = [x_1, y_1, 1]^T$; (3) Calculate the coordinate $\mathbf{P}_{1\_d} = [x_{1\_d}, y_{1\_d}, 1]^T$ of $\mathbf{P}_1$ after adding the distortion using Eq. (7); (4) Transform $\mathbf{P}_{1\_d}$ into the pixel coordinate system u-v to obtain the 2D image point $\mathbf{P}_{uv\_d}$ using the intrinsic parameter matrix $\mathbf{K}$ of fundus camera, that is, $\mathbf{P}_{uv\_d} = \mathbf{K} \cdot \mathbf{P}_{1\_d}$.

The process of mapping from 2D image points to 3D retinal points (the part displayed in green in Fig. 4) is as follows: (1) Perform distortion correction on 2D image point $\mathbf{P}'_{uv\_d}$ to obtain the undistorted 2D image point $\mathbf{P}'_{uv}$; (2) Connect the 2D image point $\mathbf{P}'_{uv}$ and the camera center point $\mathbf{c}_c$ to form a ray (Eq. (8)) that points toward the fundus, and the intersection point of the ray and the posterior hemisphere of the ellipsoid (Eq. (1)) that represents the fundus is the corresponding 3D retinal point $\mathbf{P}'_w$.

$$\begin{cases} \mathbf{P}'_w = \mathbf{P}^+ \mathbf{P}'_{uv} + \lambda \mathbf{c}_c \\ \mathbf{P}^+ = \mathbf{P}^T (\mathbf{P}\mathbf{P}^T)^{-1} \\ \mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}], \end{cases} \tag{8}$$

where, λ is the coefficient calculated by combining Eq. (8) and Eq. (1), $\mathbf{P}$ is the projection matrix of camera.

IV. RISeR Framework

Taking the corresponding points between images as a medium, the parameters of the constructed 3D spatiotemporal model are optimized under the frame-to-reference and frame-to-frame joint alignment constraints, and finally, a longitudinal 3D fundoscopy scene for registering the given retinal image sequence is obtained. The workflow of the RISeR framework is summarized in Algorithm 1. In the following subsections, we describe the key steps in this workflow in detail.

A. Keypoint Extraction and Matching

In this work, we directly adopt a commonly used method to extract and match the keypoints between retinal images. Specifically, the SuperPoint algorithm, which was proposed by DeTone et al. [30], is used to detect and describe the keypoints in retinal images, and then the SuperGlue algorithm, which was proposed by Sarlin et al. [34], is used to match the keypoint sets extracted from different retinal images. We do not perform any additional network structure design or model weight fine-tuning on SuperPoint or SuperGlue but select and set only some of their hyperparameters. Specifically, we set the confidence threshold of keypoint detection in SuperPoint to 0.001, select the weight file loaded in SuperGlue as the "outdoor" type, and adopt the default values for the remaining hyperparameters.

B. Initialization of the 3D Spatiotemporal Model

Given a retinal image sequence $\{F_0, F_1, \ldots, F_n\}$, the unknown parameters in the corresponding 3D spatiotemporal model consist of the following four parts: (1) the shapes of all ellipsoids in the dynamic eyeball model, $\{[a_i, b_i, c_i]\}_{i=0}^n$; (2) the posture of the dynamic eyeball model, $[r_a, r_b, r_c]$; (3) the poses of all remaining cameras in the camera array model except the reference camera, $\{\mathbf{R}_j, \mathbf{t}_j\}_{j=1}^n$, i.e., $\{[r_j^\theta, r_j^\phi, r_j^\omega]\}_{j=1}^n$ and $\{[t_j^x, t_j^y, t_j^z]\}_{j=1}^n$; (4) the fourth-order radial distortion coefficients of all cameras in the camera array model, $\{[k_{i\_1}, k_{i\_2}]\}_{i=0}^n$.

The above unknown parameters are initialized as follows: (1) The initial shape of each eyeball is set as a sphere with a radius of 12mm according to the Navarro eye model [51]; (2) All the eyeballs have the same posture, and the initial value is set to have no rotation relative to the world coordinate system; (3) The poses of all the remaining cameras are initialized by solving the Perspective-n-Point (PnP) problem under the RANSAC framework [33]. The PnP problem refers to the estimation of the camera pose $\{\mathbf{R}, \mathbf{t}\}$ through a set of

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 RISeR Framework

Input: A retinal image sequence $\{F_0, F_1, \ldots, F_{\mathrm{n}}\}$, F_0 is the reference image selected from the sequence.

Output: The registered retinal image sequence $\{F_0, F_{1\to0}, \ldots, F_{\mathrm{n}\to0}\}$, and the parameters of the corresponding longitudinal 3D fundoscopy scene, $\{[a_i, b_i, c_i]\}_{i=0}^{\mathrm{n}}$, $[r_a, r_b, r_c]$, $\{\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i\}_{i=0}^{\mathrm{n}}$, $\{[k_{i\_1}, k_{i\_2}]\}_{i=0}^{\mathrm{n}}$.

Compute $\mathbf{K}_0$ according to Eq. (3) and Eq. (4);
Compute $\{\mathbf{R}_0, \mathbf{t}_0\}$ according to Eq. (5);
$kp_0 = SuperPoint(F_0)$; *// Perform keypoint detection and description on* F_0
/ Perform pair-wise registration on* F_0 *and* F_1 **/*
$kp_1 = SuperPoint(F_1)$;
$mkp_{(0,1)} = SuperGlue(kp_0, kp_1)$; *// Perform keypoint matching between* F_0 *and* F_1
Compute $\mathbf{K}_1$ according to Eq. (3) and Eq. (4);
Initialization $\mathbf{\Phi}_0$**:** $[a_0, b_0, c_0] = 12.0mm$, $[k_{0\_1}, k_{0\_2}] = [-0.5623, 0.3317]$, $[r_a, r_b, r_c] = 0rad$;
Initialization $\mathbf{\Phi}_1$**:** $[a_1, b_1, c_1] = 12.0mm$, $[k_{1\_1}, k_{1\_2}] = [-0.5623, 0.3317]$, $\{\mathbf{R}_1, \mathbf{t}_1\}$ = the solution of PnP problem;
Optimization $\mathbf{\Phi}_0$ **and** $\mathbf{\Phi}_1$**:** $PSO(mkp_{(0,1)}, \text{Eq. (10)}, \mathbf{K}_0, \mathbf{R}_0, \mathbf{t}_0, \mathbf{K}_1)$;
$F_{1\to0} = Transform(F_1, \mathbf{K}_0, \mathbf{R}_0, \mathbf{t}_0, \mathbf{K}_1, \mathbf{\Phi}_0, \mathbf{\Phi}_1)$;
/ The remaining images in sequence are registered using the frame-to-reference and frame-to-frame strategy */*
if $\mathrm{n} \geq 2$ **then**
  for $(i = 2; i \leq \mathrm{n}; i++)$ **do**
    Compute $\mathbf{K}_i$ according to Eq. (3) and Eq. (4);
    $kp_i = SuperPoint(F_i)$;
    $mkp_{(0,i)} = SuperGlue(kp_0, kp_i)$;
    $mkp_{(i-1,i)} = SuperGlue(kp_{i-1}, kp_i)$;
    Initialization $\mathbf{\Phi}_i$**:** $[a_i, b_i, c_i] = 12.0mm$, $[k_{i\_1}, k_{i\_2}] = [-0.5623, 0.3317]$, $\{\mathbf{R}_i, \mathbf{t}_i\}$ = the solution of PnP problem;
    Optimization $\mathbf{\Phi}_i$**:** $PSO(mkp_{(0,i)}, mkp_{(i-1,i)}, \text{Eq. (13)}, \mathbf{K}_0, \mathbf{R}_0, \mathbf{t}_0, \mathbf{K}_{i-1}, \mathbf{K}_i, \mathbf{\Phi}_0, \mathbf{\Phi}_{i-1})$;
    $F_{i\to0} = Transform(F_i, \mathbf{K}_0, \mathbf{R}_0, \mathbf{t}_0, \mathbf{K}_i, \mathbf{\Phi}_0, \mathbf{\Phi}_i)$;
  end for
end if
return $\{F_0, F_{1\to0}, \ldots, F_{\mathrm{n}\to0}\}$, the longitudinal 3D fundoscopy scene.

2D-3D corresponding points and the intrinsic parameter matrix $\mathbf{K}$ of the camera. Here, the 2D points are the matched keypoints in the retinal image associated with the target camera. Their corresponding 3D points are obtained by mapping the matched keypoints in the reference image onto the fundus surface of the initial eyeball; (4) For the fourth-order radial distortion coefficients of all the cameras, the calibration result in our previous work [26] is used as the initial value.

C. Optimization of the 3D Spatiotemporal Model

As described in Algorithm 1, there are two different kinds of optimization involved in the RISeR framework: (1) the optimization under the single constraint of frame-to-reference; (2) the optimization under the joint constraints of frame-to-reference and frame-to-frame. The former is used to register the first image F_1 in the sequence after the reference image F_0, while the latter is used to further register subsequent images $\{F_2, F_3, \ldots, F_{\mathrm{n}}\}$ in the sequence. In both cases, the Particle Swarm Optimization (PSO) algorithm [52] is used to find the optimal candidate solution in a given search space through p particles iterating g times.

For the first kind of optimization, as shown in Fig. 2, the way of calculating its objective function is described as follows: Firstly, map the matched $m_{(0,1)}$ pairs of keypoints in F_0 and F_1 onto their respective fundus curved surfaces, and obtain $m_{(0,1)}$ sets of corresponding points $\{\mathbf{q}_t, \mathbf{p}_t\}_{t=1}^{m_{(0,1)}}$; Secondly, map $\mathbf{p}_t$ onto the fundus curved surface where $\mathbf{q}_t$ is located to obtain $\mathbf{p}'_t$; Thirdly, calculate the Euclidean distance $d_t^{(0,1)}$ of $\{\mathbf{q}_t, \mathbf{p}'_t\}$ according to Eq. (9); Finally, sort $\{d_t^{(0,1)}\}_{t=1}^{m_{(0,1)}}$ in ascending order to obtain $\{d_t^{(0,1)\prime}\}_{t=1}^{m_{(0,1)}}$, and minimize the sum of first 80% distances (to reduce the impact of mismatched keypoints) as the objective function (Eq. (10)).

$$d_t^{(0,1)} = |\mathbf{q}_t - \mathbf{p}'_t|, t = 1, 2, \ldots, m_{(0,1)}. \tag{9}$$

$$D(\mathbf{\Phi}_0, \mathbf{\Phi}_1) = \sum_{t=1}^{\lfloor 0.8 m_{(0,1)} \rfloor} d_t^{(0,1)\prime}. \tag{10}$$

For the second kind of optimization, as shown in Fig. 2, the way of calculating its objective function is described as follows: (1) map the matched $m_{(0,\mathrm{n})}$ pairs of keypoints in F_0 and F_{n} onto their respective fundus curved surfaces, and obtain $m_{(0,\mathrm{n})}$ sets of corresponding points $\{\mathbf{q}_t, \mathbf{g}_t\}_{t=1}^{m_{(0,\mathrm{n})}}$; (2) map $\mathbf{g}_t$ onto the fundus curved surface where $\mathbf{q}_t$ is located to obtain $\mathbf{g}''_t$; (3) calculate the Euclidean distance $d_t^{(0,\mathrm{n})}$ of $\{\mathbf{q}_t, \mathbf{g}''_t\}$ in 3D space according to Eq. (11); (4) map the matched $m_{(\mathrm{n}-1,\mathrm{n})}$ pairs of keypoints in $F_{\mathrm{n}-1}$ and F_{n} onto their respective fundus curved surfaces, and obtain $m_{(\mathrm{n}-1,\mathrm{n})}$ sets of corresponding points $\{\mathbf{p}_t, \mathbf{g}_t\}_{t=1}^{m_{(\mathrm{n}-1,\mathrm{n})}}$; (5) map $\mathbf{g}_t$ onto the fundus curved surface where $\mathbf{p}_t$ is located to obtain $\mathbf{g}'_t$; (6) calculate the Euclidean distance $d_t^{(\mathrm{n}-1,\mathrm{n})}$ of $\{\mathbf{p}_t, \mathbf{g}'_t\}$ in 3D space according to Eq. (12); (7) sort $\{d_t^{(0,\mathrm{n})}\}_{t=1}^{m_{(0,\mathrm{n})}}$ and $\{d_t^{(\mathrm{n}-1,\mathrm{n})}\}_{t=1}^{m_{(\mathrm{n}-1,\mathrm{n})}}$ in ascending order respectively to obtain $\{d_t^{(0,\mathrm{n})\prime}\}_{t=1}^{m_{(0,\mathrm{n})}}$ and $\{d_t^{(\mathrm{n}-1,\mathrm{n})\prime}\}_{t=1}^{m_{(\mathrm{n}-1,\mathrm{n})}}$; (8) calculate the sum of the first 80% values in two sets separately, and then add the two sums with a certain weight (to balance the two parts)

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

as the objective function (Eq. (13)) to be minimized.

$$d_t^{(0,n)} = |\mathbf{q}_t - \mathbf{g}_t''|, t = 1, 2, \ldots, m_{(0,n)}. \tag{11}$$

$$d_t^{(n-1,n)} = |\mathbf{p}_t - \mathbf{g}_t'|, t = 1, 2, \ldots, m_{(n-1,n)}. \tag{12}$$

$$\begin{cases} D(\mathbf{\Phi}_n) = \alpha \sum_{t=1}^{\lfloor 0.8m_{(0,n)} \rfloor} d_t^{(0,n)\prime} + \beta \sum_{t=1}^{\lfloor 0.8m_{(n-1,n)} \rfloor} d_t^{(n-1,n)\prime} \\ \alpha = \dfrac{\max(\lfloor 0.8m_{(0,n)} \rfloor, \lfloor 0.8m_{(n-1,n)} \rfloor)}{\lfloor 0.8m_{(0,n)} \rfloor} \\ \beta = \dfrac{\max(\lfloor 0.8m_{(0,n)} \rfloor, \lfloor 0.8m_{(n-1,n)} \rfloor)}{\lfloor 0.8m_{(n-1,n)} \rfloor}. \end{cases} \tag{13}$$

Each time an optimization is completed, some of the unknown parameters in the 3D spatiotemporal model are determined, and they are used as fixed known values in the subsequent optimization. Until all n optimizations are completed, all unknown parameters in the 3D spatiotemporal model are determined, forming a longitudinal fundoscopy scene that simulates the imaging process of a given retinal image sequence. In this scene, the retinal images to be registered in the sequence are mapped to the perspective of the reference image to obtain several sets of 2D floating point coordinates and corresponding RGB values, and then the registered retinal image sequence is generated through bilinear interpolation.

D. Generation of Synthetic Retinal Image Sequences

Benefiting from the 3D spatiotemporal simulation of the longitudinal fundoscopy scene, the proposed RISeR framework can generate many realistic synthetic retinal image sequences from a single input image through a series of transformations with actual physical significance. This approach has significant potential application value in registration method evaluation, deep learning-based model training, data augmentation, etc. Specifically, the generation of a synthetic retinal image sequence $\{F_1^s, F_2^s, \ldots, F_n^s\}$ mainly consists of the following steps: (1)Take the given retinal image as a reference image F_0, and set the relevant parameters $\{\mathbf{K}_0, \mathbf{R}_0, \mathbf{t}_0, [k_{0\_1}, k_{0\_2}], [a_0, b_0, c_0], [r_a, r_b, r_c]\}$ of the reference camera and reference eyeball in the 3D spatiotemporal model according to this image; (2) Set the parameters $\{\mathbf{K}_j, \mathbf{R}_j, \mathbf{t}_j, [k_{j\_1}, k_{j\_2}], [a_j, b_j, c_j]\}_{j=1}^n$ of the remaining cameras in the camera array model and the remaining eyeballs in the dynamic eyeball model using a certain strategy, and a complete 3D spatiotemporal model is finally constructed to simulate a longitudinal fundoscopy scene; (3) In this scene, a synthetic retinal image sequence $\{F_1^s, F_2^s, \ldots, F_n^s\}$ is generated by mapping image points between the reference camera and the remaining cameras and performing bilinear interpolation and optional color space transformations (to simulate the changes in image color, brightness, contrast, etc., caused by changes in the camera, camera settings, etc.).

As shown in Table I, according to the strategy used to set the parameters of the remaining cameras and eyeballs in the 3D spatiotemporal model and whether color space transformations are performed when generating images, five different types of longitudinal fundoscopy scenes can be constructed: (1) *P* simulates a series of retinal images captured as the eye moves during the same fundus photography; (2) *PC* simulates a series of retinal images captured from an eye that does not present any shape changes using the same fundus camera during different examinations; (3) *PCI* simulates a series of retinal images captured from an eye that does not present any shape changes using different fundus cameras during different examinations; (4) *PCS* simulates a series of retinal images captured from an eye that presents shape changes using the same fundus camera during different examinations; (5) *PCSI* simulates a series of retinal images captured from an eye that presents shape changes using different fundus cameras during different examinations. For example, taking a retinal image as input, the corresponding five different types of synthetic retinal image sequences generated through the above scenes are shown in Fig. 5.

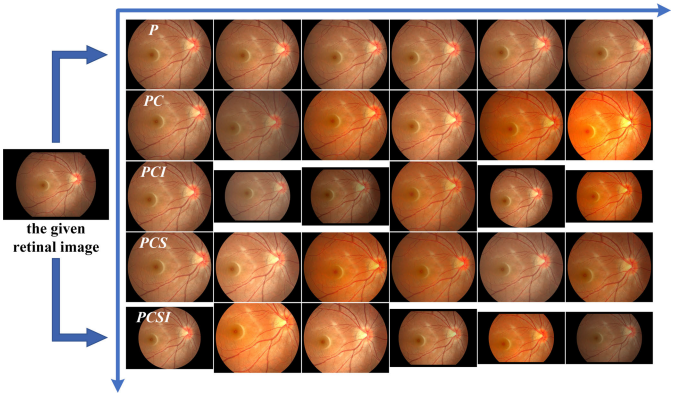

Fig. 5. Taking a retinal image as input, five different types of synthetic retinal image sequences are generated.

TABLE I
CONSTRUCTION OF FIVE DIFFERENT TYPES OF LONGITUDINAL FUNDOSCOPY SCENES

Types	Camera pose (P)	Camera intrinsic params. (I)	Eyeball shape (S)	Color space transform (C)
	$\mathbf{R}_j, \mathbf{t}_j$	$\mathbf{K}_j, [k_{j\_1}, k_{j\_2}]$	$[a_j, b_j, c_j]$	
P	✓			
PC	✓			✓
PCI	✓	✓		✓
PCS	✓		✓	✓
PCSI	✓	✓	✓	✓

Note: In the process of setting up the remaining cameras and remaining eyeballs one by one ($j = 1, 2, \ldots, n$), the items selected above are changed, and the remaining items are fixed.

V. EXPERIMENTS

A. Datasets

Five datasets are used in our experiments. The first two are publicly available retinal image pair registration datasets, including the Fundus Image Registration (FIRE) dataset [53] shared by Hernandez-Matas et al. and the Fundus Image Myopia Development (FIMD) dataset [26] shared by us in our previous work. The next two are our newly created Retinal Image Sequence Registration datasets, including a real dataset (RISeR-real) and a synthetic dataset (RISeR-syn). These two datasets are also made publicly available along with this paper, with the goal of addressing the lack of such shared data. The

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

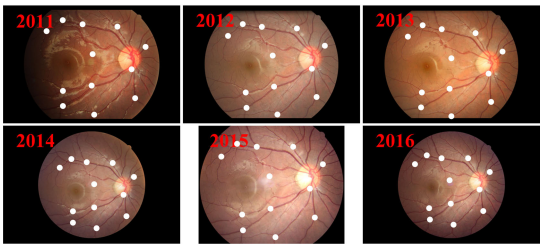

Fig. 6. A sequence of retinal images with annotated control points in RISeR-real.

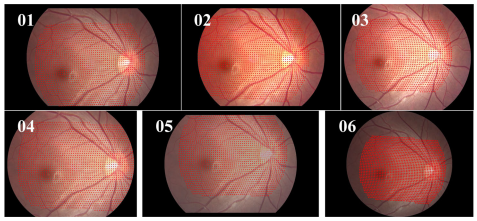

Fig. 7. A sequence of synthetic retinal images with annotated control points in RISeR-syn.

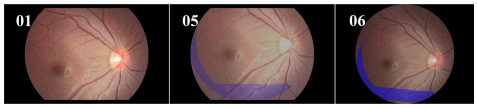

Fig. 8. The annotated control points that exist only in the common retinal area between two adjacent frames but not in the reference frame.

last dataset is a publicly available retinal image registration dataset with progression of retinopathy of prematurity (ROP), called COph100 [54], shared by Hu et al.

FIRE contains 134 retinal image pairs, which are divided into three categories: (1) 71 pairs of images with large overlapping areas and without any anatomical changes are classified as category $\mathcal{S}$; (2) 49 pairs of images with small overlapping areas and without any anatomical changes are classified as category $\mathcal{P}$; (3) 14 pairs of images with large overlapping areas and large anatomical changes are classified as category $\mathcal{A}$, these anatomical changes are caused by the progression of retinopathy, such as diabetic retinopathy [55]. Each pair of retinal images is manually annotated with 10 pairs of widely and uniformly distributed ground-truth corresponding points (also known as control points) as a medium for quantitatively evaluating the registration accuracy. However, the annotation of the 6-th pair of control points in the 37-th retinal image pair in category $\mathcal{P}$ is incorrect, so we uniformly exclude this pair of control points from the evaluation in subsequent experiments.

FIMD consists of 70 retinal image pairs, each of which has obvious anatomical changes due to myopia progression. Similar to FIRE, each pair of images is manually annotated with 12 pairs of control points. In addition, the time interval between capturing a pair of images is longer, and the fundus camera used each time is different, resulting in different resolution, brightness, contrast and FOV between two images.

RISeR-real contains 10 longitudinal retinal image sequences, which are from 10 different eyes. Each sequence consists of 5 or 6 retinal images captured using different fundus cameras at intervals of 1 year or 2 years. Each retinal image sequence has no presence of any other fundus diseases except obvious myopia progression. As shown in Fig. 6, we manually annotate 12 sets of control points for each retinal image sequence. In addition, RISeR-real also records the UCVA data corresponding to each retinal image in all sequences, which makes this dataset potentially valuable for the longitudinal study of the fundus retina associated with myopia progression. The study on the RISeR-real dataset was approved by the Medical Ethics Committees of Tianjin Eye Hospital, Tianjin, China, with the study number KY-2023067, and fully complied with the Declaration of Helsinki.

RISeR-syn is generated using the method described in Section IV-D. We select 10 retinal images from different eyes and take one of them as a given reference image each time to generate 5 different types of synthetic retinal image sequences, resulting in a total of 50 sequences, each consisting of 6 synthetic retinal images. Compared with RISeR-real, which is time-consuming and prone to introducing errors when manually annotating control points, RISeR-syn can easily obtain densely distributed and accurately corresponding control points when it is generated, thereby achieving a more objective and effective registration evaluation. Fig. 7 shows all the control points annotated in a sequence, which cover the common retinal area of all images. Moreover, under the premise of taking the first image in sequence as the reference frame, we also annotate control points that exist only in the common retinal area between two adjacent frames but not in the reference frame (as shown in Fig. 8) to more comprehensively evaluate the inter-frame continuity of the sequence registration results. Due to the lack of longitudinal studies on pattern learning and simulated generation of fundus anatomical deformation during disease progression in existing studies, no anatomical changes related to disease progression are added to RISeR-syn.

COph100 is a retinal image registration dataset that focuses specifically on the progression of ROP. This dataset consists of 100 eyes, each with 2 to 9 retinal images collected at different examination times. These images exhibit significant anatomical changes due to the progression of ROP. Similarly, all retinal images of each eye are manually annotated with 10 sets of control points, which are located mainly around the vessel intersections. We select 24 eyes that underwent 4 or more follow-up examinations from COph100 for the experiment. We name this part of the data **COph100-45689** to distinguish it from the complete dataset that contains all eyes.

B. Experimental Setup

The proposed RISeR framework is implemented via PyTorch and OpenCV. All the registration experiments related to RISeR are conducted on an Intel(R) Core(TM) i5-12500 CPU @ 3.00 GHz with 16.0 GB RAM memory.

For the PSO algorithm involved in the proposed RISeR framework, we use the official implementation *pyswarms.single.GlobalBestPSO()* [52] and set its hyperparameters in the experiments as follows: (1) number of particles $p = 10000$, iterations $g = 300$; (2) cognitive parameter $c_1 = 1.6$, social parameter $c_2 = 1.4$, inertia parameter $w = 0.6$, their update strategies are "nonlin_mod", "lin_variation", and "exp_decay", respectively; (3) the search spaces of eyeball

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

TABLE II
AVERAGE REGISTRATION ERROR e_{avg} AND AUC ON FIRE AND FIMD FOR THE PROPOSED METHOD AND THE COMPARISON METHODS

Methods	FIRE-$\mathcal{S}$		FIRE-$\mathcal{P}$		FIRE-$\mathcal{A}$		FIRE		FIMD	
	$e_{avg}\downarrow$	AUC $\uparrow$	$e_{avg}\downarrow$	AUC $\uparrow$	$e_{avg}\downarrow$	AUC $\uparrow$	$e_{avg}\downarrow$	AUC $\uparrow$	$e_{avg}\downarrow$	AUC $\uparrow$
GDB-ICP	8.561	0.690	–	0.338	–	0.200	–	0.510	–	0.215
SURF-PIIFD-RPM	3.216	0.888	–	0.583	–	0.463	–	0.732	–	0.385
GFEMR	4.718	0.833	–	0.558	19.070	0.400	–	0.687	–	0.361
REMPE	1.620	**0.958**	12.349	0.522	12.936	0.640	6.725	0.765	–	0.529
SuperRetina	1.761	0.950	11.612	0.558	5.944	0.786	5.800	0.790	7.773	0.711
GeoFormer	2.394	0.926	11.042	0.580	6.765	0.746	6.013	0.781	7.777	0.711
RIRMD	1.736	0.950	8.533	0.682	5.740	**0.794**	4.640	0.836	7.821	0.710
RISeR	**1.584**	0.957	**4.072**	**0.859**	**5.671**	**0.794**	**2.921**	**0.904**	**7.534**	**0.719**

Bold font represents the best result, and underline represents the second-best result.
"–" means that e_{avg} cannot be calculated or the value is too large because some image pairs fail to be registered.

shape parameters, eyeball posture parameters, camera rotation parameters, camera translation parameters and camera distortion parameters are 4.0 *mm*, 2.0 *rad*, 1.0 *rad*, 2.0 *mm*, and 1.0, respectively, their maximum search steps are 0.1 *mm*, 0.01 *rad*, 0.01 *rad*, 0.1 *mm*, and 0.01, respectively; (4) the position and velocity adjustment strategies of out-of-bounds particles are "nearest" and "zero", respectively. We do not set any additional optimization termination criteria, and the optimization stops only when the preset maximum number of iterations is reached.

C. Experimental Results on FIRE and FIMD

1) Evaluation Metrics: As shown in Eq. (14), the registration error e of each retinal image pair $\{F_0, F_1\}$ is obtained by calculating the average Euclidean distance $\| \ \|_2$ of the annotated control points $\{\mathbf{p}_0^i, \mathbf{p}_1^i\}_{i=1}^{n_c}$ after applying the registration transformation $T(\ )$, where n_c represents the number of annotated control point pairs. On this basis, we further use the following two ways to statistically analyze the registration errors $\{e_j\}_{j=1}^N$ of all retinal image pairs $\{F_0^j, F_1^j\}_{j=1}^N$ in the dataset, and take these as metrics to evaluate the registration results of the algorithm on this dataset.

$$e = \frac{1}{n_c}\sum_{i=1}^{n_c}\left|\left|\mathbf{p}_0^i - T\left(\mathbf{p}_1^i\right)\right|\right|_2. \tag{14}$$

Firstly, we calculate the average value of $\{e_j\}_{j=1}^N$ to obtain the **Average Registration Error** e_{avg}, which represents the overall registration accuracy of a method on a dataset.

$$e_{avg} = \frac{1}{N}\sum_{j=1}^{N} e_j, \tag{15}$$

where N is the number of retinal image pairs in a dataset.

Secondly, we count the registration success rates of $\{e_j\}_{j=1}^N$ under different error thresholds e_t (1-25 pixels), and then calculate the **Area Under Curve (AUC)** to evaluate the overall registration success rate of a method on a dataset.

$$\text{AUC} = \frac{1}{25}\sum_{e_t=1}^{25}\left[\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}\left(e_j < e_t\right)\right], \tag{16}$$

where $\mathbb{1}(\ )$ is the indicator function.

2) Registration Methods: We select four classic methods (GDB-ICP [36], SURF-PIIFD-RPM [37], GFEMR [27], REMPE [11]) and three newly proposed advanced methods (SuperRetina [17], GeoFormer [18], RIRMD [26]) for comparison. We use the publicly available implementations and keep the default parameter settings to perform registration experiments. As to SuperRetina and GeoFormer, we use the model weights shared by the authors and run them on the NVIDIA GeForce RTX 4090 GPU. The remaining five comparison methods are run on the same device as RISeR.

3) Results and Discussion: The experimental results are shown in Table II. The proposed method achieves the best registration results on both datasets (FIRE: e_{avg} = **2.921** pixels, AUC = **0.904**; FIMD: e_{avg} = **7.534** pixels, AUC = **0.719**). In addition, compared with three advanced methods (SuperRetina, GeoFormer, RIRMD) proposed in recent years and one classic method (REMPE) that still performs well, only the proposed method (RISeR) all achieves the best or second-best results in four significantly different registration tasks (FIRE-$\mathcal{S}$, FIRE-$\mathcal{P}$, FIRE-$\mathcal{A}$, FIMD), which further indicates that the proposed method also has good stability and robustness.

In fact, the core work of both SuperRetina and GeoFormer is designing a more effective keypoint extraction and matching algorithm, and the experiments in [17] and [18] have shown that the keypoint extraction and matching results of both of these two algorithms on retinal image pairs are significantly better than the combination of SuperPoint and SuperGlue used in the proposed method, RISeR. However, in this case, the registration evaluation results of RISeR on FIRE and FIMD datasets can still comprehensively surpass SuperRetina and GeoFormer with significant advantages. This illustrates the effectiveness of the transformation model based on longitudinal fundoscopy scene modeling proposed in this paper for the retinal image registration task. Moreover, the proposed method, RISeR, and the most competitive comparison method, RIRMD, adopt the same keypoint extraction and matching algorithm and optimization algorithm, and basically only the transformation model is different in pairwise registration. As shown in the last two rows of Table II, the evaluation results of RISeR on four registration tasks are all better than those of RIRMD. Especially on FIRE-$\mathcal{P}$,

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

which has higher requirements for the complexity and degree of freedom of the transformation model (caused by the unique curved surface shape of fundus combined with a large degree of eye rotation), RISeR achieves significantly greater improvements (e_{avg}: 8.533 → **4.072** pixels; AUC: 0.682 → **0.859**). This more directly demonstrates the effectiveness and progressiveness of the transformation model designed by the proposed method. Finally, it is worth noting that the image pairs in FIMD and FIRE-$\mathcal{A}$ contain anatomical changes caused by myopia progression and retinopathy progression, respectively, and the proposed method, RISeR, achieves the best registration results on both datasets. This preliminarily demonstrates that RISeR has good registration applicability for retinal images with anatomical changes caused by myopia progression or non-myopic fundus disease progression.

D. Experimental Results on RISeR-Real

1) Evaluation Metrics: For a retinal image sequence $\{F_0, F_1, \ldots, F_{\mathrm{n}}\}$, all remaining images $\{F_1, \ldots, F_{\mathrm{n}}\}$ are registered to the perspective of the selected reference image F_0 by estimating a set of transformations $\{T_{1\to0}(), \ldots, T_{\mathrm{n}\to0}()\}$. In order to more comprehensively evaluate the registration result of $\{F_0, F_1, \ldots, F_{\mathrm{n}}\}$, we calculate the following two metrics with the help of the n_c sets of annotated control points $\{\mathbf{p}_0^i, \mathbf{p}_1^i, \ldots, \mathbf{p}_{\mathrm{n}}^i\}_{i=1}^{n_c}$. First of all, as shown in Eq. (17), we calculate the **Average Registration Error** $\boldsymbol{E_{seq}}$ between the transformed control points $\{T_{1\to0}(\mathbf{p}_1^i), \ldots, T_{\mathrm{n}\to0}(\mathbf{p}_{\mathrm{n}}^i)\}_{i=1}^{n_c}$ of $\{F_1, \ldots, F_{\mathrm{n}}\}$ and the corresponding control points $\{\mathbf{p}_0^i\}_{i=1}^{n_c}$ of F_0, which reflects the overall registration accuracy of an algorithm on $\{F_0, F_1, \ldots, F_{\mathrm{n}}\}$.

$$E_{seq} = \frac{1}{\mathrm{n}} \sum_{j=1}^{\mathrm{n}} \left(\frac{1}{n_c} \sum_{i=1}^{n_c} \left|\left| \mathbf{p}_0^i - T_{j\to0}\left(\mathbf{p}_j^i\right) \right|\right|_2 \right). \tag{17}$$

Secondly, as shown in Eq. (18), we calculate the **Average Drift Distance** $\boldsymbol{D_{seq}}$ of the transformed control points $\{T_{1\to0}(\mathbf{p}_1^i), \ldots, T_{\mathrm{n}\to0}(\mathbf{p}_{\mathrm{n}}^i)\}_{i=1}^{n_c}$ between all adjacent frames in $\{F_1, \ldots, F_{\mathrm{n}}\}$, which reflects the inter-frame continuity of the registration result $\{F_0, F_{1\to0}, \ldots, F_{\mathrm{n}\to0}\}$. The smaller the value of this metric, the better the consistency of the registration result, the smoother and more natural the switching between adjacent frames, that is, the better the continuity.

$$D_{seq} = \frac{1}{\mathrm{n}-1} \sum_{j=1}^{\mathrm{n}-1} \left(\frac{1}{n_c} \sum_{i=1}^{n_c} \left|\left| T_{j\to0}(\mathbf{p}_j^i) - T_{j+1\to0}(\mathbf{p}_{j+1}^i) \right|\right|_2 \right). \tag{18}$$

2) Registration Methods: Among the existing methods for global registration, there are few studies on image sequence registration. The common practice is to directly apply the pair-wise registration method multiple times to independently register the remaining images to the selected reference image. The few global registration methods that are specifically designed for image sequences are also difficult to use for comparison because their implementations are not publicly available. Therefore, we choose the most competitive comparison method SuperRetina on FIMD dataset and the highly relevant methods (REMPE, RIRMD), adopt the common practice just mentioned above to perform registration on RISeR-real dataset. In addition, to explore the effect of the frame-to-reference and frame-to-frame joint registration strategy adopted by the proposed method RISeR, we also conduct experiment using the ablation version of RISeR (RISeR-F2R) which adopt the single registration strategy of frame-to-reference.

TABLE III

AVERAGE REGISTRATION ERROR E_{seq} AND AVERAGE DRIFT DISTANCE D_{seq} ON RISER-REAL FOR THE PROPOSED METHOD, ITS ABLATION VERSION, AND THE COMPARISON METHODS

Methods	RISeR-real	
	E_{seq} (pixels) ↓	D_{seq} (pixels) ↓
REMPE	6.165	6.858
SuperRetina	5.368	6.374
RIRMD	5.086	6.022
RISeR-F2R	5.063	5.632
RISeR	**4.588**	**4.485**

Bold font represents the best result.

3) Results and Discussion: The experimental results are shown in Table III. RISeR achieves the highest registration accuracy and the best inter-frame continuity (E_{seq} = **4.588** pixels, D_{seq} = **4.485** pixels), both of which are significantly better than those of SuperRetina (E_{seq} = 5.368 pixels, D_{seq} = 6.374 pixels), especially the improvement of D_{seq} reaching nearly 30%. Compared with the highly relevant method REMPE (E_{seq} = 6.165 pixels, D_{seq} = 6.858 pixels), RISeR also achieves significant improvements in two error metrics, reaching 25.6% and 34.6%, respectively. Even compared with the most competitive comparison method RIRMD (E_{seq} = 5.086 pixels, D_{seq} = 6.022 pixels), RISeR still has significant advantages, especially in improving the inter-frame continuity metric by 25.5%. These findings demonstrate the superior performance of the proposed method in the task of retinal image sequence registration.

Moreover, we compare the ablation experimental results between RISeR and RISeR-F2R: (1) In terms of D_{seq}, the evaluation result of RISeR is significantly reduced by approximately 20% compared with that of RISeR-F2R (D_{seq}: 5.632 pixels → **4.485** pixels). This indicates that applying the frame-to-reference and frame-to-frame joint alignment constraints can significantly improve the inter-frame continuity of retinal image sequence registration results; (2) In terms of E_{seq}, RISeR and RISeR-F2R are completely consistent in keypoint extraction and matching, transformation model, optimization algorithm, frame-to-reference alignment constraint, etc., but RISeR still achieves an improvement in the evaluation result of nearly 10% compared with that of RISeR-F2R (E_{seq}: 5.063 pixels → **4.588** pixels). This finding indicates that the joint application of a frame-to-frame alignment constraint and a frame-to-reference alignment constraint helps the transformation model simulate the fundoscopy scene more accurately, thereby achieving better registration results.

Finally, we qualitatively compare the registration results of RISeR and REMPE on the RISeR-real dataset. As shown in Fig. 9, the left two columns correspond to RISeR, the right

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

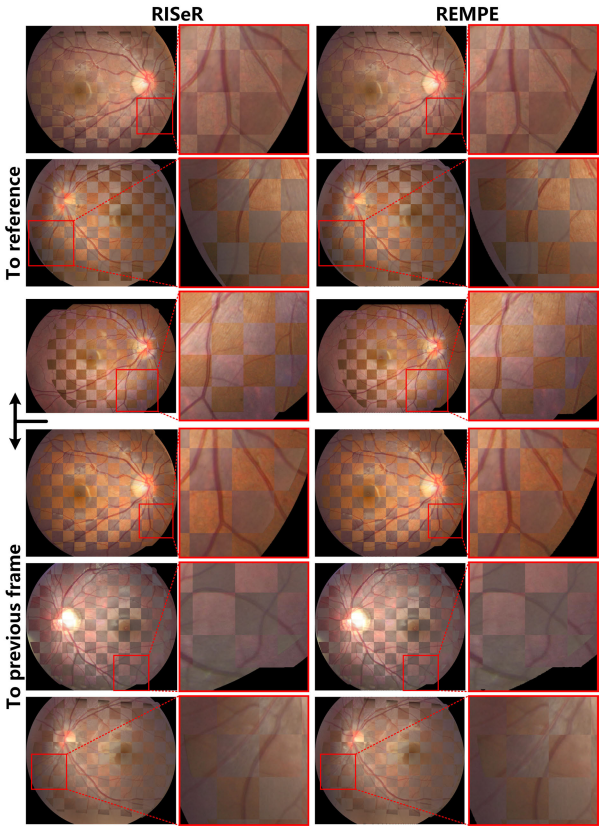

Fig. 9. Registration results of RISeR and REMPE on the RISeR-real dataset. Generate checkerboard images by alternating the corresponding patches in two images to display the registration results. Enlarge the region of interest marked with the red box to further improve readability.

two columns correspond to REMPE, the first three rows are the registration results from the current frame to the reference frame, and the last three rows are the registration results from the current frame to the previous frame. Generally, for blood vessels that pass through multiple patches, the smoother and more natural the connections between adjacent patches are, the better the registration effect. Compared with REMPE, the proposed method RISeR can achieve better registration results in both thicker main blood vessel regions and thinner branch blood vessel regions.

E. Experimental Results on RISeR-Syn

1) Evaluation Metrics: In addition to the **Average Registration Error** $\boldsymbol{E_{seq}}$ calculated by Eq. (17) and the **Average Drift Distance** $\boldsymbol{D_{seq}}$ calculated by Eq. (18), we also calculate the **Average Alignment Deviation** $\boldsymbol{D_{adj}}$ using n_a sets of additionally annotated control points (as shown in Fig. 8) $\{\mathbf{p}^i_{(j,j+1)\_j}, \mathbf{p}^i_{(j,j+1)\_j+1}\}_{i=1}^{n_a}$ that exist only in the common retinal area between two adjacent frames $\{F_j, F_{j+1}\}$ of $\{F_1, \ldots, F_n\}$ and not in the retinal area of reference frame F_0. The specific calculation process is shown in Eq. (19), where len(**J**) represents the length of set **J**. It should be noted that not every pair of adjacent frames in $\{F_1, \ldots, F_n\}$ has such control points between them, so **J** is a subset of $\{1, 2, \ldots, n-1\}$. In fact, this metric evaluates the inter-frame continuity of sequence registration results in a more stringent way compared to D_{seq}, so its value is usually larger than D_{seq}.

$$D_{adj} = \frac{1}{\text{len}(\mathbf{J})} \sum_j \left(\frac{1}{n_a} \sum_{i=1}^{n_a} \left|\left| T_{j\to 0}(\mathbf{p}^i_{(j,j+1)\_j}) - T_{j+1\to 0}(\mathbf{p}^i_{(j,j+1)\_j+1}) \right|\right|_2 \right), \quad j \in \mathbf{J} \subseteq \{1, 2, \ldots, n-1\}. \tag{19}$$

2) Registration Methods: Choose the registration methods used in the experiment of Section V-D except REMPE, which performs relatively poorly on the RISeR-real dataset.

3) Results and Discussion: The experimental results are shown in Table IV. RISeR still achieves the highest registration accuracy and the best inter-frame continuity (E_{seq} = **0.527** pixels, D_{seq} = **0.525** pixels, D_{adj} = **0.996** pixels), and all metrics are significantly improved compared with SuperRetina (E_{seq} = 0.986 pixels, D_{seq} = 1.089 pixels, D_{adj} = 2.255 pixels), with improvements reaching 46.6%, 51.8%, and 55.8% respectively. Even compared with the most competitive method (RIRMD: E_{seq} = 0.637 pixels, D_{seq} = 0.845 pixels, D_{adj} = 1.929 pixels), RISeR still has significant advantages, with improvements of 17.3%, 37.9%, and 48.4% in three error metrics, respectively. Moreover, RISeR achieves the best sequence registration results in five significantly different simulated fundoscopy scenes (*P*, *PC*, *PCI*, *PCS*, *PCSI*), which indicates that the proposed method also has good stability and robustness.

Similarly, we compare and analyze the ablation experimental results between RISeR and RISeR-F2R as follows: (1) In terms of D_{seq}, the evaluation result of RISeR is significantly reduced by 23.4% compared with that of RISeR-F2R (D_{seq}: 0.685 pixels → **0.525** pixels). This once again indicates that applying the joint alignment constraints of frame-to-reference and frame-to-frame can significantly improve the inter-frame continuity of sequence registration results; (2) In terms of D_{adj}, the evaluation result of RISeR is significantly reduced by 36.5% compared with that of RISeR-F2R (D_{adj}: 1.569 pixels → **0.996** pixels). This more strongly demonstrates the effectiveness of the registration strategy adopted by the proposed method; (3) In terms of E_{seq}, the evaluation results of RISeR and RISeR-F2R are close (E_{seq}: 0.552 pixels → **0.527** pixels). We think the main reason is that the impact of frame-to-frame alignment constraint on E_{seq} is indirect, and the registration difficulty of the synthetic data is relatively low. Therefore, for the RISeR-syn dataset, it is understandable that the improvement in E_{seq} is relatively small after RISeR introduces the frame-to-frame alignment constraint.

F. Experimental Results on COph100-45689

1) Evaluation Metrics: Due to the high registration difficulty of some retinal image sequences in COph100-45689, some methods may encounter registration failures, where the registration errors are particularly large (>50 pixels) or the registration results cannot be output at all. Therefore, we first count the **number of failed** registration sequences. Then, we use Eq. (17) and Eq. (18) to calculate the **Average**

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

AVERAGE REGISTRATION ERROR E_{seq}, AVERAGE DRIFT DISTANCE D_{seq}, AND AVERAGE ALIGNMENT DEVIATION D_{adj} ON RISER-SYN FOR THE PROPOSED METHOD, ITS ABLATION VERSION, AND THE COMPARISON METHODS

Methods	*P*			*PC*			*PCI*			*PCS*			*PCSI*			RISeR-syn		
	E_{seq}	D_{seq}	D_{adj}	E_{seq}	D_{seq}	D_{adj}	E_{seq}	D_{seq}	D_{adj}	E_{seq}	D_{seq}	D_{adj}	E_{seq}	D_{seq}	D_{adj}	E_{seq}	D_{seq}	D_{adj}
SuperRetina	0.716	0.749	1.661	0.714	0.866	1.921	1.207	1.409	2.380	0.979	1.080	2.899	1.315	1.341	2.413	0.986	1.089	2.255
RIRMD	0.540	0.766	1.693	0.645	0.868	1.721	0.554	0.738	1.700	0.717	0.914	2.480	0.727	0.942	2.051	0.637	0.845	1.929
RISeR-F2R	0.444	0.514	1.064	0.532	0.725	1.756	**0.485**	0.659	1.599	0.669	0.752	1.742	0.632	0.776	1.686	0.552	0.685	1.569
RISeR	**0.388**	**0.376**	**0.664**	**0.506**	**0.513**	**0.947**	0.494	**0.525**	**0.878**	**0.650**	**0.639**	**1.329**	**0.596**	**0.570**	**1.162**	**0.527**	**0.525**	**0.996**

Bold font represents the best result.

TABLE V

NUMBER OF FAILED, AVERAGE REGISTRATION ERROR E_{seq} AND AVERAGE DRIFT DISTANCE D_{seq} ON COPH100-45689 FOR THE PROPOSED METHOD, AND THE COMPARISON METHODS

Methods	COph100-45689		
	Failed / All	E_{seq} (pixels) ↓	D_{seq} (pixels) ↓
REMPE	14 / 24	6.090	8.156
SuperRetina	5 / 24	6.144	6.303
RIRMD	**0 / 24**	3.436	4.043
RISeR	**0 / 24**	**3.306**	**3.344**

Bold font represents the best result.

TABLE VI

THE IMPACT OF THE NUMBER OF MATCHED KEYPOINTS ON THE REGISTRATION PERFORMANCE OF THE PROPOSED METHOD

Method	Average number of matched keypoints	RISeR-real	
		E_{seq} (pixels) ↓	D_{seq} (pixels) ↓
RISeR	304	**4.588**	**4.485**
	227	4.918	4.766
	154	4.973	4.684
	123	4.827	4.835
	101	5.057	5.118

Bold font represents the best result.

Registration Error E_{seq} and **Average Drift Distance D_{seq}** for the successfully registered retinal image sequences.

2) Registration Methods: Use the same comparison methods (REMPE, SuperRetina, RIRMD) as in the experiment of Section V-D.

3) Results and Discussion: The experimental results are shown in Table V. RISeR achieves not only 0-failed registration but also the highest registration accuracy and the best inter-frame continuity (E_{seq} = **3.306** pixels, D_{seq} = **3.344** pixels). Compared with REMPE (E_{seq} = 6.090 pixels, D_{seq} = 8.156 pixels) and SuperRetina (E_{seq} = 6.144 pixels, D_{seq} = 6.303 pixels), RISeR achieves significant improvements of nearly or more than 50% on both error metrics. Compared with the most competitive method RIRMD (E_{seq} = 3.436 pixels, D_{seq} = 4.043 pixels), RISeR still achieves a significant improvement of 17.3% in inter-frame continuity. These further demonstrate that the proposed method also has good applicability for the progression of non-myopic fundus diseases.

G. The Impact of the Number of Matched Keypoints

The matched keypoints are used as a medium by RISeR for registration, and their quantity usually affects the final registration performance. Therefore, to explore the robustness of the proposed method to the number of matched keypoints, we conduct the following experiment on the RISeR-real dataset: reducing the number of matched keypoints generated by SuperPoint and SuperGlue, and evaluating the registration results of RISeR in this situation. The experimental results are shown in Table VI, which records the average number of matched keypoints in the entire dataset, as well as the corresponding registration evaluation results. As the number of matched keypoints decreases, the overall registration performance of RISeR shows a downward trend. However, the decline process is relatively slow. When the number of matched keypoints decreases by 66.8% (304 → 101), the two error metrics of RISeR increase by only 10.2% and 14.1%, respectively (E_{seq}: **4.588** pixels → 5.057 pixels; D_{seq}: **4.485** pixels → 5.118 pixels). Moreover, comparing Table VI with Table III, we can be surprised to find that even the worst result in Table VI is still in the second-best position in Table III. These findings indicate that the proposed method is robust to changes in the number of matched keypoints.

However, we also have to consider the more extreme scenario in which no usable matched keypoints can be extracted between the images to be registered due to extensive cataracts, severe illumination artifacts, or the absence of any overlapping vascular structures. In this case, the proposed method will undoubtedly fail in registration, as RISeR performs strict pixel-level point-to-point alignment only through explicit hard correspondences of matched keypoints, lacking flexible overall adjustment capabilities. To this end, in future work, we will further introduce implicit soft correspondences of regional feature similarity as an auxiliary. For instance, calculating the local normalized cross-correlation (NCC) [43] between images as part of the objective function provides regional soft correspondence constraints with a wider receptive field to complement point-to-point hard correspondence constraints.

H. The Impact of the Device Specifications of Fundus Camera

The device specifications of fundus camera include two important parameters: (1) the lens-to-cornea distance l; (2) the field of view k. In the model constructed by RISeR, they are known parameters that need to be preset. However, there are significant differences in the device specifications of different fundus cameras. Therefore, to explore the

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

TABLE VII

THE IMPACT OF THE DEVICE SPECIFICATIONS OF FUNDUS CAMERA ON THE REGISTRATION PERFORMANCE OF THE PROPOSED METHOD

Method	l	k	RISeR-real	
			E_{seq} (pixels) ↓	D_{seq} (pixels) ↓
RISeR	45.7mm	45°	**4.588**	**4.485**
	34.8mm	30°	4.947	4.834
	25mm	133°	4.866	4.607

Bold font represents the best result.

TABLE VIII

THE IMPACT OF PSO COMPUTATIONAL BUDGET ON THE REGISTRATION PERFORMANCE OF THE PROPOSED METHOD

E_{seq}/D_{seq} of RISeR on RISeR-real dataset					
		g			
		100	200	300	400
p	1000	6.296/7.221	6.017/6.153	5.909/5.955	5.351/5.675
	5000	5.378/5.926	5.261/5.359	5.545/5.459	4.927/4.932
	10000	5.643/5.867	5.108/5.218	**4.588/4.485**	4.801/4.886
	15000	5.447/5.771	4.812/4.679	5.190/5.109	4.763/4.603

Bold font represents the best result.

robustness of the proposed method to the device specifications of fundus cameras, we conduct the following experiment on the RISeR-real dataset: using the device specifications of another narrow-field fundus camera (l = 34.8mm, k = 30°) and a wide-field fundus camera (l = 25mm, k = 133°), respectively, and evaluating the registration results of RISeR under these parameter settings. The experimental results are shown in Table VII, which records the device specifications of different fundus cameras, and the corresponding registration evaluation results. As we can see, using the device specifications of other fundus cameras results in a slight decrease in registration performance. Interestingly, the performance degradation caused by using a smaller k is greater than that caused by using a larger k. We think the main reasons are that: (1) The pixels originally distributed in the 45° FOV are compressed into the 30° FOV, which leads to the loss of some information. This directly affects the sensitivity of the model to alignment errors between 3D corresponding points during the optimization process; (2) Although being magnified into the 133° FOV can easily introduce greater distortion, this can be compensated by adaptively adjusting the distortion parameters in the model during the optimization process. Moreover, comparing Table VII with Table III, we can also find that even the worst result in Table VII can still occupy the second-best position with a significant advantage in Table III. This indicates that the proposed method is robust to the device specifications of fundus cameras.

I. The Impact of PSO Computational Budget

The computational budget of PSO includes two parts: the number of particles p and the iterations g, which are two key parameters that affect the solution. To investigate the impact of the PSO computational budget on the registration performance of the proposed method, we conduct the following experiment on the RISeR-real dataset: setting p = 1000, 5000, 10000, 15000; setting g = 100, 200, 300, 400; evaluating the registration results of RISeR under these 16 different parameter combinations, respectively. The experimental results are shown in Table VIII, which records the different parameter settings and the corresponding registration evaluation results. The parameter combination of p = 10000 and g = 300 achieves the best registration performance. Moreover, among the eight neighbors of the bolded result in Table VIII, half of them can still occupy the second-best position in Table III, and almost all of them can compete head-on with the comparison methods in Table III, especially the second error metric D_{seq}. It is worth noting that, except for the two parameter combinations with lower computational budget, p = 1000, g = 100 and p = 1000, g = 200, the remaining 14 parameter combinations all can outperform the three comparison methods (REMPE, SuperRetina, RIRMD) in Table III in terms of the error metric D_{seq}. These results indicate that the proposed method has good robustness to the parameters p and g of PSO computational budget.

TABLE IX

THE RUNTIME OF THE PROPOSED METHOD AND THE COMPARISON METHODS

Methods	REMPE	RIRMD	RISeR
Runtime (s)	488	600	826

TABLE X

THE RUNTIME FOR DIFFERENT STAGES OF THE PROPOSED METHOD

Stages		Runtime (s)
Frame-to-reference	Keypoint extraction and matching	1
	PSO	235
	Image formation	71
Frame-to-reference and frame-to-frame	Keypoint extraction and matching	1
	PSO	461
	Image formation	57

J. Runtime Analysis

This experiment compares the differences in runtime between the proposed method (RISeR) and two highly relevant methods (REMPE, RIRMD), and analyzes the runtime for different stages of the proposed method to guide further improvement directions. Considering that the minimum number of images required to form a sequence is 3, we randomly select three images from a sequence in the RISeR-real dataset as the experimental data. The resolutions of these three images are 2544 × 1696, 2592 × 1728, and 1924 × 1556, respectively. The experimental results are shown in Table IX, which records the runtime of REMPE, RIRMD, and RISeR, in seconds. These results are obtained by running on a computer equipped only with CPU, as mentioned in Section V-B. Although both REMPE and RIRMD have shorter runtimes than RISeR, Table III shows that their accuracy is significantly worse than that of the proposed method. Furthermore, Table X shows the runtime of each stage of the proposed method. The fastest stage is keypoint extraction and matching. The next fastest

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

is the image formation stage, which involves two mapping processes between 2D image points and 3D retinal points, resulting in a slower speed. The specific runtime depends mainly on the image resolution. Finally, the PSO stage is the most time-consuming, and the joint alignment constraints take longer than the single alignment constraint because the increase in the number of constraints leads to an increase in computation.

Although the proposed method is slower than the comparison methods, it provides more accurate and stable registration with a significant improvement, which is the main focus of retinal image global registration algorithms used for longitudinal studies. More importantly, the focus of this paper is to develop a new transformation model and significantly improve the upper limit of the performance of retinal image global registration, to promote further study based on the model proposed in this paper. In future work, we will conduct two studies targeting the identified bottlenecks: (1) comprehensive parallelization acceleration of image formation and PSO optimization for the proposed model on GPU; (2) integrating the proposed model with deep learning networks to develop an algorithm that combines both accuracy and real-time.

VI. Conclusion

In this paper, we propose a retinal image sequence registration method for longitudinal study, which contains two key elements: 3D spatiotemporal model and joint registration strategy. The former simulates a longitudinal fundoscopy scene by constructing the dynamic eyeball model and the camera array model. And benefiting from this, we can efficiently generate many realistic synthetic retinal image sequences. The latter specifically considers the inter-frame continuity that is also of concern in the sequence registration task, and develops an algorithmic framework for registering retinal image sequences based on the joint alignment constraint strategy of frame-to-reference and frame-to-frame. The experimental results show that the proposed method achieves state-of-the-art registration accuracy, reliability, and applicability on two publicly available retinal image pair registration datasets, two newly constructed retinal image sequence registration datasets, and a publicly available retinal image sequence registration dataset.

Like investigating and solving cases, exploring the physical models inside and simulating the corresponding scene based on them can often yield satisfactory results that are surprising yet completely reasonable and interpretable. In the future, we will further develop the deep learning algorithm embedded with the 3D spatiotemporal model for direct registration of the entire retinal image sequence. At last, in order to further promote the study in related fields, we make the code of the proposed method,[1] the code for generating synthetic data,[2] the RISeR-real[3] dataset, and the RISeR-syn[4] dataset publicly available to the scientific community.

[1]https://github.com/ZengshuoWang/RISeR

[2]https://github.com/ZengshuoWang/RISeR-syn-generate

[3]https://dx.doi.org/10.21227/xpe1-e969

[4]https://dx.doi.org/10.21227/qcs2-aw02

References

[1] H. Narasimha-Iyer, A. Can, B. Roysam, H. L. Tanenbaum, and A. Majerovics, "Integrated analysis of vascular and nonvascular changes from color retinal fundus image sequences," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 8, pp. 1436–1445, Aug. 2007.

[2] J. Zhang et al., "Two-step registration on multi-modal retinal images via deep neural networks," *IEEE Trans. Image Process.*, vol. 31, pp. 823–838, 2022.

[3] C. Samarawickrama et al., "Myopia-related optic disc and retinal changes in adolescent children from Singapore," *Ophthalmology*, vol. 118, no. 10, pp. 2050–2057, Oct. 2011.

[4] J. B. Jonas, Q. Zhang, L. Xu, W. B. Wei, R. A. Jonas, and Y. X. Wang, "Change in the ophthalmoscopical optic disc size and shape in a 10-year follow-up: The Beijing eye study 2001–2011," *Brit. J. Ophthalmology*, vol. 107, no. 2, pp. 283–288, Feb. 2023.

[5] Y. Guo et al., "Parapapillary gamma zone and progression of myopia in school children: The Beijing children eye study," *Investigative Opthalmology Vis. Sci.*, vol. 59, no. 3, pp. 1609–1616, Mar. 2018.

[6] J. B. Jonas, K. Ohno-Matsui, R. F. Spaide, L. Holbach, and S. Panda-Jonas, "Macular Bruch's membrane defects and axial length: Association with gamma zone and delta zone in peripapillary region," *Investigative Opthalmology Vis. Sci.*, vol. 54, no. 2, pp. 1295–1302, Feb. 2013.

[7] R. A. Jonas, Y. N. Yan, Q. Zhang, Y. X. Wang, and J. B. Jonas, "Elongation of the disc-fovea distance and retinal vessel straightening in high myopia in a 10-year follow-up of the Beijing eye study," *Sci. Rep.*, vol. 11, no. 1, p. 9006, Apr. 2021.

[8] P. Xue, Y. Fu, H. Ji, W. Cui, and E. Dong, "Lung respiratory motion estimation based on fast Kalman filtering and 4D CT image registration," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 2007–2017, Jun. 2021.

[9] L. Chen, Y. Xiang, Y. Chen, and X. Zhang, "Retinal image registration using bifurcation structures," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2169–2172.

[10] X. Zhang, X. Lu, J. Li, Y. Gong, Q. Wang, and Y. Yin, "Two-phase parametric registration for retinal images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2024, pp. 1–6.

[11] C. Hernandez-Matas, X. Zabulis, and A. A. Argyros, "REMPE: Registration of retinal images through eye modelling and pose estimation," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 12, pp. 3362–3373, Dec. 2020.

[12] J. Chen, J. Tian, N. Lee, J. Zheng, R. T. Smith, and A. F. Laine, "A partial intensity invariant feature descriptor for multimodal retinal image registration," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1707–1718, Jul. 2010.

[13] N. Ryan, C. Heneghan, and P. de Chazal, "Registration of digital retinal images using landmark correspondence by expectation maximization," *Image Vis. Comput.*, vol. 22, no. 11, pp. 883–898, Sep. 2004.

[14] C. V. Stewart, C.-L. Tsai, and B. Roysam, "The dual-bootstrap iterative closest point algorithm with application to retinal image registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 11, pp. 1379–1394, Nov. 2003.

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

[16] A. Can, C. V. Stewart, B. Roysam, and H. L. Tanenbaum, "A feature-based, robust, hierarchical algorithm for registering pairs of images of the curved human retina," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 347–364, Mar. 2002.

[17] J. Liu, X. Li, Q. Wei, J. Xu, and D. Ding, "Semi-supervised keypoint detector and descriptor for retinal image matching," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 593–609.

[18] J. Liu and X. Li, "Geometrized transformer for self-supervised homography estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9556–9565.

[19] Y. Wang et al., "SuperJunction: Learning-based junction detection for retinal image registration," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 1, pp. 292–300, Mar. 2024.

[20] H. Ji et al., "A non-rigid image registration method based on multi-level B-spline and L2-regularization," *Signal, Image Video Process.*, vol. 12, no. 6, pp. 1217–1225, Sep. 2018.

[21] P. Xue, E. Dong, and H. Ji, "Lung 4D CT image registration based on high-order Markov random field," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 910–921, Apr. 2020.

[22] A. Zakeri et al., "DragNet: Learning-based deformable registration for realistic cardiac MR sequence generation from a single frame," *Med. Image Anal.*, vol. 83, Jan. 2023, Art. no. 102678.

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.

[23] D. Mahapatra, B. Antony, S. Sedai, and R. Garnavi, "Deformable medical image registration using generative adversarial networks," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1449–1453.
[24] Y. Liu et al., "Progressive retinal image registration via global and local deformable transformations," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2024, pp. 2183–2190.
[25] M. Nakao, K. Kobayashi, J. Tokuno, T. Chen-Yoshikawa, H. Date, and T. Matsuda, "Deformation analysis of surface and bronchial structures in intraoperative pneumothorax using deformable mesh registration," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102181.
[26] Z. Wang et al., "Retinal image registration method for myopia development," *Med. Image Anal.*, vol. 97, Oct. 2024, Art. no. 103242.
[27] J. Wang et al., "Gaussian field estimator with manifold regularization for retinal image registration," *Signal Process.*, vol. 157, pp. 225–235, Apr. 2019.
[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[29] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis.*, vol. 3951. May 2006, pp. 404–417.
[30] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.
[31] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 12405–12415.
[32] C. An, Y. Wang, J. Zhang, and T. Q. Nguyen, "Self-supervised rigid registration for multimodal retinal images," *IEEE Trans. Image Process.*, vol. 31, pp. 5733–5747, 2022.
[33] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[34] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4938–4947.
[35] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 10–5244.
[36] G. Yang, C. V. Stewart, M. Sofka, and C.-L. Tsai, "Registration of challenging image pairs: Initialization, estimation, and decision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 11, pp. 1973–1989, Nov. 2007.
[37] G. Wang, Z. Wang, Y. Chen, and W. Zhao, "Robust point matching method for multimodal retinal image registration," *Biomed. Signal Process. Control*, vol. 19, pp. 68–76, May 2015.
[38] 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 Med. Imag. Graph.*, vol. 37, nos. 7–8, pp. 597–606, Oct. 2013.
[39] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, Sep. 1997.
[40] Y. Bizais, C. Barillot, and R. Di Paola, "Automated multimodality image registration using information theory," in *Proc. Inf. Process. Med. Imag.*, 1995, pp. 263–274.
[41] R. Kolar, V. Harabis, and J. Odstrcilik, "Hybrid retinal image registration using phase correlation," *Imag. Sci. J.*, vol. 61, no. 4, pp. 369–384, May 2013.
[42] K. M. Adal et al., "A hierarchical coarse-to-fine approach for fundus image registration," in *Proc. 6th Int. Workshop Biomed. Image Registration*, 2014, pp. 93–102.
[43] Y. R. Rao, N. Prathapani, and E. Nagabhooshanam, "Application of normalized cross correlation to image registration," *Int. J. Res. Eng. Technol.*, vol. 3, no. 17, pp. 12–16, May 2014.
[44] L. Ding, T. D. Kang, A. E. Kuriyan, R. S. Ramchandran, C. C. Wykoff, and G. Sharma, "Combining feature correspondence with parametric chamfer alignment: Hybrid two-stage registration for ultra-widefield retinal images," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 2, pp. 523–532, Feb. 2023.
[45] T. Eun Choe and I. Cohen, "Registration of multimodal fluorescein images sequence of the retina," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2005, pp. 106–113.
[46] Y. Lin and G. Medioni, "Retinal image registration from 2D to 3D," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[47] A. Perez-Rovira, R. Cabido, E. Trucco, S. J. McKenna, and J. P. Hubschman, "RERBEE: Robust efficient registration via bifurcations and elongated elements applied to retinal fluorescein angiogram sequences," *IEEE Trans. Med. Imag.*, vol. 31, no. 1, pp. 140–150, Jan. 2012.
[48] C.-L. Tsai, H.-C. Hsu, X.-C. Wu, S.-J. Chen, and W.-Y. Lin, "Accurate joint-alignment of indocyanine green and fluorescein angiograph sequences for treatment of subretinal lesions," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 3, pp. 785–793, May 2017.
[49] J. B. Jonas, K. Ohno-Matsui, L. Holbach, and S. Panda-Jonas, "Association between axial length and horizontal and vertical globe diameters," *Graefe's Arch. for Clin. Experim. Ophthalmology*, vol. 255, no. 2, pp. 237–242, Feb. 2017.
[50] J. B. Jonas, R. F. Spaide, L. A. Ostrin, N. S. Logan, I. Flitcroft, and S. Panda-Jonas, "IMI—Nonpathological human ocular tissue changes with axial myopia," *Investigative Ophthalmology Vis. Sci.*, vol. 64, no. 6, p. 5, 2023.
[51] R. Navarro, J. Santamaría, and J. Bescós, "Accommodation-dependent model of the human eye with aspherics," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 8, pp. 1273–1280, 1985.
[52] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. ICNN*, vol. 4, 1995, pp. 1942–1948.
[53] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A. A. Argyros, "FIRE: Fundus image registration dataset," *Model. Artif. Intell. Ophthalmology*, vol. 1, no. 4, pp. 16–28, Jul. 2017.
[54] Y. Hu et al., "COph100: A comprehensive fundus image registration dataset from infants constituting the 'RIDIRP' database," *Scientific Data*, vol. 12, no. 1, p. 99, Jan. 2025.
[55] C. Y. Cheung, F. Tang, D. S. W. Ting, G. S. W. Tan, and T. Y. Wong, "Artificial intelligence in diabetic eye disease screening," *Asia–Pacific J. Ophthalmology*, vol. 8, no. 2, pp. 158–164, 2019.

Authorized licensed use limited to: NANKAI UNIVERSITY. Downloaded on August 27,2026 at 09:24:09 UTC from IEEE Xplore. Restrictions apply.