Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-024-01154-2

RDLR: A Robust Deep Learning-Based Image Registration Method for Pediatric Retinal Images

Hao Zhou[1] · Wenhan Yang[1] · Limei Sun[1] · Li Huang[1] · Songshan Li[1] · Xiaoling Luo[1] · Yili Jin[1] · Wei Sun[2] · Wenjia Yan[1] · Jing Li[3] · Xiaoyan Ding[1] · Yao He[1] · Zhi Xie[1]

Received: 4 March 2024 / Revised: 24 May 2024 / Accepted: 24 May 2024

Abstract

Retinal diseases stand as a primary cause of childhood blindness. Analyzing the progression of these diseases requires close attention to lesion morphology and spatial information. Standard image registration methods fail to accurately reconstruct pediatric fundus images containing significant distortion and blurring. To address this challenge, we proposed a robust deep learning–based image registration method (RDLR). The method consisted of two modules: registration module (RM) and panoramic view module (PVM). RM effectively integrated global and local feature information and learned prior information related to the orientation of images. PVM was capable of reconstructing spatial information in panoramic images. Furthermore, as the registration model was trained on over 280,000 pediatric fundus images, we introduced a registration annotation automatic generation process coupled with a quality control module to ensure the reliability of training data. We compared the performance of RDLR to the other methods, including conventional registration pipeline (CRP), voxel morph (WM), generalizable image matcher (GIM), and self-supervised techniques (SS). RDLR achieved significantly higher registration accuracy (average Dice score of 0.948) than the other methods (ranging from 0.491 to 0.802). The resulting panoramic retinal maps reconstructed by RDLR also demonstrated substantially higher fidelity (average Dice score of 0.960) compared to the other methods (ranging from 0.720 to 0.783). Overall, the proposed method addressed key challenges in pediatric retinal imaging, providing an effective solution to enhance disease diagnosis. Our source code is available at https://github.com/wuwusky/RobustDeepLeraningRegistration.

Keywords Image registration · Automatic registration annotation framework · Panoramic fundus imaging · Refinement module

Introduction

Vision is integral to a child's development and learning, with most visual functions dependent on a healthy retinal structure [1, 2]. Globally, retinal disease is a predominant cause of childhood visual impairment. The World Health Organization reported that approximately 13 out of every 1000 children suffer from vision loss, largely due to retinal abnormalities [3]. The major pediatric retinal diseases include retinopathy of prematurity (ROP), pediatric retinal detachment, congenital retinal diseases, and pediatric retinal tumors [4–7]. Left undiagnosed and untreated, these conditions can result in irreversible blindness or vision loss in children. To reduce this risk, many nations recommend fundus screening within the first 6 weeks after birth to facilitate prompt diagnosis and treatment of retinal issues [8, 9].

Accurate diagnosis of pediatric retinal disease necessitates a thorough examination of entire retinal area, including

Hao Zhou, Wenhan Yang, and Limei Sun equally contributed.

✉ Xiaoyan Ding
dingxiaoyan@gzzoc.com

✉ Yao He
scheyao@hotmail.com

✉ Zhi Xie
xiezhi@gmail.com

[1] State Key Laboratory of Ophthalmology, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

[2] Department of Ophthalmology, Guangdong Eye Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China

[3] Department of Ophthalmology, Guangdong Women and Children Hospital, Guangzhou, China

Published online: 14 June 2024

the peripheral, posterior pole, and macular regions. This comprehensive view is constructed from multiple wide-field images captured at different orientations to ensure complete coverage [8–12]. Examining the periphery is particularly important, as abnormalities typically manifest first in the extreme periphery before affecting the posterior pole [13]. However, capturing pediatric retinal images poses significant challenge due to their inability to maintain steady fixation, leading to distortion, blurring, variable illumination, and incomplete coverage [8, 9, 13]. Furthermore, existing retinal abnormalities, such as hemorrhages and scarring, can complicate imaging by affecting camera settings and lighting effects, potentially introducing distortions and artifacts. These factors degrade image quality and hinder accurate alignment and fusion of individual images to reconstruct a panoramic view essential for precision diagnosis [14–19].

To reconstruct a panoramic view, accurate image registration technology is required. The conventional registration pipeline (CRP) is a standard method that automatically achieves image registration without manual intervention, following a "detection-matching-screening" strategy [20–27]. However, this approach is not suitable for pediatric fundus images, which are typically of low quality with limited features, suffering from issues such defocus, light leakage, artifacts, insufficient contrast, halo, and avascular areas. These issues can lead to unrealistic transformations and degrade registration accuracy [19, 28].

More recently, deep learning registration methods, such as voxel morph (VM) [29] and self-supervised (SS) techniques, have been introduced [30–35]. VM learns a deformation field in an unsupervised manner and finds applications in various medical imaging tasks [32, 35, 36], while SS employs synthetically warped images as training data to overcome manual labeling challenges [37–41]. However, these methods heavily rely on abundant coherent pattern features within the images to effectively constrain the warping process, which are often insufficient in low-quality pediatric retinal images [33, 40]. Moreover, these features are necessary for determining the image orientation order and reconstructing the panoramic view.

Moreover, while deformation fields, including affine transformations, are utilized to maximize overlap, an over-reliance on local features for deformation can result in the distortion of image geometry. This distortion may be incompatible with diagnostic applications [29, 32]. It is essential to supplement additional feature information (e.g., global semantic feature) and more reliable constraint information to supervise and optimize the method, with the aim of ensuring that the results derived from deformation information are accurate and acceptable. On the other hand, supervised methods exhibit superior performance but demand a large number of manually annotated pairs for training [42], making it impractical for pediatric fundus image registration. Additionally, inspired by large Language model (LLM), recent research [43] has dismantled internet videos into extensive datasets and developed models with numerous parameters, resulting in a trained general image registration model. However, this approach awaits verification within the relevant fields, including pediatric fundus images.

Therefore, we developed the RDLR, a deep learning-based method, to achieve stable and accurate registration of low-quality, feature-sparse images. The method was trained under supervision using self-annotated data by CRP to guarantee dependability.

Related Work

The conventional registration pipeline (CRP) typically involve "feature detection, feature match, pair selection," using algorithms like Speeded Up Robust Features (SURF), Scale Invariant Feature Transform (SIFT), Oriented Fast and Rotated BRIEF (ORB) [20, 44, 45], and integration features [46]. However, these methods rely on local texture information [20, 47] and are not suitable for low-quality pediatric fundus images, which often lack reliable landmarks and discriminative features.

To overcome these limitations, researchers have leveraged deep learning techniques for image registration. Initial efforts focused on enhancing local feature extraction using learned approaches like learned invariant feature transform (LIFT) [48], deep learning-based feature-based registration (DL-FBR) [49], and other works [50, 51]. However, these methods still operate within the fragmented conventional framework [52] and fail to capture global context.

Recently, end-to-end CNN architectures have been explored for image registration. One approach utilizes deformation fields to predict alignments, exemplified by voxel morph [29] and other methods [53, 54], while others directly predict registration results, as exemplified by prior works [31, 34, 38, 40, 55, 56]. The training strategy for these methods predominantly relies on the self-supervised or unsupervised approach. However, these methods may struggle with discontinuous images, significant displacements, and imaging noise. Additionally, certain representative methods [54, 57, 58] required supplementary information, such as vessel segmentation, for precise registration, and these methods could not be transferred without the necessary segmentation models.

While existing deep learning-based methods have achieved considerable success on high-quality images, they often fall short on low-quality pediatric retinal images. This limitation highlights the need for a more comprehensive approach that can effectively handle a wider range of image types and quality levels.

Concurrently, the leveraging of large language model (LLM) has led to the integration of models with a vast number of parameters and massive datasets as one of the preferred approaches for solving general image processing problems. The research presented in Ref. 43 is the latest representative of this kind, utilizing 100 h of internet video (amounting to tens of millions of frame images when decomposed) to train a generalizable image registration model. This method has consistently outperformed other methodologies across diverse image registration tasks. However, it has not yet been validated using the pediatric fundus image data employed in the current study.

In summary, an effective registration solution that integrates global and local information is essential for multi-angle pediatric fundus images, addressing the scarcity of real training data and the challenges associated with registering discontinuous data and low-quality images with high noise.

Methods

Robust Deep Learning-Based Image Registration Method (RDLR)

RDLR is a manual-label free deep learning framework designed to accurately reconstruct panoramic retinal views from multiple images captured at different orientations (Fig. 1a). The framework operates through a sequential integration of two key modules: the Registration Module (RM) (Fig. 1b and c) and the Panoramic View Module (PVM) (Fig. 1d).

The process begins with the RM, which estimates pairwise alignment between images. This module comprises a deep neural network-based registration model (Fig. 1b) and a refinement module (Fig. 1c). Once the images are aligned, the PVM takes over to seamlessly fuse the registered images into a complete panoramic view (Fig. 1d). This final step produces the desired panoramic retinal image, which is the output of the RDLR framework.

To create training data with registration annotations, a novel automatic registration annotation generation framework was developed (Fig. 1e).

Registration Model of Registration Module (RM)

The core of the RM is a learning regression model tailored to predict the registration information and transformation relationship between two input retinal images. The model unfolds in two distinctive phases: feature extraction phase and registration prediction phase, and the detailed information of model architecture is provided in the Supplementary Material (Fig. S1).

Feature Extraction Phase This initial phase focuses on extracting essential features from the input images, laying the foundation for subsequent registration predictions. The architecture involves the following:

Dual branches design: two branches are designed for feature extraction, incorporating the efficientNet-B1 architecture as the backbone. This design, reminiscent of a Siamese network model, enhances the model's ability to capture intricate text features.

$$f_n = F_e(I_n) \tag{1}$$

Since the input is two, $f_n(f_1,f_2)$ are the extracted features of input images $I_n(I_1, I_2)$ from the branches F_e.

Context attention module: an additional context attention module is introduced to augment the extraction of relevant texture features. This module refines the network's focus on discriminative information.

$$I_{att} = I * \left(F_{ca}(I) + 1\right) \tag{2}$$

I_{att} is the texture augmentation of input image I, and F_{ca} is the context attention module which is regularized by six layers of convolution and concludes with the final sigmoid activation layer. Therefore, by combining Eqs. (1) and (2), the feature extraction phase can be expressed as

$$f_n = F_e\left(I * \left(F_{ca}\left(I_n\right) + 1\right)\right) \tag{3}$$

As depicted in Fig. 1b, the input images are processed by two branches to extract features independently. To enhance efficiency, a weight-sharing strategy is employed, enabling the branches to share weights. This strategy effectively reduces the number of model parameters by utilizing a single shared branch to compute features for multiple inputs.

Registration Prediction Phase In the second phase, the network leverages the extracted features to predict registration parameters. The architecture encompasses the following:

Grid estimation module: a five-layer convolution network is employed to fuse features from both branches, enhancing the model's capacity to grasp intricate details. This is followed by a two-layer convolution network responsible for predicting the offset.

$$\text{out}_{\text{offset}} = F_g\left(f_1 \oplus f_2\right) \tag{4}$$

F_g is the grid estimation module and $f_1 \oplus f_2$ means concatenating the two feature vectors extracted by Eq. (3) along the dimension, and $\text{out}_{\text{offset}}$ is the estimated value of the offset in the x and y directions, each containing 1024 grid points (1024 * 1 * 2).

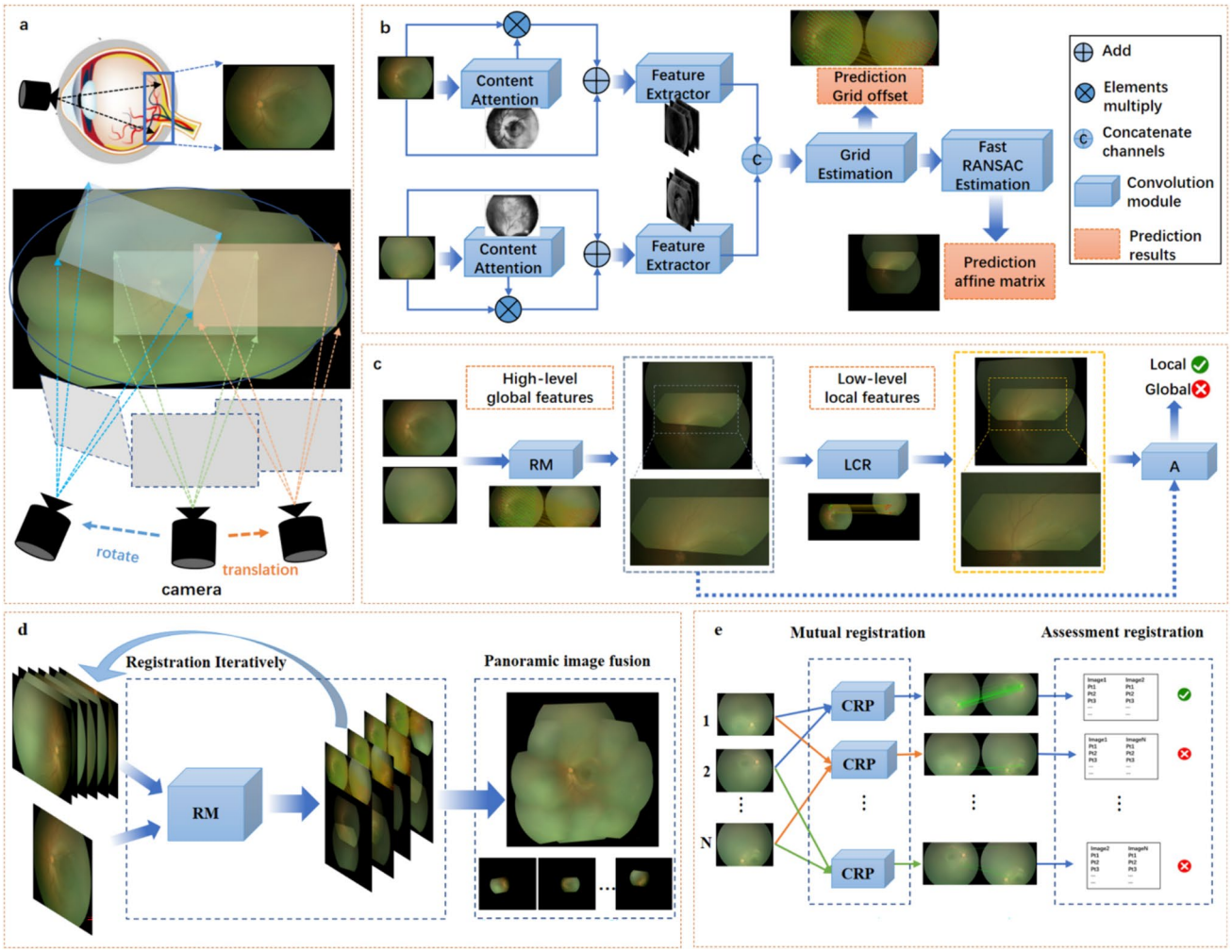

Fig. 1 Overview of robust deep learning-based image registration method (RDLR). **a** Top: A single image captures a portion of the retina. Bottom: Images adequately cover the retinal region through displacement and rotation using affine transformation. **b** Overview of the registration module (RM) of the RDLR. From left to right, the RM module is composed of the context attention module, feature extractor module, grid estimation module, and fast RANSAC estimation module. These modules enable the prediction and estimation of grid point offsets and affine transformation matrix. **c** Overview of the RM with refinement module. From left to right, RM refers to the registration model, and LCR stands for local context registration based on confined search area and CRP (conventional registration pipeline). A denotes the analysis module used to assess the reliability of local features. **d** Overview of the panoramic view module (PVM) of the RDLR, and the resulting panoramic image. Left: RM represents the registration model consisting of the registration module with refinement. PVM utilizes this module to recursively align multiple images with the target image. Right: After merging the registration results, the images are synthesized into a panoramic view, which can also be split into images with specific spatial orientations. **e** Overview of the automatic framework for registration annotation based on conventional registration pipeline (CRP)

Fast RANSAC Estimation Module After the aforementioned modules, global feature information from the input images is extracted and fused. Subsequently, the grid point offset matrix is estimated. The fast RANSAC estimation module derives the six variable values of the transformation matrix from the predicted offset. Two implementations are introduced: matrix estimation head with local information based on 128 randomly sampled grid points, employed for inference and fine tuning.

$$\mathrm{pts}_{\mathrm{combine}} = \mathrm{combine}\left(\mathrm{pts}_{\mathrm{initial}}, \left(\mathrm{pts}_{\mathrm{initial}} + \mathrm{out}_{\mathrm{offset}}\right)\right) \tag{5}$$

$\mathrm{pts}_{\mathrm{initial}}$ is the known 1024 grid points, $\mathrm{pts}_{\mathrm{initial}} + \mathrm{out}_{\mathrm{offset}}$ is the target 1024 grid points, and $\mathrm{pts}_{\mathrm{combine}}$ is the combination of 128 pairs of points randomly sampled from $\mathrm{pts}_{\mathrm{initial}}$ and $\mathrm{pts}_{\mathrm{initial}} + \mathrm{out}_{\mathrm{offset}}$.

$$\mathrm{h}_{\mathrm{combine}} = F_{\mathrm{homograph}}\left(\mathrm{pts}_{\mathrm{combine}}\right) \tag{6}$$

$F_{\text{homograph}}$ means polynomial solving which is employed to compute the corresponding homograph matrices h_{combine} for the pairs of randomly sampled points,

$$E_h = \text{abs}\left(\text{pts}_{\text{initial}} * h_{\text{combine}} - (\text{pts}_{\text{initial}} + \text{out}_{\text{offset}})\right) \tag{7}$$

The coordinates of known grid points, $\text{pts}_{\text{initial}}$, were multiplied by the estimated h_{combine} to obtain the transformed coordinates of the grid points. The known grid points, added to the estimated offsets, were used as the target grid point coordinates. The Euclidean distance between the transformed set and the target set was then computed as the all estimation homograph matrix errors E_h.

$$M = \text{argmax}\left(\text{count}\left(E_h < \text{threshold}\right)\right) \tag{8}$$

Finally, the number of point pairs corresponding to error matrices that satisfy a present threshold is tallied, and the estimation matrix corresponding to the index with the maximum count is selected as the final prediction result M.

Additionally, a two-layer linear full-connected layer module F_{out} estimating the matrix M with full grid offset information $\text{out}_{\text{offset}}$, designed for training only which is called matrix estimation head with global information.

$$M = F_{\text{out}}(\text{out}_{\text{offset}}) \tag{9}$$

Refinement Module of Registration Module (RM)

The refinement module is the final step in the RM following registration model. This involved the following steps.

1. The overlapping area of the registration was detected by comparing the predicted mask and initial mask,

$$\begin{aligned}\text{Mask}_{\text{insection}} &= \text{Mask}_{\text{Initial}} \cap \text{Mask}_{\text{Predicted}} \\ &= \text{Mask}_{\text{Initial}} \cap \left(\text{Mask}_{\text{Initial}} * M\right)\end{aligned} \tag{10}$$

   Here, $\text{Mask}_{\text{Initial}}$ is the unified initial mask for the input image pair, $\text{Mask}_{\text{Initial}} * M$ is the predicted mask $\text{Mask}_{\text{Predicted}}$ obtained by multiplying the initial mask by the transformation matrix M from Eq. (9), and $\text{Mask}_{\text{insection}}$ represents the intersected region between them, as shown in Figs. 1c and S2a.
2. Local features of the overlapping area from both images were extracted using the SURF algorithm, and matched into pairs and sorted by the similarity distance of the local features,

$$Kp_t, Kp_a = \text{SURF}_{\text{DetectFeature}}(I_t * \text{Mask}_{\text{insection}}, I_a * \text{Mask}_{\text{insection}})$$

$$E_{\text{pair}} = Sort\left(F_e\left(Kp_t, Kp_a\right), 512\right) \tag{11}$$

   After obtaining the intersected mask $\text{Mask}_{\text{insection}}$, the feature information within the overlapping region is preserved for both input images I_t and I_a using the mask. The SURF algorithm is then applied to detect texture feature points and calculated the local feature vectors of the feature points within the overlapping region of two input images. This process results in the coordinates and feature vectors Kp_t and Kp_a. Subsequently, each feature point in Kp_t and Kp_a is pairwise matched, and the Euclidean distance between the corresponding feature vectors is computed. The distance was then sorted, and the indices of the top 512 pairs, E_{pair}, with the smallest distances are selected.
3. The top 512 pairs were used to estimate a new transformation matrix to optimize the registration results using the fast RANSAC estimation module,

$$M_{\text{new}} = F_{\text{homograph}}\left(Kp_{\text{com}}\left(E_{\text{pair}}\right)\right) \tag{12}$$

   The reliable point pairs are extracted from all point pairs Kp_{com} based on the top 512 indices E_{pair}, and a new homograph matrix M_{new} is estimated using these reliable pairs to optimize the registration results.
4. We evaluated the reliability of the optimization results to determine whether to accept or reject them. The criterion for acceptance was that if the magnitude of the refinement result's movement or rotation was too large compared to the original predicted result from the registration model, the refinement was considered unreliable.

$$\text{ratio} = \frac{\text{count}\left(\text{Mask}_{\text{predicted}}\right)}{\text{count}\left(\text{Mask}_{\text{refinement}}\right)} = \frac{\text{count}\left(\text{Mask}_{\text{Initial}} * M\right)}{\text{count}\left(\text{Mask}_{\text{Initial}} * M_{\text{new}}\right)}$$

$$\begin{cases} M_{\text{result}} = M_{\text{new}}, & \textbf{if ratio} > \theta_{\max} \textbf{ or ratio} < \theta_{\min} \\ M_{\text{result}} = M, & \textbf{if } \theta_{\min} \le \text{ratio} \le \theta_{\max} \end{cases} \tag{13}$$

The two transformation matrices M_{new} and M obtained from formulas (9) and (12) are used to calculate the transformed masks $\text{Mask}_{\text{refinement}}$ and $\text{Mask}_{\text{predicted}}$ separately (Fig. S2b). The non-zero pixel count in both masks is computed, and the ratio of the two counts is calculated. Based on the predefined thresholds, if the ratio is either greater than $\theta_{\max}$ or less than $\theta_{\min}$, the refined result M_{new} is accepted. Otherwise, if $\theta_{\min} \le \text{ratio} \le \theta_{\max}$, the original predicted result M is retained.

Panoramic View Module (PVM)

PVM uses the trained RM to align multiple images onto a target orientation image in a recursive manner. The following steps are taken:

1. The target orientation image is selected manually, usually the image with the optic disk and fovea in the center.
2. RM is used to register all other images onto the target orientation image, resulting in their corresponding transformation matrix and grid offsets.
3. To accommodate images that are outside the target orientation image, they are padded with pixels (in the study, 600 pixels were added in each direction), and the transformation matrix is recalculated based on the grid offset estimation using the fast RANSAC estimation module.
4. The padded images are aligned to the padded target orientation image using the recalculated matrix.
5. The images are combined by taking the maximum value of the three channels at each pixel location among all images.

Automatic Registration Annotation Framework

To train the RDLR, all image files in the training data were grouped according to examination IDs. Each group's files were pairwise combined, resulting in a total of 13,261,064 pairs of image combinations. Then, with the help of the CRP, registration information was generated for these pairs of image combination data. The steps in the pipeline are as follows:

1. Resize each pair of images to 640×480 pixels using bilinear interpolation, which is a common method for up-sampling or down-sampling images that preserves smoothness.
2. Detect local landmark points in each image of the pair using the speeded up robust feature (SURF) algorithm, which is a robust feature detection and description method that extracts key points and computes their descriptors, enabling the identification of similar features across images.
3. Embed corresponding local feature information from each image pair, and then combine the landmark points and features. Calculate the Manhattan distance of the corresponding feature vectors and obtain the coarse matching coordinate pairs by sorting these distances, which provides an initial estimate of corresponding points.
4. Filter the coarse match pairs using the grid-based motion statistics (GMS) algorithm, which is an efficient method for rejecting outliers in feature matching by analyzing the consistency of motion vectors within a grid structure.
5. Use the random sample consensus (RANSAC) algorithm to estimate the robust match pairs and solve the affine transformation matrix based on the pairs.

However, manual inspection revealed a high incidence of errors in the data pairs, primarily due to the limitations of the CRP's registration performance. Consequently, we introduced a quality control module aimed at refining the registration outcomes. This module analyzed the registration results, selectively retaining reliable matches while discarding unreliable ones. The process utilized robust landmark point pairs, which were rigorously filtered through the RANSAC and GMS algorithms, ensuring the integrity and accuracy of annotated dataset.

1. Filter the robust match pairs based on their coordinating information using GMS again to filter consistent pairing between point coordinates and their neighborhood point coordinates in both graphs.
2. Count and assess the quantity of consistent robust pairs based on quantity, with a quantity threshold of 30 (Fig. S3).

The filtering process yielded a subset of 968,886 high-quality labeled pairs of training (Table 1). This pipeline was validated using both internal and external validation datasets, with over 99% of the filtered pairs achieving a Dice score greater than 0.9, indicating its efficacy (Fig. S3). Furthermore, to assess the generalization ability of the model trained on the quality-controlled data on low-quality test data, we conducted an evaluation using specifically low-quality data in a subsequent experiment section.

Dataset

Data Preparation

We collected a comprehensive dataset of 280,107 pediatric fundus images from around 9781 examinations conducted at two hospitals in China: Zhongshan Ophthalmic Center, Sun Yat-sen University (ZOC), and Guangdong Women and Children Hospital at Panyu (PY) (Fig. S4). The ZOC dataset, collected from April, 2018, to March, 2022, was divided into training and validation subsets. The PY dataset, collected from January, 2015, to March, 2018, served as an independent external validation dataset (Table 1). All data were exported directly from the same device (RetCam [8, 9]) and saved as PNG (ZOC) and JPG (PY) image files, respectively. Each examination had a unique 36-bit machine code ID consisted of letters, numbers, and symbols. In addition to the above information, we have anonymized all the information for privacy protection.

Table 1 Description of the datasets

Datasets	Training set	Validation set	External set
Source	ZOC		PY
Examination date	2018.4–2021.12	2022.1–2022.3	2015.1–2018.3
No. of examination	9546	50	100
No. of image	275,644	1159	1416
No. of annotated pairs	968,886	3828	7133

Evaluation Annotation

We developed an annotation tool for labeling the internal and external validation datasets using CRP (Fig. S5). This tool allowed technicians to manually mark landmark points and simultaneously generate the registration results. Landmark points refer to the pairs of coordinates indicating the same spatial location in the two images to be registered. All validation data adhered to the following rules during the annotation process: 1) Priority was given to bifurcation and convex points formed by the bending of blood vessels; 2) in cases where vessels were straight, preference was given to the points where vessels connected to the optic disc; 3) points that were widely distributed across the image were selected whenever possible; 4) points with unclear textures or artifacts were avoided; 5) a minimum of three pairs of points were required. Technicians were able to preview the registration results and added more landmark points until precision was achieved.

To evaluate the performance of proposed method, we randomly selected 50 examinations from the ZOC dataset as the internal validation set and 100 examinations from the PY dataset as the external validation set. The datasets were annotated by a clinical technician, and the annotations were subsequently reviewed by a clinical retinal expert to ensure accuracy. The annotation process involved manually correcting the joining positions of stitched image pairs, a task performed by human. This manual annotation was crucial for the creation of a reliable ground truth for registration accuracy assessment. Ultimately, the same clinical technician annotated 3828 stitched pairs for the internal (ZOC) dataset and 7133 stitched pairs for the external (PY) dataset, with annotations reviewed and verified by the same clinical expert.

For panoramic view annotation, each view was composed of multiple image pairs. The image with the highest pairing connectivity to other images in each examination was selected as the starting target orientation and then other orientation images were paired with the starting target orientation image. After manual annotation of all the pairs, the panoramic view could be assembled based on the pair annotations. Similarly, the panoramic view annotations were generated for the internal and external validation datasets.

Metrics

Registration Accuracy of Pairs

We compared the overlap of the predicted mask and the annotation mask of two images, A and B, in terms of scale and orientation by using the Sørensen–Dice coefficient (Dice score), given by

$$\text{Dice score} = \frac{2 * (\text{pred} \cap \text{true})}{\text{pred} \cup \text{true}} \tag{14}$$

where pred is the estimation mask generated by RDLR's result and true is the annotation mask (Fig. 2a). Additionally, to evaluate the precision of local content registration, we selected nine landmark points using a uniform distribution and calculated the mean square error (MSE) between the annotated and predicted coordinates. The MSE reflects the accuracy of the method in capturing detailed content, given by

$$\text{pts}_p = \text{pts} * M_p$$

$$\text{pts}_g = \text{pts} * M_g$$

$$\text{Dist(MSE)} = \left|\text{pts}_p - \text{pts}_g\right|^2 \tag{15}$$

where pts is a matrix of 9 * 2 containing nine-point coordinates at fixed positions (Fig. S6a), and M_p and M_g are affine transformation matrices of 3 * 2 from RDLR prediction and annotation, respectively. pts_p and pts_g are the coordinate matrices after affine transformation.

Proportion of Acceptable Pairs

The acceptance rate of the registration methods was calculated based on the DICE score under various threshold conditions. Finally, a graph was plotted to show the relationship between the DICE threshold and the acceptance proportion.

Registration Accuracy of Panoramic View

We compared the overlap of the predicted mask and the annotation mask of the panoramic view of an examination in terms of scale and orientation by using DICE score.

Statistical Analysis

We evaluated all results using DICE, expressed as percentage (that is, 0–1.0) with 95% CIs computed by using of normal approximate method [59]. The MSE, expressed as percentage (that is, 0–∞) with 95% CIs were computed by sample method. The $P \leq 0.05$ was set to determine significance, and P values were two-sided test (T-test). All plots were generated using Python package Seaborn (v.0.11.2) and matplotlib (v.3.3.1). Python package Scipy (v.1.5.2) was used to compute P values of the t-test.

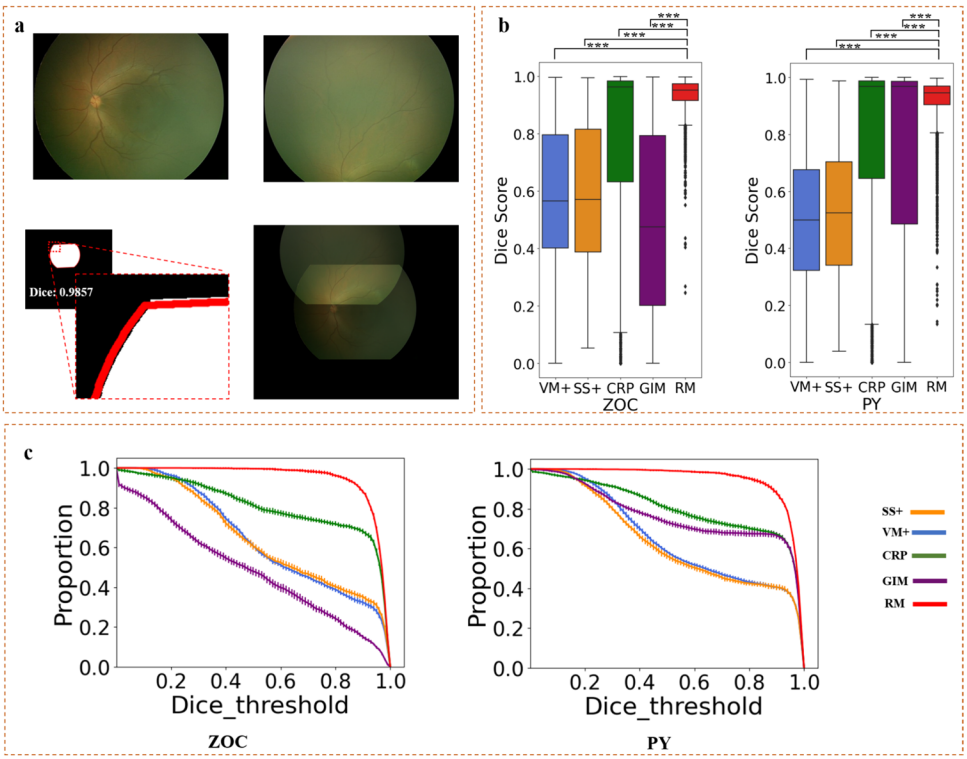

Fig. 2 Performance evaluation of the RM of RDLR in comparison analysis. **a** Dice score between the prediction (red) and the annotation information (white) for an annotated pair. The white region represents the target mask obtained from the annotated information, while the red contour indicates the contour of the mask predicted by the model. **b** Evaluation of the registration performance comparing VM+, SS+, CRP, GIM, and RM. Left: Dice score of the internal (ZOC) validation dataset. Right: Dice score of the external (PY) validation dataset (VM+, voxel morph with refinement; SS+, the RM trained with a self-supervised training strategy; CRP, the conventional registration pipeline consisting of multiple registration algorithms; GIM, generalizable image matcher; RM, the proposed registration module). **c** Evaluation of the registration performance with proportion of acceptable pairs. Left: Proportion of acceptable registrations in the internal (ZOC) validation dataset. Right: Proportion of acceptable registrations in the external (PY) validation dataset. (***: $P<0.001$. All error bars represent 95% confidence intervals computed using Bootstrap)

Implementation Details

Our proposed model was implemented using the PyTorch (1.11) in Python (3.9.12). It was trained on Nvidia Tesla V100 GPU with 32 GB memory * 6, and Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz. The inference of model was run on Nvidia RTX 3090 GPU with 24 GB memory, and Intel(R) i9-11900K @ 3.5 GHz.

To train the model, we designed a training strategy, the bidirectional supervised training (BS), which could accumulate the convergence speed of model. This strategy involves registering input images B to A and A to B, as illustrated in Fig. S6. In terms of train data, we randomly extracted 10,000 data for validation and set the number of rounds of model training based on the metrics. Ultimately, we trained the registration module for 13 epochs with a batch size of 48, and set the initial learning rate of as $1e-4$ and, then, decays the learning rate by 0.1 once in epoch 5 and 10. The AdamW algorithm was chosen as the optimizer function. During the process, from epoch 1 to 10, the registration module was trained with matrix estimation head with global information. After epoch 10, the matrix estimation head with local information was chosen to finetune model for 3 epochs. The entire training process took about 123 h (Fig. S6).

Experiments and Results

Comparison Analysis of RM

We conducted a comparative analysis among various registration methods, including the proposed registration module (RM), CRP, the deep learning registration model voxel morph (VM), the RM trained with a self-supervised (SS) training strategy, and the generalizable image matcher (GIM), using a total of 10,961 annotated pairs from both the internal (ZOC) and external (PY) validation datasets. It is important to note that both VM and SS were trained on the same dataset as RM. To ensure a fair comparison and to enhance the performance of VM and SS, we have

incorporated the refinement module proposed in our study into both VM and SS. The modified versions of these solutions were designated as SS+ and VM+, respectively. We compared the registration of the annotated and predicted labels, measured by Dice score (Fig. 2a). The distribution of Dice scores and the proportion of acceptable pairs over different Dice thresholds allowed us to gain insights into the precision and robustness of the registration methods.

Across the two validation datasets of internal (ZOC) and external (PY), the average Dice scores of RM were 0.948 (95% CI: 0.946, 0.950) and 0.947 (95% CI: 0.945, 0.949), respectively. In contrast, CRP achieved Dice scores of 0.802 (95% CI: 0.792, 0.811) and 0.806 (95% CI: 0.799, 0.812), respectively. The Dice scores of VM+ and SS+ were even lower, with values of only 0.648 (95% CI: 0.639, 0.656) and 0.651 (95% CI: 0.641, 0.661) in the internal (ZOC) dataset, and 0.646 (95% CI: 0.639, 0.653) and 0.643 (95% CI: 0.636, 0.650) in the external (PY) dataset, respectively. For GIM, there was a discrepancy in performance at the two validation datasets. The average Dice scores were 0.491 (95% CI:0.480, 0.501) and 0.778(95%CI: 0.771, 0.785), respectively. For both validation datasets, Dice scores of RM increased significantly higher than those of other methods ($P < 0.001$, t-test) (Fig. 2b and Table 2). As the RM, VM+, and SS+ models contain the refinement module proposed in our study, we also compared them without the refinement module (named as RM-, VM, SS). We observed that the refinement module consistently improved accuracy for other methods. Furthermore, even in the absence of the refinement module, the RM module still achieves a significantly higher level of performance compared to the other two methods (Fig. S7 and Table S1).

The box plot showed that RM exhibited high accumulations of Dice scores, indicating its stable performance across different images, even for low-quality ones. On the other hand, the box plots of the other three methods (CRP, VM+, SS+, and GIM) had scores spread over a wider range. Notably, the median of the box plot for CRP indicated that the local feature could ensure the maximum of the registration performance. However, the scores of VM+ and SS+ had similar distributions, implying that they were not well adapted to handle image registration tasks with incoherent features and substantial displacements (Fig. S8). In contrast, RM, which incorporates both global and local features in its registration model and refinement module, outperformed other methods in terms of both the highest accuracy and the stability of the distribution range of scores. This superiority is also evident from the density graph of the distribution (Fig. S9).

Furthermore, RM maintained nearly 100% acceptable pairs for Dice thresholds below 0.7 and started to decline at Dice threshold of around 0.85, while CRP decreased to 80% before Dice threshold of 0.5, demonstrating a more reliable performance of RM. The curves of GIM, VM+, and SS+ were considerably lower than those of the other methods, indicating generally poor scores for these two methods. However, above a Dice threshold of 0.9, the curve of RM (the percentage of acceptable pairs for the RM method) experienced a rapid decline, surpassing CRP, and CRP approached RM at a Dice threshold of 0.95. This suggests that when local features were reliable, CRP could provide acceptable results, although not as good as RM.

RM also outperformed GIM, CRP, VM+, and SS+ in terms of MSE by sampled points (Table 2 and Fig. S7). To summarize, RM combined the benefits of both local and global features and significantly outperformed the other methods over the entire Dice threshold range (Fig. 2c).

Ablation Analysis of RM

To determine the impact of local and global features, we conducted ablation analysis. We masked the corresponding of one of the input image pairs based on the overlapping and non-overlapping areas, of the annotated pairs, and evaluated the change in average Dice scores. Masking the overlapping region caused a decrease for all methods, but the decline was smaller in RM, VM+, and SS+, compared to CRP, indicating that CRP relied solely on overlapping features while other methods could use non-overlapping features to compensate for the loss of overlapping features (Fig. 3a).

Masking the non-overlapping region caused minimal changes in RM, but improved performance of CRP and SS+ (Fig. 3b), indicating that the loss of non-overlapping features made methods focus more on the overlapping

Table 2 Performance evaluation of the RM of RDLR in stitched pair

Datasets	Internal validation dataset		External validation dataset	
Metric	DICE (95% CI)	MSE (95% CI)	DICE (95% CI)	MSE (95% CI)
VM+ [29]	0.648 (0.639, 0.656)	169.72 (165.21, 174.23)	0.646 (0.639, 0.653)	172.95 (169.32, 176.58)
SS+ [41]	0.651 (0.641, 0.661)	163.54 (158.72, 168.37)	0.643 (0.636, 0.650)	170.26 (166.60, 173.91)
CRP [20, 47]	0.802 (0.792, 0.811)	94.55 (89.60, 99.51)	0.806 (0.799, 0.812)	93.53 (90.05, 97.01)
GIM [43]	0.491 (0.480, 0.501)	265.12 (258.07, 272.16)	0.778 (0.771, 0.785)	108.14 (104.41, 111.88)
RM	0.948 (0.946, 0.950)	20.16 (19.16, 21.16)	0.947 (0.945, 0.949)	21.36 (20.47, 22.26)

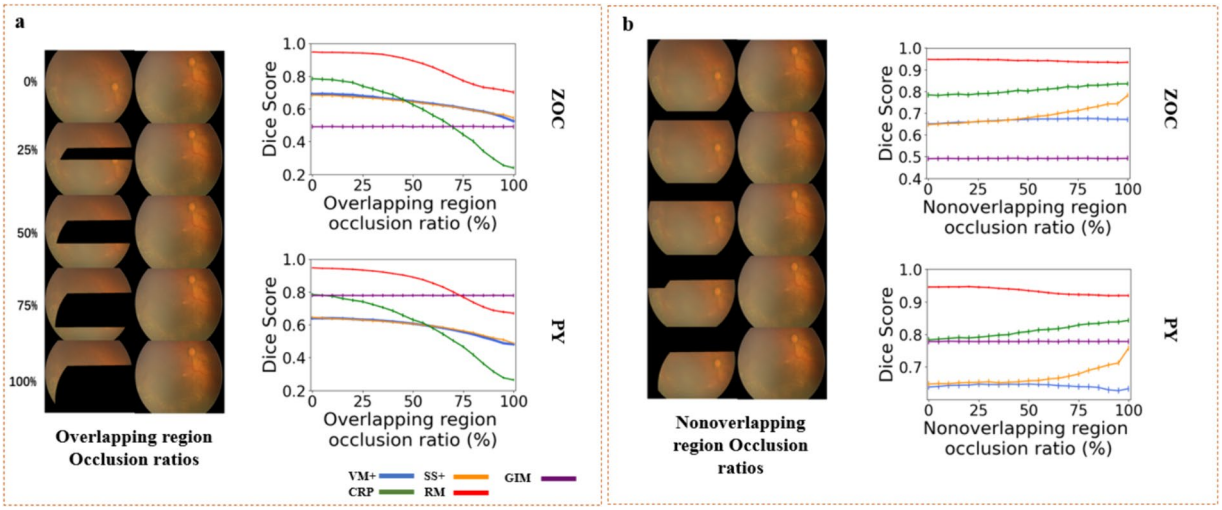

Fig. 3 Performance evaluation of the RM of RDLR in ablation analysis. **a** Ablation analysis of registration under various occlusion conditions with overlapping regions. **b** Ablation analysis of registration under various occlusion conditions with non-overlapping regions

region. This aligns with the design principle of the refinement module to perform local alignment with limited search areas determined by RM. The VM+'s score was also less affected. The possible reason is that the deformation field tends to distort the image as much as possible to maximize the overlap area of the registration result, even though it contradicts the actual situation. Moreover, the average Dice score of CRP was always lower than that of RM, indicating that considerable local features were unreliable, even with the confined search area, so that determination of the reliability of local features by the refinement module was necessary. Ablation analysis of MSE showed similar trends (Fig. S10a, b).

Comparison Analysis in Low-Quality Data

To evaluate the performance of RM of RDLR in low-quality images, we conducted an additional analysis of the CRP results. We sorted the samples based on their metric distribution in the two validation datasets and observed that the worst-performing subset predominantly consisted of low-quality samples (Fig. S11a, b). Subsequently, we extracted the top 25% and 12.5% of all worst-performing samples, respectively, to create a low-quality sample validation set. These sets were named according to the extraction ratio, i.e., internal set (4rd), external set (4rd), internal set (8rd), and external set (8rd). We then compared the performance of RM, CRP, GIM, VM+, and SS+ on these sets. The results clearly indicate that RM demonstrated superior performance across all validation sets (Tables 3 and 4). The box plots revealed that in case of low-quality sample, the performance metrics of the comparative methods were generally in a lower range, whereas RM significantly outperformed the other methods in terms of median, extremes, and the distribution range of the metrics (Figs. 4a, b and S11c).

Additionally, we conducted a comparative experiment simulating data quality degradation by testing the impact of down-sampling the pixel resolution of pediatric fundus images. In study, the down-sampling process involved reducing the image resolution and then reconstructing the original resolution image using linear interpolation. During compression and reconstruction, varying degrees of information loss occurred in the image, thereby simulating the degradation of image quality. We found

Table 3 Performance evaluation of the RM of RDLR in stitched pair with validation set (4rd)

Datasets	Internal set (4rd)		External set (4rd)	
Metric	DICE (95% CI)	MSE (95% CI)	DICE (95% CI)	MSE (95% CI)
VM+ [29]	0.406 (0.394, 0.419)	305.32 (298.72, 311.92)	0.362 (0.353, 0.371)	323.29 (317.96, 328.63)
SS+ [41]	0.382 (0.369, 0.395)	305.94 (299.35, 312.53)	0.376 (0.367, 0.385)	310.62 (305.78, 315.47)
CRP [20, 47]	0.349 (0.337, 0.361)	323.71 (316.41, 331.00)	0.351 (0.343, 0.359)	333.97 (328.14, 339.81)
GIM [43]	0.273 (0.257, 0.289)	437.33 (417.77, 456.88)	0.526 (0.511, 0.542)	242.90 (234.58, 251.23)
RM	0.882 (0.877, 0.888)	53.53 (50.85, 56.21)	0.859 (0.854, 0.864)	65.15 (62.64, 67.66)

Table 4 Performance evaluation of the RM of RDLR in stitched pair with validation set (8rd)

Datasets	Internal set (8rd)		External set (8rd)	
Metric	DICE (95% CI)	MSE (95% CI)	DICE (95% CI)	MSE (95% CI)
VM+ [29]	0.329 (0.315, 0.344)	346.46 (338.44, 354.49)	0.294 (0.283, 0.305)	363.49 (356.58, 370.39)
SS+ [41]	0.299 (0.284, 0.314)	347.17 (339.17, 355.17)	0.302 (0.292, 0.312)	351.77 (345.73, 357.80)
CRP [20, 47]	0.212 (0.200, 0.224)	407.99 (400.01, 415.96)	0.201 (0.193, 0.209)	428.20 (421.10, 435.30)
GIM [43]	0.210 (0.191, 0.230)	501.31 (471.12, 531.48)	0.416 (0.396, 0.437)	303.80 (292.60, 315.01)
RM	0.880 (0.872, 0.887)	55.48 (51.67, 59.29)	0.846 (0.838, 0.854)	70.54 (66.74, 74.34)

that RM were robust to down-sampling up to 25% and maintained their performance, while CRP showed a steady decline during down-sampling (Fig. 4c). While large down-samplings beyond 25% did lead to decrease in the performance of RM, they still outperformed CRP, highlighting their capacity to handle even extremely low-quality images. Conversely, VM+ and SS+ consistently performed poorly under all conditions. Ablation analysis of MSE revealed similar trends, further confirming the superiority of RM (Fig. S11d).

Comparison Analysis of PVM in Panoramic View

Benefiting from the panoramic view module (PVM), in collaboration with the registration module (RM), a comprehensive panoramic view of the fundus can be generated (furthermore, PVM can employ various registration methods to produce panoramic images). To evaluate the precision of the panoramic reconstruction, we compared it to manual annotation using Dice score (Fig. 5a and b). We also calculated the mean number of error pairs per examination across various Dice threshold levels (Fig. 5c).

We compared the performance of RM, CRP, GIM, VM+, and SS+. RM exhibited superior performance with average Dice scores of 0.960 (95% CI: 0.955, 0.965) and 0.963 (95% CI: 0.959, 0.967) for the internal (ZOC) and external (PY) datasets, respectively. In contrast, CRP achieved lower Dice scores of 0.766 (95% CI: 0.725, 0.807) and 0.792 (95% CI: 0.763, 0.821) for the two datasets, respectively. The Dice scores of GIM were 0.783 (95%CI: 0.751, 0.815) and 0.874 (95%CI: 0.858, 0.890) for two validation sets, respectively. Similarly, VM+ and SS+ achieved only 0.720 (95% CI: 0.699, 0.742) and 0.726 (95% CI: 0.703, 0.750) in internal (ZOC) and 0.736 (95% CI: 0.719, 0.753) and 0.756 (95% CI: 0.737, 0.775) in external (PY), respectively. Dice scores of RM significantly outperformed all the other methods on both validation datasets ($P < 0.001$, t-test) (Fig. 5b and Table 5). The comparison results of additional methods (included RM- vs. RM, VM vs. VM+, SS vs. SS+) can be found in the Supplementary Material (Table S2).

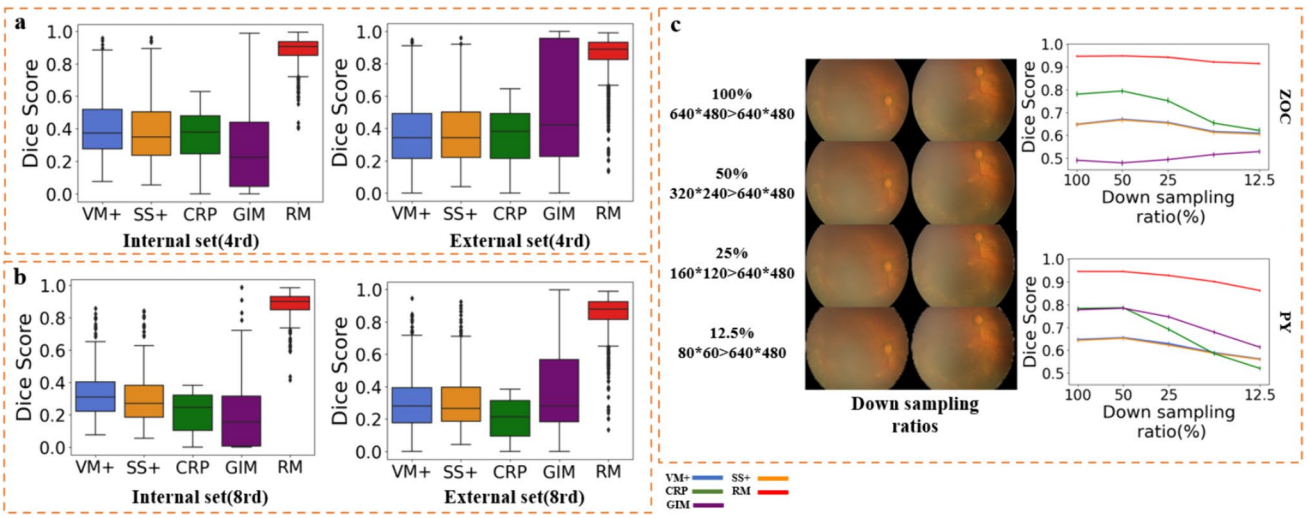

Fig. 4 Performance evaluation of the RM of RDLR in low-quality data. **a** Evaluation of the registration performance comparing VM+, SS+, CRP, GIM, and RM in validation datasets (4rd). Left: Dice score of the internal (ZOC) set (4rd). Right: Dice score of the external (PY) set (4rd). **b** Evaluation of the registration performance comparing VM+, SS+, CRP, GIM, and RM in validation datasets (8rd). Left: Dice score of the internal (ZOC) set (8rd). Right: Dice score of the external (PY) set (8rd). **c** Down-sampling analysis evaluating the model performance across different pixel resolutions and image qualities

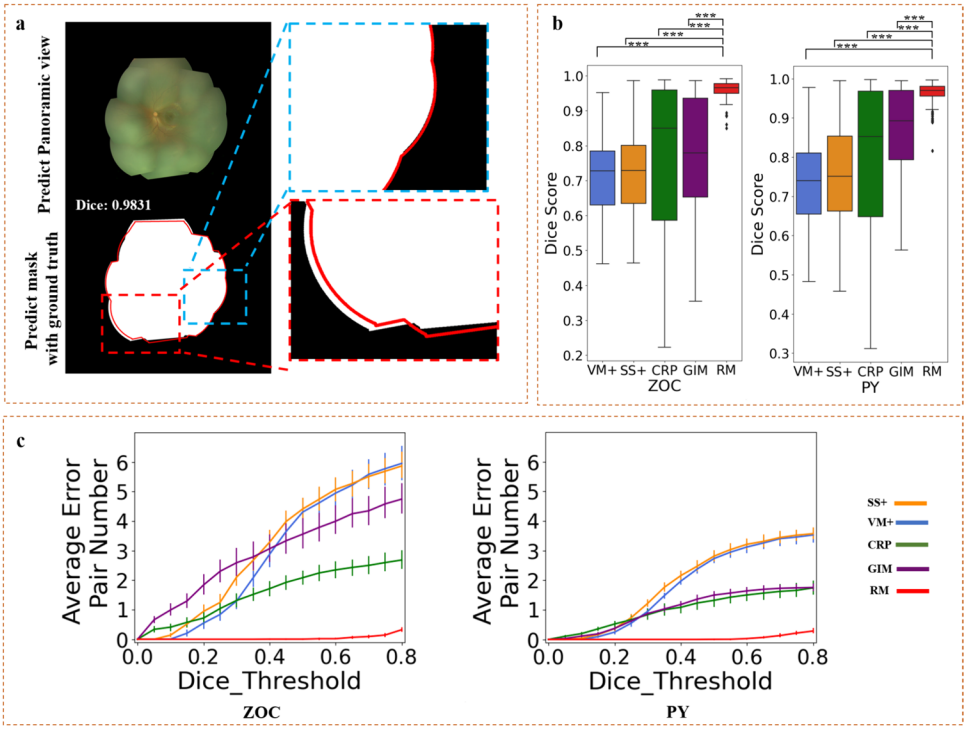

Fig. 5 Performance evaluation of the PVM of RDLR in comparison analysis. **a** Dice score between the prediction (red) and the annotation information (white) for the panoramic view. **b** Evaluation of the accuracy of panoramic view estimation on the internal (ZOC) and external (PY) validation datasets. Left: Dice score of the internal (ZOC) validation dataset. Right: Dice score of the external (PY) validation dataset. **c** Evaluation of the average number of error pairs for each panoramic view estimation on the internal (ZOC) and external (PY) validation datasets. Left: Average number of error pairs for each panoramic view in the internal (ZOC) validation dataset. Right: Average number of error pairs for each panoramic view in the external (PY) validation dataset. (***: $P < 0.001$. All error bars represent 95% confidence intervals computed using Bootstrap)

The box plot showed that CRP's unstable registration resulted in low Dice scores. Moreover, the error-dice plot revealed that CRP distorted images in nearly every examination, making it unsuitable for clinical use (Fig. 5c). GIM, VM+, and SS+ showed better lowest Dice scores than CRP, but they produced more distorted images per examination than CRP due to limited high Dice scores. In contrast, the panoramic view construction of RM achieved consistently high Dice scores for both the validation datasets (Figs. 5b, c and S12). This highlights RM's strong potential in facilitating disease diagnosis.

Comparison Analysis in Other Modality Data

To assess the applicability of RDLR beyond pediatric fundus images, we conducted tests on different types of images without requiring additional data or fine tuning, including fundus angiography (FA) of infants, conventional fundus photography (CFP) of adults, anterior segment (AS) images of infants, optical coherence tomography angiography (OCTA) images of adults, whole-slide pathology images (WSI) of rat brain, and four-dimensional computerized tomography (4D-CT) images of human chest. Additionally,

Table 5 Performance evaluation of the PVM of RDLR in panoramic view

Datasets	Internal validation dataset		External validation dataset	
Metrics	DICE (95% CI)	MSE (95% CI)	DICE (95% CI)	MSE (95% CI)
VM+ [29]	0.720 (0.699, 0.742)	179.72 (161.51, 197.83)	0.736 (0.719, 0.753)	175.06 (162.32, 187.80)
SS+ [41]	0.726 (0.703, 0.750)	183.22 (165.58, 200.85)	0.756 (0.737, 0.775)	174.29 (161.46, 187.11)
CRP [20, 47]	0.766 (0.725, 0.807)	98.57 (79.13, 118.0)	0.792 (0.763,0.821)	90.83 (78.64, 103.02)
GIM [43]	0.779 (0.747, 0.810)	199.34 (168.58, 230.11)	0.871 (0.855, 0.887)	109.98 (96.28, 123.68)
RM	0.954 (0.945, 0.962)	21.57 (17.58, 25.57)	0.963 (0.959,0.967)	20.70 (18.04, 23.36)

we explored cross-modal registration between FA and fundus photograph (FP) of infants. Our results demonstrated RDLR's capability for registration across different imaging modalities (Fig. S13, and Tables S3 and S4). Given that VM and SS necessitate additional data for training and adaptation to specific modalities, we only compared RDLR to the general method: GIM and CRP. These findings underscore the versatility and potential of RDLR in facilitating clinical diagnosis and research across diverse imaging modalities.

Discussion

In this study, we introduced the RDLR to accurately register image pairs and generate panoramic views for pediatric fundus images. Our experiments demonstrate the superior performance of RDLR in terms of registration accuracy and stability, outperforming other methods on both internal and external validation datasets of registration task. Additionally, the RDLR can be completed within 1 s in our test environment, suggesting potential for real-time application during data acquisition while maintaining clinically acceptable levels of robustness and accuracy. This could allow immediate update to panoramic images and provide timely feedback to clinicians on the progress of image collection.

Through the comparison of self-supervised training strategies (SS), unsupervised training strategies (VM), and RDLR in our study, we found that models trained with annotation information generated from real data are more stable. The proposed self-annotation framework not only addresses the lack of training data annotation but also provides reliable supervision, which is essential for RDLR's high registration accuracy. By comparing the training curves of the unidirectional registration training strategy (US) and the bidirectional registration training strategy (BS), we observed that BS strategy enhances the model's generalization capability under the same number of training iterations (Fig. S6). The strategy completes forward and backward predictions in a single iteration, with dual supervision ensuring prediction accuracy and satisfying the transformation relationship between directions, effectively adding new constraints and improving the model's training ceiling. The introduction of the refinement module further enhances the accuracy of RDLR. Comparisons between RM-, RM, VM, VM+, SS, and SS+ in terms of registration pair and panoramic registration effects clearly demonstrate that the refinement module improves the precision of these methods. With a sufficiently accurate initial registration, the refinement module can maximize the precision.

The ablation experiment on occluded regions showcased the stability of RDLR and its ability to efficiently utilize features. Even when all overlapping informative region (local information) was obscured, RDLR was still able to predict based on global information outside the overlap, demonstrating the participation of global information in registration predictions. The performance degradation observed after occluding non-overlapping region further supported this finding. In the low-quality data comparison experiment, validation datasets were created by selecting poorly performing samples based on metric distributions, and additional low-quality data were simulated through down sampling. RDLR's result on both real and simulated low-quality data showed significant superiority over other methods, attributable to the data-driven supervised training and the model's effective feature utilization.

Focusing on pediatric fundus images, the study also demonstrated the extensibility of the approach through experiments on other medical modalities images. RDLR, trained on single-modality data without additional fine tuning, outperformed other methods, indicating its ability to focus on universal image features rather than relying solely on modality-specific characteristics (e.g., color).

However, RDLR does have limitations. Firstly, the registration model in RM was based on assumption of affine transformation, which may not fully accommodate the slight distortions that can occur at the edges of real images. This can lead to imperfect alignment of both central and peripheral features in certain extreme cases. Additionally, the iterative registration process of PVM relies on a central orientation as the starting point, with other orientations aligned toward it. If a peripheral orientation is used as the starting point, there is a risk that other orientations may lack overlapping regions for alignment, resulting in erroneous registration. Finally, the threshold selection of quality control module for the automatic annotation module in the study could be more refined and simultaneously evaluate the accuracy trends of RM under different train data under different thresholds, which will lead to a more reliable quality control threshold in subsequent work.

Future research should focus on several key areas. Improving alignment techniques to better account for both central and peripheral features is one area, which could involve techniques such as local affine transformations, perspective transformations, or image distortion correction. Another area is to refine the PVM's iterative process by implementing multi-level iterations, dynamically specifying the starting orientation at each level to avoid misregistration due to the absence of overlapping regions. Lastly, developing a scoring system to evaluate registration results could help in identifying and filtering out clearly erroneous registrations. This scoring system and the RM can be combined and integrated into the self-annotation process, using data generated by the older RM to iteratively refine and improve the RM.

Conclusion

The proposed RDLR method, which integrated the registration module (RM) for precise registration and the panoramic view module (PVM) for panoramic image generation enhance the utility of imaging data, achieved exceptional performance in pediatric retinal image registration. It addressed the scarcity of training data for deep learning through automatic annotation. The RM model, trained on real-world data, effectively extracts reliable local and global semantic feature information, reducing noise interference and enabling accurate registration predictions even with limited overlapping regions. The refinement module optimized results by utilizing local information, improving registration accuracy. Extensive experimental evaluation consistently demonstrated RDLR's superior registration accuracy and reliability, making it a promising solution for clinical and research application across diverse imaging modalities.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-024-01154-2.

Acknowledgements We thank the State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, and the Center for Precision Medicine, Sun Yat-sen University, for the long-term support.

Author Contribution All authors contributed to the study conception and design. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability Data will be made available on request.

Code Availability The source code of RDLR in this study can be found on GitHub (https://github.com/wuwusky/RobustDeepLeraningRegistration) and the comparison methods are publicly available on GitHub (VM, https://github.com/voxelmorph/voxelmorph; GIM, https://github.com/xuelunshen/gim); other methods, such as CRP and SS, were implemented based on OpenCV (https://github.com/opencv/opencv) and Pytorch (https://github.com/pytorch/pytorch).

Declarations

Ethical Approval Ethical approval for research used was obtained from each center (the ethics committee of Zhongshan Ophthalmic Center, Sun Yat-sen University (2020KYPJ175), and the ethics committee of Guangdong Women and Children Hospital (202201057)).

Consent to Participate Informed consent was obtained from all individual participants included in the study. Written informed consent was obtained from the parents.

Consent for Publication The authors affirm that human research participants provided informed consent for publication of the images in all figures and tables.

Competing Interests The authors declare no competing interests.

References

1. Ripley, D.L. & Politzer, T. Vision disturbance after TBI. *NeuroRehabilitation* **27**, 215-216 (2010).
2. Fox, S.M., Koons, P. & Dang, S.H. Vision Rehabilitation After Traumatic Brain Injury. *Phys Med Rehabil Clin N Am* **30**, 171-188 (2019).
3. Blindness, G.B.D., Vision Impairment, C. & Vision Loss Expert Group of the Global Burden of Disease, S. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. *Lancet Glob Health* **9**, e130-e143 (2021).
4. Liu, D., Zheng, J. & Lu, Y. Fundus Examination of 23,861 Newborns by Digital Imaging in Ningbo. *J Ophthalmol* **2021**, 6620412 (2021).
5. Gundlach, B.S., *et al.* Real-world visual outcomes of laser and anti-VEGF treatments for retinopathy of prematurity. *Am J Ophthalmol* (2021).
6. Munson, M.C., *et al.* Autonomous early detection of eye disease in childhood photographs. *Sci Adv* **5**, eaax6363 (2019).
7. Yan, H.X., *et al.* Analysis of fundus examination results in 8 808 pediatric patients in Northwest China. *Zhonghua Yan Ke Za Zhi* **57**, 777-783 (2021).
8. RetCam - a useful adjunctive tool to evaluate and manage paediatric glaucomas. *Asian Journal of Ophthalmology* (2008).
9. Park, J.W., Park, S.W. & Heo, H. RetCam image analysis of the optic disc in premature infants. *Eye (Lond)* **27**, 1137-1141 (2013).
10. Vinekar, A., *et al.* Universal ocular screening of 1021 term infants using wide-field digital imaging in a single public hospital in India - a pilot study. *Acta Ophthalmol* **93**, e372-376 (2015).
11. Mayro, E.L., Wang, M., Elze, T. & Pasquale, L.R. The impact of artificial intelligence in the diagnosis and management of glaucoma. *Eye (Lond)* **34**, 1-11 (2020).
12. L PINELLO, M.M. Use of wide field digital retinal imaging (RET CAM II) in paediatric retinal diseases. *Acta Ophthalmol* **90**(2012).
13. Karp, K.A., *et al.* Training retinal imagers for retinopathy of prematurity (ROP) screening. *J AAPOS* **20**, 214-219 (2016).
14. Fielder, A.R., *et al.* Describing Retinopathy of Prematurity: Current Limitations and New Challenges. *Ophthalmology* **126**, 652-654 (2019).
15. Chiang, M.F., *et al.* International Classification of Retinopathy of Prematurity, Third Edition. *Ophthalmology* **128**, e51-e68 (2021).
16. Campbell, J.P., *et al.* Artificial Intelligence for Retinopathy of Prematurity: Validation of a Vascular Severity Scale against International Expert Diagnosis. *Ophthalmology* (2022).
17. Gschliesser, A., *et al.* Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. *Am J Ophthalmol* **160**, 553–560 e553 (2015).
18. Bolon-Canedo, V., *et al.* Dealing with inter-expert variability in retinopathy of prematurity: A machine learning approach. *Comput Methods Programs Biomed* **122**, 1-15 (2015).
19. Chan-Ling, T., Gole, G.A., Quinn, G.E., Adamson, S.J. & Darlow, B.A. Pathophysiology, screening and treatment of ROP: A multidisciplinary perspective. *Prog Retin Eye Res* **62**, 77-119 (2018).
20. Herbert Bay, T.T., and Luc Van Gool. SURF: Speeded Up Robust Features. (2006).
21. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding* **110**(2008).
22. Herrmann, C., *et al.* Robust Image Stitching with Multiple Registrations. *Computer Vision - Eccv 2018, Pt Ii* **11206**, 53–69 (2018).
23. Schlagenhauf, T., Brander, T. & Fleischer, J. A stitching algorithm for automated surface inspection of rotationally symmetric components. *Cirp Journal of Manufacturing Science and Technology* **35**, 169-177 (2021).
24. Kerkech, M., Hafiane, A. & Canals, R. Vine disease detection in UAV multispectral images using optimized image registration

and deep learning segmentation approach. *Comput Electron Agr* **174**(2020).
25. MiroslavTrajković, M. Fast corner detection. *Image and Vision Computing* **16**(1998).
26. Parkhomenko, P.M.A. Affine Transformations. *Euclidean and Affine Transformations* (1965).
27. Fischler, M.A. & Bolles, R.C. Random sample consensus. *Communications of the ACM* **24**, 381-395 (1981).
28. Ting, D.S.W.*, et al.* Deep learning in ophthalmology: The technical and clinical considerations. *Prog Retin Eye Res* **72**, 100759 (2019).
29. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J. & Dalca, A.V. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Trans Med Imaging* (2019).
30. Li, Y., Chen, C., Yang, F. & Huang, J. Hierarchical Sparse Representation for Robust Image Registration. *IEEE Trans Pattern Anal Mach Intell* **40**, 2151-2164 (2018).
31. Chang, X., Du, S., Li, Y. & Fang, S. A Coarse-to-Fine Geometric Scale-Invariant Feature Transform for Large Size High Resolution Satellite Image Registration. *Sensors (Basel)* **18**(2018).
32. Liu, R.*, et al.* Learning Deformable Image Registration from Optimization: Perspective, Modules, Bilevel Training and Beyond. *IEEE Trans Pattern Anal Mach Intell* **PP**(2021).
33. Qu, L.*, et al.* Cross-modal coherent registration of whole mouse brains. *Nat Methods* **19**, 111-118 (2022).
34. Li, A., Guo, J. & Guo, Y. Image Stitching Based on Semantic Planar Region Consensus. *IEEE Trans Image Process* **30**, 5545-5558 (2021).
35. Haskins, G., Kruger, U. & Yan, P. Deep learning in medical image registration: a survey. *Machine Vision and Applications* **31**(2020).
36. Abbasi, S.*, et al.* Medical image registration using unsupervised deep neural network: A scoping literature review. *Biomedical Signal Processing and Control* **73**(2022).
37. Han, R.*, et al.* Deformable MR-CT image registration using an unsupervised, dual-channel network for neurosurgical guidance. *Med Image Anal* **75**, 102292 (2022).
38. Nie, L., Lin, C., Liao, K., Liu, S. & Zhao, Y. Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images. *IEEE Trans Image Process* **PP**(2021).
39. Hering, A.*, et al.* CNN-based lung CT registration with multiple anatomical constraints. *Med Image Anal* **72**, 102139 (2021).
40. Fan, J., Cao, X., Yap, P.T. & Shen, D. BIRNet: Brain image registration using dual-supervised fully convolutional networks. *Med Image Anal* **54**, 193-206 (2019).
41. Nie, L., Lin, C., Liao, K., Liu, M. & Zhao, Y. A view-free image stitching network based on global homography. *Journal of Visual Communication and Image Representation* **73**(2020).
42. Krishnan, R., Rajpurkar, P. & Topol, E.J. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng* (2022).
43. Shen, X.*, et al.* GIM: Learning Generalizable Image Matcher From Internet Videos. in *The Twelfth International Conference on Learning Representations* (2023).
44. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* **60**, 91-110 (2004).
45. Rublee, E., Rabaud, V., Konolige, K. & Bradski, G.R. ORB: an efficient alternative to SIFT or SURF. in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6–13, 2011* (2011).
46. Zhang, H., Jia, N., Zhuo, K. & Zhao, W. Retinal fundus image registration framework using Bayesian integration and asymmetric Gaussian mixture model. *International Journal of Imaging Systems and Technology* **33**, 403-418 (2023).
47. Bian, J.*, et al.* GMS: Grid-Based Motion Statistics for Fast, Ultra-Robust Feature Correspondence. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2828–2837 (2017).
48. Yi, K.M., Trulls, E., Lepetit, V. & Fua, P. LIFT: Learned Invariant Feature Transform. *Springer International Publishing* (2016).
49. Rivas-Villar, D., Hervella, Á.S., Rouco, J. & Novo, J. Joint keypoint detection and description network for color fundus image registration. *Quantitative Imaging in Medicine and Surgery* **13**, 4540 (2023).
50. Xu, J.*, et al.* Reliable and stable fundus image registration based on brain-inspired spatially-varying adaptive pyramid context aggregation network. *Frontiers in Neuroscience* **16**, 1117134 (2023).
51. Kim, J.*, et al.* Fundus Image Translation with Scale-Aware Registration and Gradient-Guided GAN. *Available at SSRN 4700915*.
52. Rivas-Villar, D., Hervella, Á.S., Rouco, J. & Novo, J. Color fundus image registration using a learning-based domain-specific landmark detection methodology. *Computers in Biology and Medicine* **140**(2022).
53. Mok, T.C.W. & Chung, A.C.S. Fast Symmetric Diffeomorphic Image Registration with Convolutional Neural Networks. in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020).
54. Martínez-Río, J., Carmona, E.J., Cancelas, D., Novo, J. & Ortega, M. Deformable registration of multimodal retinal images using a weakly supervised deep learning approach. *Neural Computing and Applications* **35**, 14779-14797 (2023).
55. Nguyen, T., Chen, S.W., Shivakumar, S.S., Taylor, C.J. & Kumar, V. Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model. *Ieee Robotics and Automation Letters* **3**, 2346-2353 (2018).
56. Nianjin Ye, C.W., Shuaicheng Liu, Lanpeng Jia, Jue Wang, Yongqing Cui. DeepMeshFlow: Content Adaptive Mesh Deformation for Robust Image Registration. *arXiv* (2019).
57. Ochoa-Astorga, J.E., Wang, L., Du, W. & Peng, Y. A Straightforward Bifurcation Pattern-Based Fundus Image Registration Method. *Sensors* **23**, 7809 (2023).
58. Wang, C.-Y.*, et al.* MEMO: dataset and methods for robust multimodal retinal image registration with large or small vessel density differences. *Biomedical Optics Express* **15**, 3457-3479 (2024).
59. Land, C.E. An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics* **14**, 145-158 (1972).