


REVIEW

Open Access



Medical image registration and its application in retinal images: a review

Qiushi Nie¹, Xiaoqing Zhang^{1,2}, Yan Hu¹, Mingdao Gong¹ and Jiang Liu^{1,3,4*} 

Abstract

Medical image registration is vital for disease diagnosis and treatment with its ability to merge diverse information of images, which may be captured under different times, angles, or modalities. Although several surveys have reviewed the development of medical image registration, they have not systematically summarized the existing medical image registration methods. To this end, a comprehensive review of these methods is provided from traditional and deep-learning-based perspectives, aiming to help audiences quickly understand the development of medical image registration. In particular, we review recent advances in retinal image registration, which has not attracted much attention. In addition, current challenges in retinal image registration are discussed and insights and prospects for future research provided.

Keywords Computer-aided diagnosis, Medical image registration, Deep learning, Generative model, Transformer, Retina

Introduction

Medical image registration is a fundamental step in computer-aided diagnosis (CAD) and image-guided surgical treatment and has attracted much attention. It aligns multiple medical images by finding appropriate spatial transformation relationships to fuse their corresponding information, helping doctors make a more comprehensive and precise diagnostic conclusion. These medical images may be acquired at different times, angles, and even modalities for a certain tissue or organ of the human body. Therefore, the purpose of medical image

registration is to eliminate the interference of these factors and find consistent objects or shapes for matching.

Numerous methods have been developed to address the different transformation tasks in medical image registration. These can be grouped into two types: coarse-grained global linear registration and fine-grained local elastic registration. Coarse-grained global linear registration extracts the salient features of the input image pair, thereby matching these features and overcoming angular changes. Fine-grained local elastic registration performs pixel-level analysis of the input image pair after linear alignment and local corrections to overcome spontaneous tissue movements and deformations.

Another method to classify registration methods is based on what is used to match the images. The first and direct approach is an intensity-based method [1]. These methods consider registration as an optimization problem by iteratively disturbing the transformation parameters to maximize pixel-wise similarity. Another early but still popular approach is feature-based methods [2], which extract manually designed features and descriptors, match them, and establish a transformation based on matching. In contrast to intensity-based methods,

*Correspondence:

Jiang Liu
liuj@sustech.edu.cn

¹ Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

² Center for High Performance Computing and Shenzhen Key Laboratory of Intelligent Bioinformatics, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

³ Singapore Eye Research Institute, Singapore 169856, Singapore

⁴ State Key Laboratory of Ophthalmology, Optometry and Visual Science, Eye Hospital, Wenzhou Medical University, Wenzhou 325027, China

feature-based methods provide more robust registration by matching salient features rather than simply comparing pixels.

In the past decade, deep features have replaced hand-craft features with their ability to provide learnable, and therefore, more flexible, problem-specific feature representations for registration tasks. Later, after deep feature extractors, end-to-end registration neural networks integrated the entire registration process into a single network by applying deep learning techniques such as convolutional neural networks (CNNs), generative adversarial networks (GANs), and transformers. Once trained, these methods can obtain registration results directly from input image pairs, thereby speeding up registration. They have also been proven to have better registration performance.

Several reviews on deep learning for medical image registration have been conducted [3–5]. However, those studies only investigated the popular CNN-based methods at the time and did not mention the latest transformer-based methods. Additionally, those studies only investigated methods based on deep learning but ignored traditional methods from the early years, which can also provide significant guidance.

Among medical images, retinal images focus on a unique part of the human body that allows noninvasive observation of blood vessels *in vivo*. This noninvasive approach allows the capture of high-quality images, which facilitates the examination of the retina with minimal discomfort to patients. Longitudinal studies critical for monitoring disease progression often use a series of images captured at various time intervals. The diagnosis of retinal diseases such as age-related macular degeneration (AMD) is further facilitated by the availability of multiple imaging modalities, each serving a distinct diagnostic purpose. AMD, characterized by choroidal neovascularization (CNV), exemplifies the need for a multimodal approach: (1) color fundus (CF) photography effectively highlights areas of hemorrhage and the presence of fibrovascular tissue; (2) fluorescein angiography (FA) reveals subtle leaks associated with CNV that are not always visible to the naked eye; and (3) optical coherence tomography (OCT) provides detailed cross-sectional scans that can uncover intraretinal abnormalities. These modalities collectively assist ophthalmologists in diagnosing retinopathies and formulating strategies for ophthalmic surgery [6]. Moreover, retinal analysis is relevant not only to eye diseases, but also to various human diseases, including diabetes [7], Alzheimer's disease [8], and coronary heart disease [9]. Therefore, the retina serves as a microcosm for broader health assessments, providing a noninvasive yet informative window into a patient's overall well-being. Retinal image registration,

which combines complementary structural and functional information from the same or different modalities, is a crucial step in this process. Due to the particularity of the way retinal images are collected, they are mainly affected by three factors: illumination differences, angle differences, and variations in retinal lesions. These factors pose multiple technical challenges in the registration of retinal images: (1) Ensuring consistency in pixel values by standardizing or normalizing lighting conditions; (2) Identifying correspondences over long distances; (3) Tracking and quantifying the progression of retinal lesions.

However, in recent years, few studies have systematically reviewed retinal image registration. Although reviews have been conducted on related topics, such as retinal disease classification [10] and segmentation [11], the specific area of retinal image registration has not been thoroughly explored. Saha et al. [12], and Pan and Chen [13] addressed retinal image registration; however, they focused on a single retinal modality and did not perform a comparative analysis with mainstream medical image registration techniques. Therefore, the purpose of this paper is to review and summarize existing medical image registration works using traditional and deep learning-based methods, aiming to help audiences grasp the development of medical image registration. Moreover, retinal image registration are also surveyed and synthesized as a characteristic of this review. Finally, the current challenges in retinal image registration are also highlighted and future research directions discussed.

An initial literature search was performed using free-text searches in PubMed and Google Scholar. Papers that included the search term Medical Image Registration were considered and the publishing conference or journal and citations checked to ensure the quality of the research. Later, another search was performed using the search term Retinal Image Registration and all related papers considered. In the analysis, different temporal scopes were adopted for traditional and contemporary methods. For traditional methods the search was extended to encompass the last two decades, whereas for deep learning-based methods, the focus was narrowed to the most recent ten-year period to capture the latest advancements. Finally, the search space was iteratively increased by examining the bibliographies of the relevant papers.

The overall organization is illustrated in Fig. 1: **Background** section defines the basic concepts of image registration and briefly introduces the popular retinal image modalities. **Traditional registration methods** and **Deep learning-based registration methods** sections review the general methodology of medical image registration categorized as traditional and deep learning, respectively.

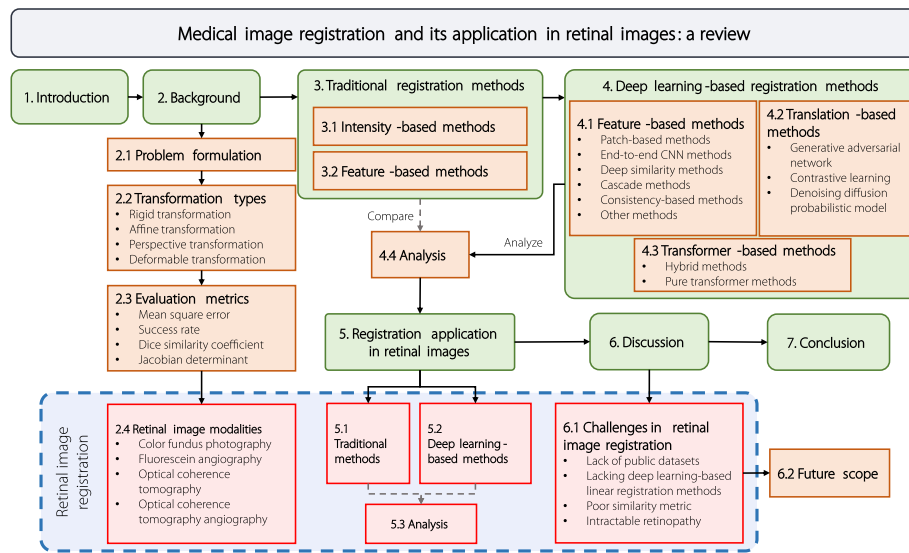


Fig. 1 Structure of the review

Registration application in retinal images section reviews the applications in retinal image registration, and compares them with the general methodology. **Discussion** section discusses the advantages and disadvantages of the reviewed methods, highlights the current challenges, and provides potential future research directions. Finally, in **Conclusion** section, the paper is summarized.

Background

Problem formulation

Image registration is a fundamental task in image processing. This involves finding correspondences between two images, namely, a moving and fixed image, and establishing a transformation between them. A fixed image is used as a reference, and the goal is to transform the moving image to match the fixed image. Registration algorithms are designed to determine the best transformation, denoted by T^* , that maximizes the similarity between two images [14]. This can be achieved by maximizing the image similarity function $\text{sim}(I_f, T(I_m))$, where I_m and I_f are the moving and fixed images, respectively, and $T(I_m)$ is the moving image transformed using the transform T .

Transformation types

This subsection introduces different transformation models, including rigid, affine, perspective, and deformable. Rigid, affine, and perspective transformations are linear, whereas deformable transformations are nonlinear. Figure 2 visually demonstrates their effects.

Rigid transformation consists of translation and rotation and preserves the original image’s size and shape. It is represented as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \mathbf{R} \begin{bmatrix} x \\ y \end{bmatrix} + \mathbf{t} \tag{1}$$

Here, (x, y) and (x', y') denote the original and transformed pixel coordinates, $\mathbf{R} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$ is the rotation matrix, and $\mathbf{t} = [t_x, t_y]^T$ is the translation vector.

Affine transformation combines translation, rotation, scaling, and shearing, offering more flexibility than rigid transformation. Affine transformation preserves straight lines and parallelism, but is not perpendicular. It is represented as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \mathbf{t} \tag{2}$$

Perspective transformation, or projective transformation, corrects perspective distortions, such as foreshortening and skew, between images. Perspective transformation maintains straightness but not parallelism or perpendicularity. This is represented in homogeneous coordinates as follows:

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = \begin{bmatrix} A & B & C \\ D & E & F \\ a & b & c \end{bmatrix} \begin{bmatrix} x \\ y \\ w \end{bmatrix} \tag{3}$$

Here, (x, y, w) is the homogeneous coordinate of the image to be transformed, (x', y', w') is the target coordinate in the transformed image. By setting $w = 1$ and

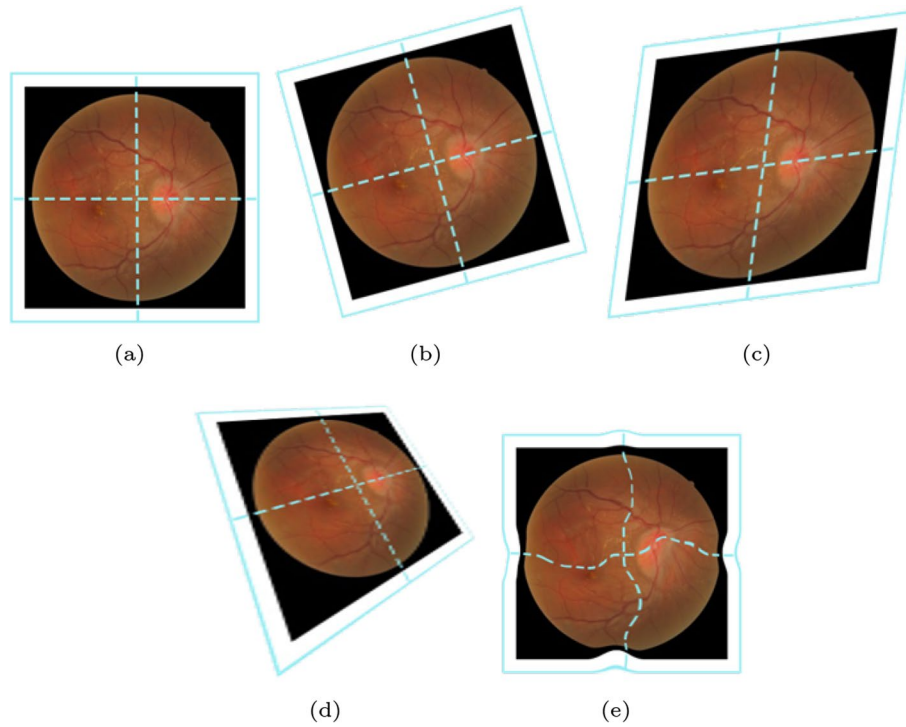


Fig. 2 Effect of different transformations. **a** Origin; **b** Rigid; **c** Affine; **d** Perspective; **e** Deformable

transforming the target $w' = 1$, the target point (x', y') is obtained in Cartesian coordinates:

$$\begin{aligned} x' &= \frac{Ax+By+C}{ax+by+c} \\ y' &= \frac{Dx+Ey+F}{ax+by+c} \end{aligned} \tag{4}$$

Deformable transformation allows nonlinear deformation, better adapting to shape variations compared to rigid or affine methods. It is represented as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \phi[x, y] \tag{5}$$

Here, ϕ represents the deformation field and $\phi[x, y]$ represents the transformation vector $(\Delta x, \Delta y)$ at (x, y) .

Evaluation metrics

Reliable evaluation metrics are crucial for assessing the medical image registration performance and guiding the design of new algorithms. Here, a brief review of four popular evaluation metrics is provided.

Mean square error (MSE) and root mean square error (RMSE) are standard metrics for measuring the quality of image registration. MSE can be calculated as

$$MSE = \frac{1}{N} \sum_{i=1}^N (I - J)^2 \tag{6}$$

RMSE simply adds an extra step to the square root based on MSE, and I and J have different meanings.

1. images: serves as image similarity measurement between warped moving image I'_M and fixed image I_F .
2. point pairs: serves as the distance measurement between corresponding point pairs.
3. transformations: serves as the difference between ground truth transformation and predicted transformation.

Success rate (SR) quantifies the proportion of successful registrations out of the total number of registration samples. It can be mathematically expressed as

$$SR = \frac{N_{\text{success}}}{N_{\text{total}}} \times 100\% \tag{7}$$

where N_{success} is the number of successful registrations and N_{total} is the total number of registration samples. However, the definition of *success* varies among studies, using different criteria or thresholds. The most



Fig. 3 Fundus photography examples using different imaging techniques. **a** CF from FIRE dataset [24]; **b** FA from CF-FA dataset [25]; **c** OCT from ref. [26]; **d** OCTA from OCTA-500 dataset [27]

commonly used criterion is the MSE between the predicted and corresponding ground truth points.

Dice similarity coefficient (DSC) quantifies the spatial overlap between two segmentations. For registration, Dice is calculated between the segmentation maps of the fixed image and the warped moving image to evaluate the overlap of the anatomical structures, which can be mathematically expressed as

$$\text{DSC} = 2 \times \frac{|S_F \cap (S_M \circ \phi)|}{|S_F| + |S_M \circ \phi|} \quad (8)$$

where S_F and S_M are the segmentations of I_M and I_F , respectively, and $S_M \circ \phi$ represents the warped segmentation of the moving image using the transformation ϕ .

Jacobian determinant quantifies the physical plausibility and invertibility of deformations by measuring how each pixel (or voxel if 3D) changes after the application of a certain deformation field. When the Jacobian determinant is non-positive, the deformation is not diffeomorphic. The percentage of pixels (or voxels) with non-positive Jacobian determinants ($|J_\phi| \leq 0$) is always used, and the Jacobian determinant J at each point (i, j) of the deformation field ϕ can be formulated as

$$\det(J_\phi(i, j)) = \begin{vmatrix} \frac{\partial i}{\partial x} & \frac{\partial j}{\partial x} \\ \frac{\partial i}{\partial y} & \frac{\partial j}{\partial y} \end{vmatrix} \quad (9)$$

For 3D images, the 3D Jacobian determinant of each point (i, j, k) is used. This can be similarly defined as:

$$\det(J_\phi(i, j, k)) = \begin{vmatrix} \frac{\partial i}{\partial x} & \frac{\partial j}{\partial x} & \frac{\partial k}{\partial x} \\ \frac{\partial i}{\partial y} & \frac{\partial j}{\partial y} & \frac{\partial k}{\partial y} \\ \frac{\partial i}{\partial z} & \frac{\partial j}{\partial z} & \frac{\partial k}{\partial z} \end{vmatrix} \quad (10)$$

Retinal image modalities

To illustrate retinal image registration, four commonly used techniques for photographing the eye are introduced: CF photography, FA, OCT, and optical coherence tomography angiography (OCTA). These techniques

provide various medical imaging tools to analyze retinal conditions.

CF photography

CF photography involves the use of a fundus camera to capture color images of the retina using white light. Equipped with a low-power microscope, the camera magnifies the interior surface of the eye. This technique is cost effective and straightforward for trained professionals [15]. The CF images (Fig. 3a) contain a broader range of fundus and rich color information, making it helpful in checking the atrophy of the retina and macular. Additionally, it helps diagnose retinopathies, such as diabetic retinopathy (DR), AMD, and glaucoma, as well as reveal signs of systemic diseases, such as diabetes and cardiovascular diseases [16].

FA

The FA, shown in Fig. 3b, involves a special dye called fluorescein and a camera to trace blood flow in the retina and choroid. It uses a special dye, fluorescein, and a camera to examine blood flow in the retina and choroid. The radiopaque dye is injected into the vein of the tester's arm, and the retinal vessels are photographed by tracing the dye before and after injection. FA can be used to detect capillary leakage [17], aneurysms, and neovascularization. However, some people may experience discomfort after the procedure [18].

OCT

OCT is an imaging technology that uses the interference between an investigated object and a local reference signal to create high-resolution cross-sectional images and 3D scans of the retina and anterior segment [19]. Figure 3c shows a cross-sectional scan of OCT. It is a non-invasive technique that enables visualization of each layer of the retina, measurement of its thickness, and provides treatment guidance for conditions such as glaucoma, DR, and AMD. Intraoperative OCT (iOCT) is necessary in many retinal therapies, including glaucoma surgery [20] and epiretinal device implantation [21], because it provides real-time visualization of the retinal layers.

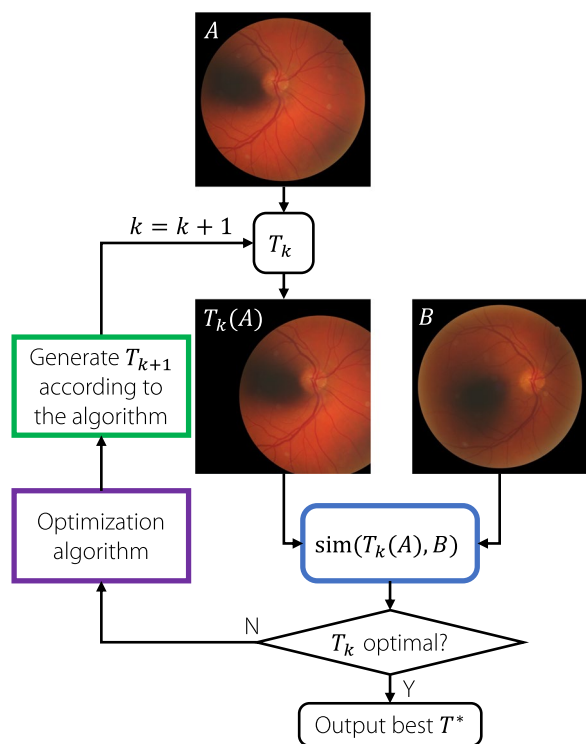


Fig. 4 General registration procedure using iterative optimization

OCTA

Figure 3d showcases OCTA, an emerging imaging technology that builds upon OCT. OCTA captures images of the vascular network with a higher resolution and smaller view than FA without invasiveness. Using the decorrelation signal produced by moving blood cells, OCTA generates an image of the microvascular network. Recent studies have demonstrated the ability of OCTA to overcome the limitations of assessing blood flow in the optic nerve, explain the vascular pathogenesis of glaucoma [22] and show impressive success in preclinical DR diagnosis [23].

Traditional registration methods

Researchers have developed increasingly sophisticated algorithms and resilient features during the initial image registration phase to achieve precise registration. This paper employs the phrase “traditional methods” to differentiate between the techniques utilized before the advent of deep learning and those implemented thereafter.

Intensity-based methods

Intensity-based methods treat this problem as an iterative optimization problem. The basic steps of the intensity-based registration are shown in Fig. 4. Initially, a random transformation T_0 is selected, and an objective

function is defined to measure the similarity between the transformed image $T_k(A)$ and another image B . The goal is to find the optimal transformation T^* to maximize similarity. At each step, the optimization algorithm applies a perturbation to the parameters in T based on the current similarity measure $\text{sim}(T_k(A), T(b))$. The process is terminated when the similarity satisfies the requirement, or converges with no further increase.

Researchers have mainly concentrated on developing various similarity functions, including (normalized) cross-correlation (CC), (normalized) mutual information (MI), and sum of squared differences (SSD). These functions are typically calculated by using the difference between each corresponding pixel in an input image pair. MI is considered the most important and widely used function. The large-deformation diffeomorphic metric mapping [28] model is based on manifold learning theory and uses the Euler-Lagrange equation for optimization. It regards the image as a point on the manifold and achieves image registration by calculating the deformation between the manifolds. This model can handle large deformations and maintain the nonlinear structure of an image.

Recently, several studies have been conducted using intensity-based methods. Lange and Heldmann [29] proposed a normalized gradient field (NGF) distance measure for 2D-3D image registration. To overcome the drawback that standard similarity measures may lead to optimization problems with many local optima, Öfverstedt et al. [30] adopted a symmetric, intensity-interpolation-free similarity measure that combines intensity and spatial information. Castillo [31] proposed an intensity-based deformable image registration optimization formulation that is easier to optimize. The similarity function is designed as a simple quadratic function that can be solved using a straightforward coordinate descent iteration.

Feature-based methods

Feature-based methods are popular methods of matching images based on their correspondence. These methods focus on the local structures and salient features of images, rather than on global information. The process is divided into three steps. First, features such as points, edges, and regions are extracted from the input images. Next, a descriptor is calculated for each feature. In the matching stage, the closest features of the two images are matched to establish potential correspondences. The idea is that the corresponding points should have similar descriptors. Finally, the transformation parameters are estimated based on the matching results. The primary challenge is to determine the most effective method for

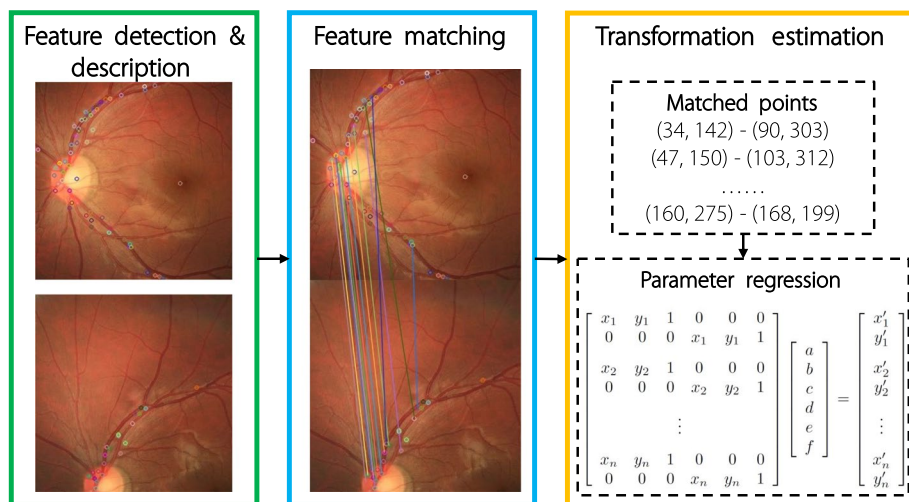


Fig. 5 General keypoint-based registration procedure

extracting and describing features. Figure 5 illustrates the keypoint-based registration process.

One pioneering work in feature-point-based registration is the scale-invariant feature transform (SIFT) [32]. SIFT transforms image data into scale-invariant coordinates, identifies stable keypoints, assigns orientations to keypoints, and generates feature descriptors for each keypoint. The extracted features are invariant under variations in scale, brightness, and angle. However, the process is computationally expensive. To address this problem, various efforts [33–37] have been made to enhance the performance and efficiency of SIFT. For instance, speeded up robust features (SURF) [33] simplify the filter function to reduce the dimensions of descriptors and improve computational efficiency. Another method, oriented FAST and rotated BRIEF (ORB) [36], integrates the FAST [38] keypoint detector and the BRIEF [39] descriptor to solve the high computational cost of SIFT features and the lack of rotation invariance, scale invariance, and sensitivity to noise in the BRIEF feature. As a result, ORB is capable of delivering a speedup of up to two significant figures compared with SIFT. Other studies have focused on edge and contour features using classic edge detection [40, 41] and image segmentation [42] algorithms for feature extraction.

Deep learning-based registration methods

Deep learning-based image segmentation has proven to be a robust tool for image segmentation since 2019 [43]. These methods can improve accuracy and efficiency by automatically learning high-level features from the input images. Registration tasks, similar to segmentation, have been developed using deep learning methods. They differ from feature-based approaches because they utilize

deep neural networks to replace feature extractors, feature matching, and transformation processes. Rather than directly optimizing the transformation parameters, these methods indirectly optimize the registration model parameters, thereby revealing the true essence of their effectiveness.

Feature-based methods

The CNN is a pioneering work in computer vision. It uses learnable convolution kernels and inductive biases, such as locality and translation equivariance, to detect learned patterns in local regions and extract high-level features. This characteristic makes CNNs particularly suitable for object detection and image registration tasks, where spatial features are essential. Table 1 displays the prominent works on CNN-based registration methods, which have become the most popular approaches in the field since 2016.

Patch-based methods

Instead of the direct regression of registration parameters from the image pair, a patch-based approach is used to divide the image into smaller patches. The patch is utilized in different ways depending on the predicted transformation type. For linear transformations, the network establishes a match that can be used to derive the registration parameters. Conversely, a local displacement field is output and combined for nonlinear transformations. Various CNN models were proposed by Zagoruyko and Komodakis [73] that output the similarity between two image patches as feature descriptors. Cao et al. [46] proposed a similarity-steered CNN regression architecture that estimates the displacement vectors at each corresponding location between linearly aligned brain MR

Table 1 Overview of feature-based image registration methods

Reference	Year	Scene	Dimension	Modality	Type	TS	MM	Net architecture	Evaluation metric	Loss function
Miao et al. [44]	2016	Virtual	2D/3D	X-ray/CT	R	S	N	CNN regressor	SR/TRE	MSE
Yang et al. [45]	2017	Brain	3D	MR	D	S	N	Dual branch	MSE/ I_{ϕ}	MAE
Cao et al. [46]	2017	Brain	3D	MR	D	S	N	CNN regressor	Dice/ASSD	MSE
de Vos et al. [47]	2017	Digits/Heart	2D	Digit/MR	D	U	N	CNN regressor	Dice/95SD/ASSD	NCC
Zheng et al. [48]	2018	Bone	2D/3D	X-ray/CT	R	S	Y	Dual branch	RMSE/Error rate/TRE	MSE/PDA
Sloan et al. [49]	2018	Brain	3D	MR	R	S	N	CNN regressor	MAE	MSE
Chen and Wu [50]	2018	Brain	3D	MR	A	S	N	Siamese	Jacc/HD	MSE
Ly et al. [51]	2018	Abdominal	3D	MR	D	S	N	CNN regressor	SNR	NCC
Hu et al. [52]	2018	Prostate gland	3D	MR/US	D	W	Y	FCN	RMSE/Dice	Dice + Smooth
Jiang and Shackleford [53]	2018	Chest	3D	CT	D	U	N	CNN regressor	SSD	-
Li and Fan [54]	2017	Brain	3D	MR	D	U	N	FCN	Dice	NCC + Smooth
Fan et al. [55]	2019	Brain	3D	MR	D	S	N	FCN	Dice	MSE
Xu and Niethammer [56]	2019	Knee/Brain	3D	MR	D	W	N	FCN	Dice	Dice + NCC + Smooth
de Vos et al. [57]	2019	Heart/Chest	3D	MR/CT	A/D	U	N	Siamese	Dice/HD/ASSD/ I_{ϕ}	NCC + Smooth
Zhao et al. [58]	2019	Liver/Brain	3D	CT/MR	A/D	U	N	FCN	Jacc/Lm. Dist	CC + Smooth + Orthogonality + Determinant
Zhao et al. [59]	2019	Liver/Brain	3D	CT/MR	A/D	U	N	FCN	Dice/Lm. Dist	CC + Smooth
Balakrishnan et al. [60]	2019	Brain	3D	MR	D	W/U	N	FCN	Dice/ I_{ϕ}	MSE/CC + Dice + Smooth
Dalca et al. [61]	2019	Brain	3D	MR	D	U	N	FCN	Dice/ I_{ϕ}	Variational inference + Surface matching
Hu et al. [62]	2019	Brain	3D	MR	D	U	N	Siamese + Pyramid	Dice	NLCC + Smooth
Wang and Zhang [63]	2020	Synth./Brain	2D/3D	Eye/MR	D	S	N	Dual-FCN	Dice	MSE + Smooth
Mansilla et al. [64]	2020	Chest	2D	X-ray	D	W	N	FCN	Dice/HD/ASSD	NCC + Smooth
Mok and Chung [65]	2020	Brain	3D	MR	D	U	N	FCN	DSC/ I_{ϕ}	NCC + Pair + Smooth + Magnitude
Kim et al. [66]	2021	Face/Brain	2D/3D	Expr/MR	D	U	N	FCN	NMSE/SSIM/Dice/ I_{ϕ}	LCC + Smooth + Cycle + Identity
Czolbe et al. [67]	2021	Brain/Cell	3D	MR/EM	D	W/U	N	FCN	Dice	Cos sim of feature extractor
Mok and Chung [68]	2022	Brain	3D	MR	D	U	N	FCN	MAE/SR	NCC + Inverse + Smooth
Kang et al. [69]	2022	Brain	3D	MR	D	U	N	Siamese + Pyramid	Dice/ASSD/HD/ I_{ϕ}	NLCC + Smooth
Tran et al. [70]	2022	Liver/Brain	3D	CT/MRI	D	U	N	FCN	Dice/Jacc	Adversarial + Discrimination + Reconstruction
Kong et al. [71]	2023	Brain	2D/3D	CT/MR	D	U	Y	FCN	Dice/HD	Evaluator + Smooth
Che et al. [72]	2023	Brain	3D	MR	D	U	N	FCN	Dice/ I_{ϕ} /ASSD	NCC + Smooth + Anti-folding

For the **TS** (Training Strategy) column, S: Supervised, W: Weakly supervised, and U: Unsupervised. In the **MM** (multi-modal) column, Y: Yes, and N: No. In the **Type** column, R: Rigid, A: Affine, P: Perspective, and D: Deformable

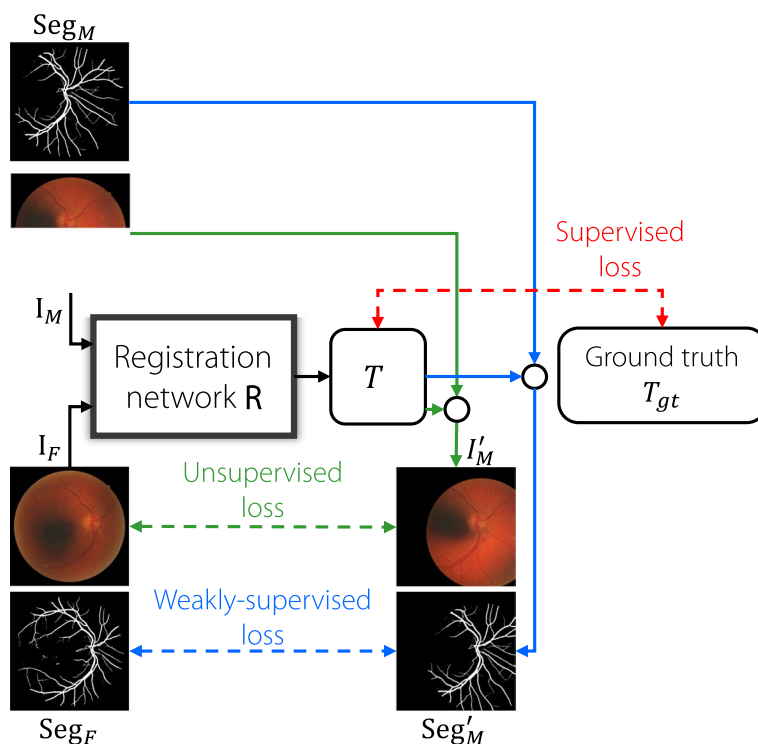


Fig. 6 The overall framework for end-to-end deep learning-based medical image registration methods. The moving image I_M and the fixed image I_F are sent into the registration network R , and the output is obtained as the predicted transformation T . Seg_M denotes the anatomical segmentation label of I_M while Seg_F denotes the anatomical segmentation label of I_F . The small circles denote performing transformation T on I_M or Seg_M using STN [75], gaining warped moving image I'_M or warped label Seg'_M . Red, blue, and green lines denote the supervised, weakly-supervised, and unsupervised training strategies, respectively

pairs. Interpolation is subsequently utilized to obtain a dense deformation. Lv et al. [51] divided the signal into three bins and used a CNN to estimate the displacement field for abdominal motion correction throughout the respiratory cycle. However, these methods typically require additional steps of patch selection and final registration, which can be time-consuming. In addition, the generation or manually labeling ground truth can be a limiting factor.

End-to-end CNN methods

Supervised end-to-end networks have been developed for direct registration owing to their increased computing power. The ground truth is obtained using traditional algorithms or manual labels. A general end-to-end deep learning registration framework is shown in Fig. 6. Miao et al. [44] employed 2D/3D CNN regressors to directly estimate rigid transformation parameters in real time. Quicksilver [45] divides 3D brain MRI into 3D patches owing to the limitations of GPU memory; however, it can directly predict the deformation field for the input patches. To improve the performance of supervised methods, Chee and Wu [50] leveraged unlabeled data to

generate a synthetic dataset, and trained an affine image registration network. BIRNet [55] was proposed as a hierarchical dual-supervised fully CNNs based on U-Net [74] in the following year, with a loss function designed as a combination of the difference in image intensity and the difference in predicted displacement and ground truth displacement in each layer of U-Net's decoder. Wang and Zhang [63] introduced a low-dimensional Fourier representation of diffeomorphic transformations to improve training and inference efficiency.

Weakly supervised registration methods take advantage of additional semantic information to ensure meaningful registration and overcome the challenge of the unavailability of ground truth transformations. These methods utilize additional information such as anatomical segmentation to perform registration. Hu et al. [52] proposed a weakly supervised registration network for multimodal 3D prostate gland images using the ground-truth segmentation labels of the gland and other anatomical landmarks. Xu and Niethammer [56] proposed a deep learning framework called DeepAtlas that jointly learns networks for image registration and segmentation, which are trained alternately and complement each

other to achieve better results with only a few labels for segmentation.

Unsupervised methods have also been studied to eliminate the ground-truth labels. The spatial transformer layer (STL) [75], which is a differentiable module that can warp an input image, is the foundation of many unsupervised registration methods. STL enables the transformation of a moving image in a differentiable manner, allowing the application of conventional similarity measurements between the transformed and fixed images during training as the loss function. In 2017, DIRNet [47] was introduced as the first end-to-end unsupervised deformable registration network that adopted STL. Subsequently, VoxelMorph [60] was proposed as a U-Net-based network that achieved faster runtime and better performance than traditional iterative-based methods, with only unsupervised training. Auxiliary anatomical segmentation can be performed under weakly supervised settings. In their subsequent study, Dalca et al. [61] adopted a probabilistic generative model to provide diffeomorphic guarantees. Dual-PRNet [62] extended VoxelMorph [60] by incorporating a pyramid registration module that uses multilevel context information and sequentially warps convolutional features. Dual-PRNet++ [69] further enhances the PR module in Dual-PRNet by computing the correlation features and using residual convolutions.

Deep similarity methods

Pixel-based similarity metrics, such as MSE and NCC, are commonly employed in deep learning. However, these metrics may encounter difficulties when dealing with low-intensity contrasts or noise. To address these issues, deep similarity methods that utilize custom similarity measures have been developed. For example, DeepSim [67] utilizes semantic information extracted by a pretrained feature extractor in a segmentation network to construct a semantic similarity metric. This specialized metric allows the network to learn and adapt to dataset-specific features, thereby improving the low-quality image performance. IMSE [71] takes this a step further with a self-supervised approach to train a modality-independent evaluator using a new data augmentation technique called shuffle remap, which can provide style enhancement. The evaluator then serves as a multimodal similarity estimator to train the multimodal registration network.

Cascade methods

Cascade methods were inspired by traditional iterative registration methods. The cascade architecture, that is, stacking networks in series, can provide progressive registration in a coarse-to-fine manner. DLIR [57]

implemented a cascade architecture by stacking an affine network followed by multiple deformable networks, with each network being trained sequentially and the weights of the previous networks fixed. By contrast, Zhao et al. [58, 59] proposed a recursive cascade architecture similar to DLIR but much more sophisticated. They jointly trained their cascade networks to learn the progressive alignments more effectively.

Consistency-based methods

Consistency-based methods add consistency constraints based on the registration or transformation properties. In 2020, Mok and Chung [65] addressed the challenge of deformable transformation invertibility by introducing a swift and symmetric diffeomorphic image-registration approach. The network was trained with an inverse-consistency constraint, which enabled it to learn the bidirectional transformations of the mean shape of two input images to produce topology-preserving and inverse-consistent transformations. In the following year, Kim et al. [66] proposed CycleMorph, which utilizes cycle consistency as an additional constraint to enhance topology preservation and reduce folding issues. To register images X to Y and Y to X , the method employs two CNNs: G_X and G_Y . The warped images from both networks are used as image pairs and sent to the networks themselves to ensure that they could be returned to their original state, maximizing the similarity between the original and reversed images.

Other methods

However, with the development of novel architectures, the number of parameters has increased significantly, making it more difficult to achieve real-time registration without high computing power. Tran et al. [70] attempted to solve this problem using knowledge distillation. They transferred meaningful knowledge of distilled deformations from a pretrained high-performance network (teacher network) to a fast, lightweight network (student network). After training, only a lightweight student network is used during the inference, allowing the model to achieve a fast inference time using only a common CPU.

Translation-based methods

Multimodal image registration can be complex, because it involves aligning images of varying modalities with unique intensity distributions. This poses a challenge for unimodal methods. However, an innovative solution to this issue is to leverage image translation techniques. This solution transforms the multimodal registration problem into a more straightforward unimodal registration

Table 2 Overview of translation-based image registration methods

Methodology	Reference	Year	Scene	Dimension	Modality	Type	Evaluation metric
GAN	Mahapatra et al. [76]	2018	Retina/Heart	2D	CF/FA/MR	D	Dice/HD/ASD
	Qin et al. [77]	2019	Lung/Brain	2D	CT/MR	D	Dice/MCD/HD/RMSE
	Xu et al. [78]	2020	Kidney/Abdomen	3D	CT/MR	D	Dice/TRE
	Han et al. [79]	2022	Brain	3D	MR/CT	D	Dice/SD/HD/TRE
	Zhang et al. [80]	2023	Liver	3D	US	D	TRE
Contrastive learning	Casamitjana et al. [81]	2021	Brain	3D	Histology/MRI	D	RMSE/Dice
	Chen et al. [82]	2022	Thorax/Abdomen/Lung	3D	CT/MRI	D	Dice/HD95
DDPM	Kim et al. [83]	2022	Face/Brain	2D/3D	Expression/MR	D	Dice/ J_{ϕ}

For the **Type** column, R: Rigid, A: Affine, P: Perspective, and D: Deformable

problem, as shown in Fig. 7. Table 2 lists the most widely available translation-based registration algorithms.

GANs

A GAN [84] consists of two subnetworks, a generator and a discriminator, trained in a game-theoretic setting to generate synthetic data that are indistinguishable from the actual data. The generator generates synthetic samples, whereas the discriminator attempts to differentiate between natural and synthetic samples. The training process continues until the generated samples are indistinguishable from the actual ones.

Mahapatra et al. [76] used a GAN to generate a registered image with a distribution identical to that of the moving image and deformation field. They also ensured that the structure of the generated image matched that of the reference image through a structural similarity loss. Qin et al. [77] proposed a method

for decomposing images into a latent shape space and separate latent appearance space for both modalities, which were used to learn a bidirectional registration function.

CycleGAN [85], which is based on GAN, enables image-to-image (i2i) translation using unpaired images. It employs cycle consistency loss to ensure that the reconstructed images are consistent with the original input images. Several multimodal registration methods [78, 79] have used CycleGAN as the primary network for image translation. Xu et al. [78] introduced two additional losses to enforce structural similarity between translated and authentic images. They also jointly trained the translated unimodal and multimodal streams to complement each other. Han et al. [79] implemented image synthesis in both directions and predicted the associated uncertainty, providing the information used in the fusion of the two direction estimations.

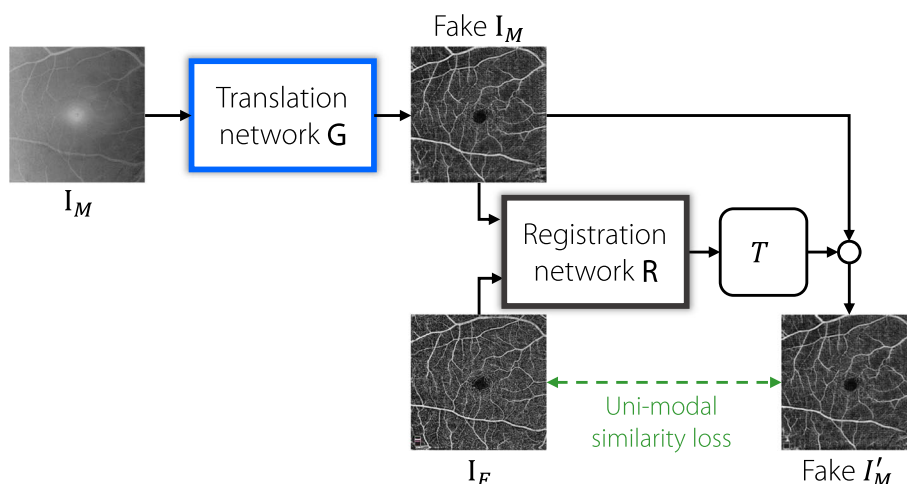


Fig. 7 Overall framework for translation-based methods. The moving image I_M is first sent into the translation network G which performs inter-modality translation and outputs the fake image $Fake I_M$. Then, $Fake I_M$ and the fixed image I_F are sent into the registration network R , and the output is obtained as the predicted transformation T . The small circles denote performing transformation T on $Fake I_M$ using STN [75], gaining warped fake moving image $Fake I'_M$

Table 3 Overview of transformer-based image registration methods

Reference	Year	Scene	Dimension	Modality	Type	Net architecture	Evaluation metric	Loss function
Chen et al. [90]	2021	Brain	3D	MR	D	Hybrid	Dice	MSE + Smooth
Zhang et al. [91]	2021	Brain	3D	MR	D	Hybrid	-	-
Mok and Chung [92]	2022	Brain	3D	MR	A	Pure	Dice/HD	NCC + Dice
Chen et al. [93]	2022	Brain/Heart	3D	MRI/XCAT/CT	A/D	Hybrid	Dice/ J_ϕ /SSIM/HD	(MSE/LNCC) + Dice + Smooth
Song et al. [94]	2022	Brain	3D	MR	D	Hybrid	Dice/ J_ϕ	(MSE/LNCC) + Smooth
Wang et al. [95]	2022	Brain	3D	MR	D	Hybrid	Dice	LCC + Smooth
Shi et al. [96]	2022	Heart	3D	CT	D	Pure	Dice/ J_ϕ	Sim + Smooth
Zhu and Lu [97]	2022	Brain	3D	MR	D	Pure	Dice/ J_ϕ	MSE + Smooth + Determinant + Inverse
Chen et al. [98]	2023	Brain	3D	MR	D	Hybrid	Dice/HD/ J_ϕ	(LNCC/MSE) + Smooth
Wang et al. [99]	2023	Brain	3D	MR	D	Hybrid	Dice/ASSD/ J_ϕ	NCC + Smooth

For the **Type** column, R: Rigid, A: Affine, P: Perspective, and D: Deformable

Contrastive learning

Contrastive learning defines positive and negative samples, and the goal is to learn a representation space where positive samples are close to each other and negative ones are far away. A recent study by Park et al. [86] explored the integration of contrastive learning into image translation by introducing an additional loss called patchNCE to a naive GAN. This loss encourages the generated output patches to be closer to their corresponding image patches than to random ones. Casamitjana et al. [81] used patchNCE loss to train an i2i translation network for transferring source images to the desired target domain. Subsequently, they applied an independently trained intramodality registration network to the target domain to predict the deformation field. Building on this work, Chen et al. [82] proposed an end-to-end architecture that jointly trains registration and translation networks without requiring a discriminator.

Denosing diffusion probabilistic model

A new generative model called the denoising diffusion probabilistic model (DDPM) [87] was recently introduced. This model is designed to learn Markov transformation from a simple Gaussian distribution to an actual data distribution. DDPM has been shown to generate images of higher quality than GAN [88]. In addition, Kim et al. [83] developed DiffuseMorph, which is the first and currently the only registration network based on diffusion. The network estimates the score function by adding a diffusion network before a standard registration network, and even shows the image registration trajectory by scaling the conditional score. However, unlike translation between modalities, DiffuseMorph

constructs a score function directly between the input image pairs.

Transformer-based methods

Recently, Google explored a method to use a pure transformer architecture in vision tasks, known as a vision transformer (ViT) [89], achieving competitive performance compared to existing CNN methods. ViTs split the image into patches and treat them as tokens, as in an NLP application, which has led to their successful application in various computer vision tasks, including image registration. Table 3 presents transformer-based image registration methods.

Hybrid methods

Initially, researchers attempted to integrate Transformers into CNN-based models. Chen et al. [90] pioneered the use of ViT on high-level features extracted from the convolutional layers of moving and fixed images. Building on this approach, Song et al. [94] proposed TD-Net, which utilizes multiple transformer blocks for down-sampling. Conversely, Zhang et al. [91] introduced a dual transformer network comprising two branches, intra-image and inter-image, with transformers embedded in both branches to enhance the features, similar to the approach in ref [90]. Wang et al. [95] enhanced the UNet [74] architecture for registration by introducing a bilevel connection and a unique transformer block. TransMorph [93] was proposed as a hybrid transformer-ConvNet model that utilizes Swin transformers [100] in the encoder and convolutional layers in the decoder. The authors demonstrated that positional embedding can be disregarded, leading to a flatter loss landscape for registration. The following year, Chen et al. [98] proposed TransMatch, emphasizing the importance of inter-image feature matching. They employed a transformer-based

encoder and matched the regions using their new local window cross-attention module. Recently, Wang et al. [99] introduced a motion decomposition transformer based on a multihead neighborhood attention mechanism that can model multiple motion modalities.

Pure transformer methods

An alternative method involves the integration of a pure transformer architecture into a network. In a recent study by Shi et al. [96], a unique X-shaped transformer architecture called XMorpher was introduced. The researchers incorporated cross-attention between two feature extraction branches and a window-size constraint to enhance the information exchange and locality of the network. In another study, Swin-VoxelMorph [97] utilized a fully Swin transformer-based 3D Swin-UNet and a bidirectional constraint to optimize both forward and inverse transformations. To fill this gap in affine image registration, Mok and Chung [92] proposed a Coarse-to-Fine vision transformer, a pure transformer architecture. The researchers transformed the image pairs into small-to-large resolutions and passed them through different stages of ViT to achieve the desired results.

Analysis

The evolution of image registration methods has been closely tied to advancements in computing power and deep-learning architectures. In the early stages, when computing power was limited, patch-based methods predominated. However, as computational capabilities and network diversity have expanded, it has become feasible to process entire images, and even 3D volumes, in a holistic manner. This shift facilitated the simultaneous and integrated performance of feature extraction and matching tasks. Concurrently, the feature extraction component of image registration has been progressively enhanced by the rapid development of deep-learning architectures.

Translation-based methods are effective in mitigating multimodal registration challenges by aligning image pairs within the same modality, thereby simplifying the registration process. Recently, there has been a surge in generative network-assisted registration methods that capitalize on the latest advancements in generative network models. Although GANs have shown promise in modality translation, their training process is notably complex and demands meticulous manual hyperparameter tuning for both the generator and discriminator components. Previously, contrastive learning dominated the unsupervised learning landscape; however, it requires extensive high-quality datasets for effective training. Diffusion models have recently emerged as promising image-generation techniques capable of producing highly

realistic effects. However, its potential application in image registration remains an open research area.

In a CNN, the convolution operations are typically localized, focusing on extracting features from within a specific neighborhood. By contrast, the transformer architecture, with its self-attention mechanism, offers a distinct advantage by facilitating the exchange of information across the entire image. This capability is a key factor driving the integration of transformer models into registration networks because it significantly enhances the feature extraction process by considering global contextual information. There is also a trend towards the development of pure transformer architectures that have exhibited remarkable performances in various visual tasks. However, adapting the attention mechanism to suit specific requirements of image registration remains a problem. Therefore, cross-attention transformers are being investigated for their potential to refine the feature extraction phase and improve the feature-matching stage. This tailored approach can lead to more effective and robust registration methods, particularly for complex multimodal imaging scenarios.

By shifting the focus to the architecture of neural networks, distinct preferences in medical-image registration were observed. For linear registration, the CNN regressor stands out as the favored architecture owing to its versatility in both feature extraction and direct regression for obtaining linear registration parameters. By contrast, fully convolutional networks (FCN), particularly those resembling the UNet architecture [74], are favored for nonlinear registration. This is because of the FCN's ability to produce a deformation field that corresponds to the size of the input image, making it exceptionally suitable for such tasks. The FCN architecture typically includes an encoder-decoder framework, with the encoder responsible for feature extraction and the decoder responsible for analyzing these features to generate results. A skip connection between the encoder and decoder facilitates the integration of the extracted features, enhancing the predictive capabilities of the network. Interestingly, more recent transformer-based models, which have had a significant impact, often adhere to this fundamental structure.

Building upon the FCN, derivative models such as the Siamese network and dual-branch network have been developed. These models employ two encoders that independently extract features from the input image and subsequently interact with and merge these features. In the context of single-modal registration tasks, the Siamese network, which shares weights between two encoders, is commonly utilized for its efficiency. However, in multimodal registration tasks, this approach diverges by employing two distinct encoders with separate weights

to adaptively extract consistent features from two different modalities. Furthermore, to achieve a more robust deformation field, certain networks have been designed to output a pyramid of multiscale deformation fields. These fields are then integrated to form the final deformation field. The advantage of this multiscale approach is that it incorporates features from various levels of detail, rather than relying solely on the features produced by the decoder's final output.

Innovation in loss functions is a critical aspect in the development of neural networks for medical image registration. In supervised training, the MSE between the predicted and true transformation parameters is the prevalent choice for the loss function. This metric provides a straightforward quantification of prediction accuracy. When shifting to unsupervised training, in which ground-truth transformation parameters are unavailable, image-similarity measures become essential. The most widely utilized image similarity losses include the MSE and CC. The CC, in particular, is calculated as follows:

$$CC(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Cov}(X, X)\text{Cov}(Y, Y)}} \quad (11)$$

where $\text{Cov}(X, Y) = \frac{1}{|\Omega|} \sum_{x \in \Omega} X(x)Y(x) - \frac{1}{|\Omega|^2} \sum_{x \in \Omega} X(x) \sum_{y \in \Omega} Y(y)$ is the covariance. In weakly supervised training scenarios, an additional loss function often comes into play—the dice loss, which is predicated on the segmentation labels of image pairs. This loss function is particularly adept at capturing spatial agreement between segmentations. Moreover, for nonlinear registration tasks, it is crucial to incorporate a smoothing penalty term into the loss function. This term encourages smoothness in the deformation field by promoting similarity in deformation quantities across adjacent positions. The most favored penalty term is the diffusion regularizer, which is mathematically expressed as:

$$R_{\text{diff}}(\phi) = \sum \|\nabla_{\phi}\|^2 \quad (12)$$

Furthermore, the incorporation of novel constraints that leverage the fundamental properties of the registration and transformation processes results in the creation of more refined output transformations. This approach is the cornerstone of consistency-based methods, which aim to ensure that the transformations generated by the network closely adhere to the underlying physical and geometrical principles of the registration task. In addition to these advancements, deep similarity methods have introduced the concept of training an evaluator or a custom similarity function to serve as a network's loss function. This approach enables the network to automatically learn an appropriate similarity metric that aligns

with the specific characteristics and requirements of the task.

While traditional methods require cumbersome iterative optimization calculations, which result in significant time consumption, deep learning-based approaches offer a notable efficiency advantage by allowing data to be input into the network during testing and providing results immediately after training. Furthermore, from a preprocessing standpoint, both traditional and deep learning-based methods require the downsampling of typically collected high-resolution medical images [101]. Utilizing the original scale would not only amplify the search space for the iterative optimization algorithm but also increase the number of parameters required by the deep learning network, imposing substantial overhead on both methodologies. Nevertheless, traditional methods have the advantage of delivering more stable outcomes and are convenient for plug-and-play applications. By contrast, deep learning-based methods require specialized training for each task. The trained model becomes obsolete when the application context shifts.

Registration application in retinal images

Traditional methods

First, intensity-based methods for retinal image registration were explored. The aforementioned intensity similarity metrics, such as MI [102–104] and CC [105], were used. Feature-based methods are more effective than intensity-based methods for retinal image registration. One popular approach is to use typical landmarks in retinal images. In 2003, Stewart et al. [106] introduced a Dual-Bootstrap Iterative Closest Point (Dual-Bootstrap ICP) algorithm for retinal image registration. This algorithm begins by matching individual vascular landmarks and aligning images based on the detected blood vessel centerlines. Other studies have utilized vascular features [107–110] and optical discs [111] for registration purposes.

One potential solution is to enhance the capabilities of keypoint detectors and feature descriptors to improve their performance. Ramli et al. [112] designed a D-saddle detector capable of detecting feature points even in low-quality regions. Yang et al. [113] built upon previous work [106] to create the generalized dual-bootstrap iterative closest point, which uses better initialization, robust estimation, and strict decision criteria to align retinal images from different modalities. Chen et al. [114] implemented a Harris detector to identify corner points, extract partial intensity-invariant feature descriptors, and perform bilateral matching between image pairs. The outliers are then removed, and the final transformation is applied. Ramli

Table 4 Overview of deep learning-based retinal image registration methods

Reference	Year	Modality	Type	TS	MM	Net architecture	Evaluation metric	Loss function
Lee et al. [119]	2019	CF/FA /OCT	A	U	Y	CNN regressor	SR/RMSE/MAE/MEE	-
Zhang et al. [120]	2019	CF/FA	D	W	Y	FCN	Dice	Style + Content + MSE + SSIM + Smooth
De Silva et al. [121]	2020	CF/F/IR	A/D	S	Y	Siamese + CNN regressor	Reg. error	Overlap + Displacement
Wang et al. [122]	2020	CF/IR	P	S	Y	Seg. + Det. and Desc. + Out. Rej.	SR/Dice	CE + MSE + Dice
Tian et al. [123]	2020	CF/OCT	D	U	Y	FCN + Pyramid	MSE/HD/MSSIM	CC + Edge + Smooth
Zou et al. [124]	2020	CF	D	U	N	FCN + Pyramid	PA/Dice/RMSE	NCC + Smooth
Wang et al. [125]	2021	CF/FA/IR	P	S	Y	Seg. + Det. and Desc. + Out. Rej.	SR/Dice	CE + MSE + Dice
Zhang et al. [126]	2021	CF/FA/IR	A/D	S	Y	Seg. + Det. and Desc. + Out. Rej. / FCN	Dice	CE + MSE + Dice/Style + MSE
Sui et al. [127]	2021	MSI	D	W	N	FCN + Pyramid	TRE/Dice	Sim + Smooth
An et al. [128]	2022	CF/FA/IR	R	U	Y	Seg. + Det. and Desc. + Out. Rej.	SR/Dice	Pos. + Desc. + Score + CE + (MSE/Dice)
Benvenuto et al. [129]	2022	CF	D	U	N	FCN	MSE/SSIM/Dice	NCC
López-Varela et al. [130]	2022	OCTA	D	U	N	FCN + Pyramid	MSE/NRMSE/SSIM/VIF	LNCC
Rivas-Villar et al. [131]	2022	CF	S	S	N	FCN	RMSE/AUC	MSE
Kim et al. [132]	2022	CF	P	S	N	FCN	AUC	CE + Focal + Smooth
Santarossa et al. [133]	2022	CF/FAF/FAG	P	S	Y	CNN regressor	AUC	Relaxed Ranking
Liu et al. [134]	2022	CF	P	S	N	Det. and Desc.	EER	Det. + Desc.
Rivas-Villar et al. [135]	2023	OCT	A+Z	U	N	Det. and Desc.	Error	Repeatability + Reliable
Liu and Li [136]	2023	CF	P	U	N	CNN + Attn.	AUC/SR	CE

For the **TS** (Training Strategy) column, S: Supervised, W: Weakly supervised, U: Unsupervised. For the **MM** (multi-modal) column, Y: Yes, N: No. For the **Type** column, R: Rigid, S: Similarity, A: Affine, P: Perspective, Z: Z-axis, and D: Deformable

et al. [112] improved the saddle detector to detect feature points in low-quality regions. Gharabaghi et al. [115] utilized affine moment invariants as shape descriptors. By combining the domain knowledge, SIFT and its variants are used in refs [116, 117]. Li et al. [118] introduced orientation-independent feature matching that uses a new circular neighborhood-based feature descriptor.

Deep learning-based methods

In this subsection, we review deep learning-based methods for retinal applications, categorized as outlined in “[Deep learning-based registration methods](#)” subsection. Table 4 summarizes the deep learning-based retinal image registration methods.

Feature-based methods

The identification of retinal landmarks has been a catalyst for the development of deep learning techniques. In particular, ref. [119] used handcrafted features, whereas [131, 132] used CNNs. Specifically, Lee et al. [119] employed a CNN to classify patches of various step patterns based on intensity changes. By contrast, Rivas-Villar et al. [131] used a CNN to produce a heatmap of blood vessels and bifurcations, and applied the maximum detection and

feature matching method RANSAC [137] during testing. Similarly, Kim et al. [132] used a vessel segmentation network and joint detection network to identify vascular landmark points for registration. The SIFT algorithm [32] is then used to compute the descriptors based on the regions around these points. Benvenuto et al. [129] used an Isotropic Undecimated Wavelet Transform to segment blood vessels and ocular shapes. Based on the segmentation, the registration network adopted from U-Net is trained to perform registration. This year, Rivas-Villar et al. [135] explored deep learning registration methods for OCT 3D Scan. They first performed affine alignment on a 2D projection, followed by z-axis registration based on layer segmentation.

Recent studies explored the potential of end-to-end methods that utilize innovative network architectures. De Silva et al. [121] developed a model that employs a VGG 16 feature extractor, a correlation matrix, and a regression network to emulate the traditional feature-based registration pipeline, encompassing feature extraction, matching, and computation of the registration transformation; the effectiveness of their model was evaluated on a multimodal retinal dataset. Tian et al. [123] enhanced the U-Net architecture [74] by

incorporating an image pyramid for multiscale input and introduced a novel edge similarity loss calculated through the correlation between the gradients of the fixed and moving images. However, based on U-Net, Sui et al. [127] further refined this approach by feeding an image pyramid of the original image and a ground truth vessel map into each layer of the encoder and decoder, respectively. Liu et al. [134] proposed Super-Retina, an end-to-end method with jointly trainable keypoint detector and descriptor.

It is worth noting that Wang et al. [120, 122, 125, 126, 128] made significant contributions to multimodal retinal image registration. They initiated their work using a deformable registration model comprising a vessel segmentation network and a deformation field estimation network, as described in ref. [120]. In their subsequent study [122], they refined the vessel segmentation network from their prior work and integrated a pretrained superpoint model [138] for feature detection and description, complemented by an outlier rejection network to facilitate perspective registration. This three-stage methodology, consisting of segmentation, detection, description, and outlier rejection, was subsequently employed in ongoing research. A notable advantage of this approach is its ability to bridge the intensity gap between different modalities; however, the complexity of the methodology remains a drawback. They further improved the segmentation network using pixel-adaptive convolution [125]. In ref. [126], the authors introduced perspective registration as a coarse step, followed by the addition of a deformable framework for fine alignment to achieve remarkable accuracy. Most recently, ref. [128] transformed the three-stage approach into a self-supervised process.

Translation-based methods

Although numerous studies have been conducted on i2i translation in various retinal modalities [139, 140], few studies on retinal image registration have used translation-based techniques. MedRegNet [133], which utilizes CycleGAN [85] as an image-translation tool, is the only available method of its kind. However, it is primarily employed as a generator of multimodal retinal data, rather than as a registration tool. The aforementioned work [120, 122, 125, 126, 128] can also be regarded as translation-based when addressing multimodal data. These studies capitalize on image segmentation to produce blood vessel segmentation maps, effectively converting different modalities into a unified ‘mask’ modality for registration purposes.

Transformer-based methods

Research on transformer-based retinal image registration methods is still in its infancy. GeoFormer [136] is the first method to adopt an advanced transformer-based attention blocks for detector-free feature matching on retinal images. It enhances coarse features by using geometrically matched regions rather than entire images, resulting in more accurate coarse matches.

Analysis

In the domain of retinal image registration, traditional approaches have extensive applications and often employ various retinal modalities. Some studies have integrated domain-specific knowledge into general registration methodologies. However, these intensity-based methods can be sensitive to variations in illumination across image pairs, which may arise from differences in camera settings, imaging modalities, or changes in the retinal background due to retinopathy. This sensitivity is a common challenge affecting feature-based methods that require robust feature descriptors to perform well. Moreover, a significant drawback of many conventional registration techniques is their long inference times.

Deep learning-based methods for retinal image registration emerged more recently in 2019 and can be categorized into two main approaches. The first approach leverages state-of-the-art network architectures within the framework of mainstream registration methods. Although these methods deliver exceptional performance, their reliance on architectural design for domain-specific insights is notable. By contrast, the second approach aims to address the registration challenge in a manner that is more tailored to the domain. This involves extracting or utilizing key features such as vessel segmentation or vascular junctions for subsequent registration processes.

It has been observed that the diversity of approaches in retinal image registration appeared to be considerably lower than that in other areas of medical image registration. This can be attributed to several factors, including differences in the imaging principles and targets. For instance, imaging modalities such as CT and MR utilize X-rays and magnetic fields to generate images with high tissue contrast while maintaining consistent intensities across various acquisitions. By contrast, retinal image registration commonly relies on CF and FA images, which depend solely on white light illumination. The unique imaging principle of retinal imaging, combined with the natural movement of the subject’s eyeballs, can result in significant brightness variations within a

Table 5 Public retinal image registration datasets

Dataset	Source	Camera specifications	Format	Modality	Resolution	Size (pairs)	Ground truth
FIRE [24]	Papageorgiou Hospital, Aristotle University of Thessaloniki, Greece	Nidek AFC-210 fundus camera	JPG	CF	2912×2912	134	Control points
FLoRI21 [142]	RECOVERY study [143]	Optos California and 200Tx cameras	TIFF	UWF FA	3900×3072	15	Control points
CF-FFA [25]	Unknown	Unknown	JPG	CF & FFA	720×576	60	None

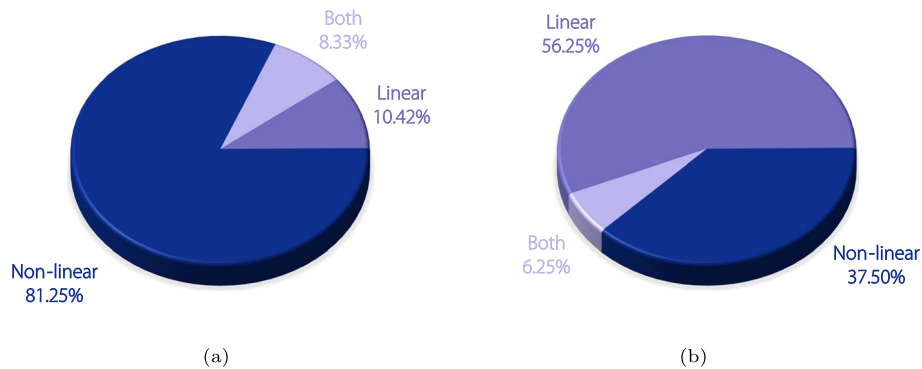


Fig. 8 Comparative analysis of deep learning-based method using different transformation types. Pie chart (a) illustrates the distribution of different transformation types in general medical image registration, while pie chart (b) displays the distribution in retinal image registration

sequence of images. Furthermore, the imaging target itself differs, while CT and MR are often used to image areas rich in features, such as the chest, abdomen, and brain. Retinal images predominantly focus on the vasculature and the optic disc, which exhibit less distinctive features. Consequently, learning the robust features for retinal image registration using deep learning is inherently challenging.

This survey revealed the scarcity of translation- or transformer-based approaches within the domain of retinal image registration. Notably, the majority of transformer-based studies have been conducted using MRI datasets. This trend can be attributed to the increased availability of public MRI datasets, which offer a wealth of data for research purposes. In addition, the ViT model, a prominent example of a transformer-based architecture, requires a substantially larger dataset to surpass the performance of conventional CNN models. While strategies such as data augmentation and the employment of pretrained models may offer provisional relief to this challenge, the crux of the solution lies in more publicly available data.

Discussion

Challenges in retinal image registration

Lack of public datasets

In artificial intelligence, many tasks rely on competition and public evaluations to make progress. These

challenges offer a comprehensive and impartial platform for researchers to compare the performance, computation time, and robustness of newly designed algorithms. The Learn2Reg challenge, for example, recently focused on registering medical imaging modalities commonly used in the brain, abdomen, and thorax [141]. The datasets currently available for retinal image registration are listed in Table 5. Sufficient public retinal image datasets have not been formed for each modality, nor has there been competition.

Different transformation type used from mainstream medical image registration

Based on articles using deep learning, the proportion of different transformation types used in the general medical image registration and the proportion of each type, specifically in the retinal image registration were calculated, as shown in Fig. 8. It was found that over 80% of studies on general medical image registration employed nonlinear transformations. On the contrary, linear transformation is the most commonly used method in retinal applications. This is because retinal images are primarily captured from a limited area of the retina while other commonly used modalities are 3D images with the subject completely contained in the image. Such difference in transformation types makes it difficult for retinal image registration to learn from mainstream medical image registration.

Poor similarity metric

Similarity metrics are used to optimize the registration network in an unsupervised manner or to evaluate the quality of the registration. The key technical challenge in medical-image registration is the selection and design of the most effective similarity measurement methods. Brightness changes may be the most significant difficulty in unimodal image registration. One of the main obstacles to multimodal image registration is that images from different modalities have different resolutions, contrasts, and luminosities. Therefore, a newly designed similarity metric, or a completely different technical route for multimodal image registration, is urgently required.

Intractable retinopathy

During clinical treatments, most patients experience eye retinopathy; therefore, their retinas may be severely damaged. Small bulges, swellings, or blood may cover the normal fundus and negatively affect photography. Some diseases alter the retinal structure. Most samples in the public datasets are retinal images from ordinary people. However, when used for clinical diagnosis, the retinas of some patients are likely to have retinopathy. In this case, a network trained using normal images does not perform well.

Future scope

In this era of large models, it can be anticipated that a general large model for registration will soon emerge. With the ability to use human-marked point pairs or corresponding mask areas as registration prompts, this model can be trained on higher quality, broader types, and more extensive image registration datasets, allowing for better generalization.

There remain many areas in which retinal image registration can be explored. With multiple imaging modalities, there is a pressing need for multimodal image registration. To address this issue, translation-based and disentangling representation methods may be new approaches. Interestingly, any pioneers attempting transformer-based retinal registration methods that could lead to even greater accuracy was not observed.

Moreover, data scarcity remains a significant challenge; however, this can be overcome through data generation or transfer learning. For instance, the dataset with image pairs could be supplemented through random translation, rotation, brightness, and contrast enhancement using retinal images from other datasets. When employing transfer learning, endoscopic images from other parts of the human body can be trained or virtual datasets can be manually generated and fine-tuned for retinal image registration.

Conclusions

This study thoroughly analyzed medical image registration, focusing on its application in retinal imaging. The review compares general medical image registration techniques and their adaptation to retinal imaging, highlights gaps in the current research, and provides advice on avenues for future research. State-of-the-art medical image registration methods were also evaluated and the advantages and disadvantages of each method. Finally, challenges specific to retinal registration were identified and potential opportunities for further advancement discussed.

Abbreviations

CAD	Computer-aided diagnosis
CNN	Convolutional neural network
GAN	Generative adversarial network
AMD	Age-related macular degeneration
CNV	Choroidal neovascularization
CF	Color fundus photography
FA	Fluorescein angiography
OCT	Optical coherence tomography
OCTA	Optical coherence tomography angiography
(R)MSE	(Root) mean square error
SR	Success rate
DSC	Dice similarity coefficient
DR	Diabetic retinopathy
CC	Cross-correlation
MI	Mutual information
SSD	Sum of squared difference
DDPM	Denoising diffusion probabilistic model
VIT	Vision transformer
SIFT	Scale-invariant feature transform
ORB	Oriented FAST and rotated BRIEF
STL	Spatial transformer layer
i2i	Image-to-image
FCN	Fully convolutional networks

Acknowledgements

Not applicable.

Authors' contributions

All the authors make substantial contribution in this manuscript. QN, YH, and MG participated in writing the first draft of the paper; Then the paper was revised carefully and completed the final manuscript by QN and XZ; JL supervised the study. All the authors have read and approved the final manuscript.

Funding

This work was supported in part by General Program of National Natural Science Foundation of China, Nos. 82102189 and 82272086; and Guangdong Provincial Department of Education, No. SJZLGC202202.

Availability of data and materials

All relevant data and material are presented in the main paper.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 26 March 2024 Accepted: 31 July 2024

Published online: 21 August 2024

References

- Oliveira FPM, Tavares JMRS (2014) Medical image registration: a review. *Comput Methods Biomech Biomed Eng* **17**(2):73–93. <https://doi.org/10.1080/10255842.2012.670855>
- Zitova B, Flusser J (2003) Image registration methods: a survey. *Image Vision Comput* **21**(11):977–1000. [https://doi.org/10.1016/S0262-8856\(03\)00137-9](https://doi.org/10.1016/S0262-8856(03)00137-9)
- Boveiri HR, Khayami R, Javidan R, Mehdizadeh A (2020) Medical image registration using deep neural networks: a comprehensive review. *Comput Electr Eng* **87**:106767. <https://doi.org/10.1016/j.compeleceng.2020.106767>
- Haskins G, Kruger U, Yan PK (2020) Deep learning in medical image registration: a survey. *Machine Vision Appl* **31**(1):18. <https://doi.org/10.1007/s00138-020-01060-x>
- Bharati S, Mondal MRH, Podder P, Prasath VBS (2022) Deep learning for medical image registration: a comprehensive review. arXiv preprint arXiv:2204.11341.
- Mokwa NF, Ristau T, Keane PA, Kirchoff B, Sadda SR, Liakopoulos S (2013) Grading of age-related macular degeneration: comparison between color fundus photography, fluorescein angiography, and spectral domain optical coherence tomography. *J Ophthalmol* **2013**:385915. <https://doi.org/10.1155/2013/385915>
- de Carlo TE, Chin AT, Bonini Filho MA, Adhi M, Branchini L, Salz DA et al (2015) Detection of microvascular changes in eyes of patients with diabetes but not clinical diabetic retinopathy using optical coherence tomography angiography. *Retina* **35**(11):2364–2370. <https://doi.org/10.1097/IAE.0000000000000882>
- Frost S, Kanagasingam Y, Sohrabi H, Vignarajan J, Bourgeat P, Salvado O et al (2013) Retinal vascular biomarkers for early detection and monitoring of alzheimer's disease. *Transl Psychiatry* **3**(2):e233. <https://doi.org/10.1038/tp.2012.150>
- Wong TY, Klein R, Sharrett AR, Duncan BB, Couper DJ, Tielsch JM et al (2002) Retinal arteriolar narrowing and risk of coronary heart disease in men and women: the atherosclerosis risk in communities study. *JAMA* **287**(9):1153–1159. <https://doi.org/10.1001/jama.287.9.1153>
- Zhang XQ, Hu Y, Xiao ZJ, Fang JS, Higashita R, Liu J (2022) Machine learning for cataract classification/grading on ophthalmic imaging modalities: a survey. *Mach Intell Res* **19**(3):184–208. <https://doi.org/10.1007/s11633-022-1329-0>
- Hoque ME, Kipli K (2021) Deep learning in retinal image segmentation and feature extraction: a review. *Int J Online Biomed Eng* **17**(14):103–118. <https://doi.org/10.3991/ijoe.v17i14.24819>
- Saha SK, Xiao D, Bhuiyan A, Wong TY, Kanagasingam Y (2019) Color fundus image registration techniques and applications for automated analysis of diabetic retinopathy progression: a review. *Biomed Signal Process Control* **47**:288–302. <https://doi.org/10.1016/j.bspc.2018.08.034>
- Pan LJ, Chen XJ (2021) Retinal OCT image registration: methods and applications. *IEEE Rev Biomed Eng* **16**:307–318. <https://doi.org/10.1109/rbme.2021.3110958>
- Khalifa F, Beache GM, Gimel'farb G, Suri JS, El-Baz AS (2011) State-of-the-art medical image registration methodologies: a survey. In: El-Baz AS, Acharya UR, Mirmehdi M, Suri JS (eds) Multi modality state-of-the-art medical image segmentation and registration methodologies. Springer, Heidelberg, pp 235–280. https://doi.org/10.1007/978-1-4419-8195-0_9
- Besenczi R, Tóth J, Hajdu A (2016) A review on automatic analysis techniques for color fundus photographs. *Comput Struct Biotechnol J* **14**:371–384. <https://doi.org/10.1016/j.csbj.2016.10.001>
- Abràmoff MD, Garvin MK, Sonka M (2010) Retinal imaging and image analysis. *IEEE Rev Biomed Eng* **3**:169–208. <https://doi.org/10.1109/RBME.2010.2084567>
- Baek J, Lee MY, Kim B, Choi A, Kim J, Kwon H et al (2021) Ultra-widefield fluorescein angiography findings in patients with macular edema following cataract surgery. *Ocul Immunol Inflammation* **29**(3):610–614. <https://doi.org/10.1080/09273948.2019.1691739>
- Kornblau IS, El-Annan JF (2019) Adverse reactions to fluorescein angiography: A comprehensive review of the literature. *Surv Ophthalmol* **64**(5):679–693. <https://doi.org/10.1016/j.survophthal.2019.02.004>
- Podoleanu AG (2012) Optical coherence tomography. *J Microsc* **247**(3):209–219. <https://doi.org/10.1111/j.1365-2818.2012.03619.x>
- Ang BCH, Lim SY, Dorairaj S (2020) Intra-operative optical coherence tomography in glaucoma surgery—a systematic review. *Eye* **34**(1):168–177. <https://doi.org/10.1038/s41433-019-0689-3>
- Grewal DS, Carrasco-Zevallos OM, Gunther R, Izatt JA, Toth CA, Hahn P (2017) Intra-operative microscope-integrated swept-source optical coherence tomography guided placement of argus II retinal prosthesis. *Acta Ophthalmol* **95**(5):e431–e432. <https://doi.org/10.1111/aos.13123>
- Werner AC, Shen LQ (2019) A review of OCT angiography in glaucoma. *Semin Ophthalmol* **34**(4):279–286. <https://doi.org/10.1080/08820538.2019.1620807>
- Shaikh NF, Vohra R, Balaji A, Azad SV, Chawla R, Kumar V et al (2021) Role of optical coherence tomography-angiography in diabetes mellitus: utility in diabetic retinopathy and a comparison with fluorescein angiography in vision threatening diabetic retinopathy. *Indian J Ophthalmol* **69**(11):3218–3224. https://doi.org/10.4103/ijjo.IJO_1267_21
- Hernandez-Matas C, Zabalus X, Triantafyllou A, Anyfanti P, Douma S, Argyros AA (2017) FIRE: Fundus Image Registration Dataset. *J Model Ophthalmol* **1**(4):16–28. <https://doi.org/10.35119/maio.v1i4.42>
- Alipour SHM, Rabbani H, Akhlaghi MR (2012) Diabetic retinopathy grading by digital curvelet transform. *Comput Math Methods Med* **2012**:761901. <https://doi.org/10.1155/2012/761901>
- Mooney P (2017) Retinal OCT Images (optical coherence tomography). <https://www.kaggle.com/datasets/paultimothymooney/kernany2018>. Accessed 25 Feb 2024
- Li MC, Chen YR, Ji ZX, Xie KR, Yuan ST, Chen Q et al (2020) Image projection network: 3D to 2D image segmentation in OCTA images. *IEEE Trans Med Imaging* **39**(11):3343–3354. <https://doi.org/10.1109/TMI.2020.2992244>
- Beg MF, Miller MI, Trounev A, Younes L (2005) Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int J Comput Vision* **61**(2):139–157. <https://doi.org/10.1023/b:visi.0000043755.93987.a>
- Lange A, Heldmann S (2020) Multilevel 2D-3D intensity-based image registration. In: Špiclin Ž, McClelland J, Kybic J, Goksel O (eds) Biomedical image registration. 9th international workshop, WBIR 2020, Portorož, Slovenia, December 2020. Lecture notes in computer science, vol 12120. Springer, Heidelberg, pp 57–66. https://doi.org/10.1007/978-3-030-50120-4_6
- Öfverstedt J, Lindblad J, Sladoje N (2019) Fast and robust symmetric image registration based on distances combining intensity and spatial information. *IEEE Trans Image Process* **28**(7):3584–3597. <https://doi.org/10.1109/TIP.2019.2899947>
- Castillo E (2019) Quadratic penalty method for intensity-based deformable image registration and 4DCT lung motion recovery. *Med Phys* **46**(5):2194–2203. <https://doi.org/10.1002/mp.13457>
- Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* **60**(2):91–110. <https://doi.org/10.1023/b:visi.0000029664.99615.94>
- Bay H, Tuytelaars T, Van Gool L (2006) SURF: Speeded up robust features. In: Leonardis A, Bischof H, Pinz A (eds) Computer vision - ECCV 2006. 9th European conference on computer vision, Graz, Austria, May 2006. Lecture notes in computer science, vol 3951. Springer, Heidelberg, pp 404–417. https://doi.org/10.1007/11744023_32
- Ke Y, Sukthankar R (2004) PCA-SIFT: A more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, IEEE, Washington, 27 June–2 July 2004. <https://doi.org/10.1109/cvpr.2004.1315206>
- Tola E, Lepetit V, Fua P (2008) A fast local descriptor for dense matching. In: Proceedings of 2008 IEEE conference on computer vision and pattern recognition, IEEE, Anchorage, 23–28 June 2008. <https://doi.org/10.1109/cvpr.2008.4587673>
- Ruble E, Rabaud V, Konolige K, Bradski G (2011) ORB: An efficient alternative to SIFT or SURF. In: Proceedings of 2011 international conference on computer vision, IEEE, Barcelona, 6–13 November 2011. <https://doi.org/10.1109/ICCV.2011.6126544>
- Cai GR, Jodoin PM, Li SZ, Wu YD, Su SZ, Huang ZK (2013) Perspective-SIFT: an efficient tool for low-altitude remote sensing image registration. *Signal Process* **93**(11):3088–3110. <https://doi.org/10.1016/j.sigpro.2013.04.008>

38. Rosten E, Drummond T (2006) Machine learning for high-speed corner detection. In: Leonardis A, Bischof H, Pinz A (eds) *Computer Vision – ECCV 2006*. 9th European conference on computer vision, Graz, Austria, May 2006. Lecture notes in computer science, vol 3951. Springer, Heidelberg, pp 430–443. https://doi.org/10.1007/11744023_34
39. Calonder M, Lepetit V, Strecha C, Fua P (2010) BRIEF: binary robust independent elementary features. In: Daniilidis K, Maragos P, Paragios N (eds) *Computer vision – ECCV 2010*. 11th European conference on computer vision, Heraklion, Crete, Greece, September 2010. Lecture notes in computer science, vol 6314. Springer, Heidelberg, pp 778–792. https://doi.org/10.1007/978-3-642-15561-1_56
40. Canny J (1986) A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* **8**(6):679–698. <https://doi.org/10.1016/b978-0-08-051581-6.50024-6>
41. Marr D, Hildreth E (1980) Theory of edge detection. *Proc Roy Soc B: Biol Sci* **207**(1167):187–217. <https://doi.org/10.1098/rspb.1980.0020>
42. Pal NR, Pal SK (1993) A review on image segmentation techniques. *Pattern Recognit* **26**(9):1277–1294. [https://doi.org/10.1016/0031-3203\(93\)90135-J](https://doi.org/10.1016/0031-3203(93)90135-J)
43. Hesamian MH, Jia WJ, He XJ, Kennedy P (2019) Deep learning techniques for medical image segmentation: achievements and challenges. *J Digit Imaging* **32**(4):582–596. <https://doi.org/10.1007/s10278-019-00227-x>
44. Miao S, Wang ZJ, Zheng YF, Liao R (2016) Real-time 2D/3D registration via CNN regression. In: *Proceedings of 2016 IEEE 13th international symposium on biomedical imaging, IEEE, Prague, 13–16 April 2016*. <https://doi.org/10.1109/isbi.2016.7493536>
45. Yang X, Kwitt R, Styner M, Niethammer M (2017) Quicksilver: fast predictive image registration - a deep learning approach. *NeuroImage* **158**:378–396. <https://doi.org/10.1016/j.neuroimage.2017.07.008>
46. Cao XH, Yang JH, Zhang J, Nie D, Kim M, Wang Q et al (2017) Deformable image registration based on similarity-steered CNN regression. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins D, Duchesne S (eds) *Medical image computing and computer assisted intervention - MICCAI 2017*. 20th international conference, Quebec City, QC, Canada, September 2017. Lecture notes in computer science, vol 10433. Springer, Heidelberg, pp 300–308. https://doi.org/10.1007/978-3-319-66182-7_35
47. de Vos BD, Berendsen FF, Viergever MA, Staring M, Išgum I (2017) End-to-end unsupervised medical image registration with a convolutional neural network. In: Cardoso MJ, Arbel T, Carneiro G, Syeda-Mahmood T, Tavares JMRS, et al (eds) *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Third international workshop, DLMIA 2017, and 7th international workshop, ML-CDS 2017, held in conjunction with MICCAI 2017, Québec City, QC, Canada, September. Lecture notes in computer science, vol 10553. Springer, Heidelberg, pp 204–212. https://doi.org/10.1007/978-3-319-67558-9_24
48. Zheng JN, Miao S, Wang ZJ, Liao R (2018) Pairwise domain adaptation module for CNN-based 2-D/3-D registration. *J Med Imaging* **5**(2):021204. <https://doi.org/10.1117/1.jmi.5.2.021204>
49. Sloan JM, Goatman KA, Siebert JP (2018) Learning rigid image registration-utilizing convolutional neural networks for medical image registration. In: *Proceedings of the 11th international joint conference on biomedical engineering systems and technologies, SciTePress, Funchal, 19–21 January 2018*. <https://doi.org/10.5220/0006543700890099>
50. Chee E, Wu ZZ (2018) AIRNet: self-supervised affine registration for 3D medical images using neural networks. *arXiv preprint arXiv: 1810.02583*
51. Lv J, Yang M, Zhang J, Wang XY (2018) Respiratory motion correction for free-breathing 3D abdominal MRI using CNN-based image registration: a feasibility study. *Br J Radiol* **91**(1083):20170788. <https://doi.org/10.1259/bjr.20170788>
52. Hu YP, Modat M, Gibson E, Li WQ, Ghavami N, Bonmati E et al (2018) Weakly-supervised convolutional neural networks for multimodal image registration. *Med Image Anal* **49**:1–13. <https://doi.org/10.1016/j.media.2018.07.002>
53. Jiang PG, Shackelford JA (2018) CNN driven sparse multi-level B-spline image registration. In: *Proceedings of 2018 IEEE/CVF conference on computer vision and pattern recognition, IEEE, Salt Lake City, 18–23 June 2018*. <https://doi.org/10.1109/cvpr.2018.00967>
54. Li HM, Fan Y (2018) Non-rigid image registration using self-supervised fully convolutional networks without training data. In: *Proceedings of 2018 IEEE 15th international symposium on biomedical imaging, IEEE, Washington, 4–7 April 2018*. <https://doi.org/10.1109/isbi.2018.8363757>
55. Fan JF, Cao XH, Yap PT, Shen DG (2019) BIRNet: brain image registration using dual-supervised fully convolutional networks. *Med Image Anal* **54**:193–206. <https://doi.org/10.1016/j.media.2019.03.006>
56. Xu ZL, Niethammer M (2019) DeepAtlas: joint semi-supervised learning of image registration and segmentation. In: Shen DG, Liu TM, Peters TM, Staib LH, Essert C, Zhou SA et al (eds) *Medical image computing and computer assisted intervention - MICCAI 2019*. 22nd international conference, Shenzhen, China, October 2019. Lecture notes in computer science, vol 11765. Springer, Heidelberg. https://doi.org/10.1007/978-3-030-32245-8_47
57. de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Išgum I (2019) A deep learning framework for unsupervised affine and deformable image registration. *Med Image Anal* **52**:128–143. <https://doi.org/10.1016/j.media.2018.11.010>
58. Zhao SY, Lau T, Luo J, Chang EIC, Xu Y (2020) Unsupervised 3D end-to-end medical image registration with volume twinning network. *IEEE J Biomed Health Inform* **24**(5):1394–1404. <https://doi.org/10.1109/JBHI.2019.2951024>
59. Zhao SY, Dong Y, Chang EIC, Xu Y (2019) Recursive cascaded networks for unsupervised medical image registration. In: *Proceedings of 2019 IEEE/CVF international conference on computer vision, IEEE, Seoul, 27 October–2 November 2019*. <https://doi.org/10.1109/iccv.2019.01070>
60. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV (2019) VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans Med Imaging* **38**(8):1788–1800. <https://doi.org/10.1109/tmi.2019.2897538>
61. Dalca AV, Balakrishnan G, Guttag J, Sabuncu MR (2019) Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Med Image Anal* **57**:226–236. <https://doi.org/10.1016/j.media.2019.07.006>
62. Hu XJ, Kang M, Huang WL, Scott MR, Wiest R, Reyes M (2019) Dual-stream pyramid registration network. In: Shen DG, Liu TM, Peters TM, Staib LH, Essert C, Zhou SA et al (eds) *Medical image computing and computer assisted intervention - MICCAI 2019*. 22nd international conference, Shenzhen, China, October 2019. Lecture notes in computer science, vol 11765. Springer, Heidelberg, pp 382–390. https://doi.org/10.1007/978-3-030-32245-8_43
63. Wang J, Zhang MM (2020) DeepFLASH: an efficient network for learning-based medical image registration. In: *Proceedings of 2020 IEEE/CVF conference on computer vision and pattern recognition, IEEE, Seattle, 13–19 June 2020*. <https://doi.org/10.1109/cvpr42600.2020.00450>
64. Mansilla L, Milone DH, Ferrante E (2020) Learning deformable registration of medical images with anatomical constraints. *Neural Netw* **124**:269–279. <https://doi.org/10.1016/j.neunet.2020.01.023>
65. Mok TCW, Chung ACS (2020) Fast symmetric diffeomorphic image registration with convolutional neural networks. In: *Proceedings of 2020 IEEE/CVF conference on computer vision and pattern recognition, IEEE, Seattle, 13–19 June 2020*. <https://doi.org/10.1109/cvpr42600.2020.00470>
66. Kim B, Kim DH, Park SH, Kim J, Lee JG, Ye JC (2021) CycleMorph: cycle consistent unsupervised deformable image registration. *Med Image Anal* **71**:102036. <https://doi.org/10.1016/j.media.2021.102036>
67. Czolbe S, Krause O, Feragen A (2021) Semantic similarity metrics for learned image registration. In: *Proceedings of the medical imaging with deep learning, PMLR, Lübeck, 7–9 July 2021*. <https://doi.org/10.1016/j.media.2023.102830>
68. Mok TCW, Chung ACS (2022) Robust image registration with absent correspondences in pre-operative and follow-up brain MRI scans of diffuse glioma patients. In: Bakas S, Crimi A, Baid U, Malec S, Pytlarz M, Baheti B et al (eds) *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries*. 8th International workshop, BrainLes 2022, held in conjunction with MICCAI 2022, Singapore, September 2022. Lecture notes in computer science, vol 13769. Springer, Heidelberg, pp 231–240. https://doi.org/10.1007/978-3-031-33842-7_20
69. Kang M, Hu XJ, Huang WL, Scott MR, Reyes M (2022) Dual-stream pyramid registration network. *Med Image Anal* **78**:102379. <https://doi.org/10.1016/j.media.2022.102379>

70. Tran, MQ, Do, T, Tran, H, Tjiputra, E, Tran, QD, Nguyen, A (2022) Light-weight deformable registration using adversarial learning with distilling knowledge. *IEEE Trans Med Imaging* **41**(6):1443–1453. <https://doi.org/10.1109/tmi.2022.3141013>
71. Kong LK, Qi XS, Shen QJ, Wang JC, Zhang JY, Hu YL et al (2023) Indescribable multi-modal spatial evaluator. In: Proceedings of 2023 IEEE/CVF conference on computer vision and pattern recognition, IEEE, Vancouver, 17–24 June 2023. <https://doi.org/10.1109/cvpr52729.2023.00950>
72. Che TT, Wang XY, Zhao K, Zhao Y, Zeng DB, Li QL et al (2023) AMNet: Adaptive multi-level network for deformable registration of 3D brain MR images. *Med Image Anal* **85**:102740. <https://doi.org/10.1016/j.media.2023.102740>
73. Zagoruyko S, Komodakis N (2015) Learning to compare image patches via convolutional neural networks. In: Proceedings of 2015 IEEE conference on computer vision and pattern recognition, IEEE, Boston, 7–12 June 2015. <https://doi.org/10.1109/cvpr.2015.7299064>
74. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention - MICCAI 2015. 18th International conference, Munich, Germany, October 2015. Lecture notes in computer science, vol 9351. Springer, Heidelberg, pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
75. Jaderberg M, Simonyan K, Zisserman A (2015) Spatial transformer networks. In: Proceedings of the 28th international conference on neural information processing systems, MIT Press, Montreal, 7–12 December 2015. <https://doi.org/10.48550/arXiv.1506.02025>
76. Mahapatra D, Antony B, Sedai S, Garnavi R (2018) Deformable medical image registration using generative adversarial networks. In: Proceedings of 2018 IEEE 15th international symposium on biomedical imaging, IEEE, Washington, 4–7 April 2018. <https://doi.org/10.1109/isbi.2018.8363845>
77. Qin C, Shi BB, Liao R, Mansi T, Rueckert D, Kamen A (2019) Unsupervised deformable registration for multi-modal images via disentangled representations. In: Chung ACS, Gee JC, Yushkevich PA, Bao SQ (eds) Information processing in medical imaging. 26th International conference, IPMI 2019, Hong Kong, China, June 2019. Lecture notes in computer science, vol 11492. Springer, Heidelberg, pp 249–261. https://doi.org/10.1007/978-3-030-20351-1_19
78. Xu Z, Luo J, Yan JP, Pulya R, Li X, Wells W, et al (2020) Adversarial uni-and multi-modal stream networks for multimodal image registration. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, et al (eds) Medical image computing and computer assisted intervention - MICCAI 2020. 23rd international conference, Lima, Peru, October 2020. Lecture notes in computer science, vol 12263. Springer, Heidelberg, pp 222–232. https://doi.org/10.1007/978-3-030-59716-0_22
79. Han R, Jones CK, Lee J, Wu P, Vagdari P, Uneri A, et al (2022) Deformable mr-ct image registration using an unsupervised, dual-channel network for neurosurgical guidance. *Med Image Anal* **75**:102292. <https://doi.org/10.1016/j.media.2021.102292>
80. Zhang JJ, Fu TY, Wang YY, Li JS, Xiao DQ, Fan JF, et al (2023) An alternately optimized generative adversarial network with texture and content constraints for deformable registration of 3D ultrasound images. *Phys Med Biol* **68**(14):145006. <https://doi.org/10.1088/1361-6560/ace098>
81. Casamitjana A, Mancini M, Iglesias JE (2021) Synth-by-Reg (SbR): contrastive learning for synthesis-based registration of paired images. In: Svoboda D, Burgos N, Wolterink JM, Zhao C (eds) Simulation and synthesis in medical imaging. 6th international workshop, SASHIMI 2021, held in conjunction with MICCAI 2021, Strasbourg, France, September 2021. Lecture notes in computer science, vol 12965. Springer, Heidelberg, pp 44–54. https://doi.org/10.1007/978-3-030-87592-3_5
82. Chen ZK, Wei J, Li R (2022) Unsupervised multi-modal medical image registration via discriminator-free image-to-image translation. In: Proceedings of the thirty-first international joint conference on artificial intelligence, ijcai.org, Vienna, 23–29 July 2022. <https://doi.org/10.24963/ijcai.2022/117>
83. Kim B, Han I, Ye JC (2022) DiffuseMorph: unsupervised deformable image registration using diffusion model. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (eds) Computer vision - ECCV 2022. 17th European conference, Tel Aviv, Israel, October 2022. Lecture notes in computer science, vol 13691. Springer, Heidelberg, pp 347–364. https://doi.org/10.1007/978-3-031-19821-2_20
84. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S et al (2014) Generative adversarial nets. In: Proceedings of the 27th international conference on neural information processing systems, MIT Press, Montreal, 8–13 December 2014. https://doi.org/10.1007/978-3-658-40442-0_9
85. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of 2017 IEEE international conference on computer vision, IEEE, Venice, 22–29 October 2017. <https://doi.org/10.1109/iccv.2017.244>
86. Park T, Efros AA, Zhang R, Zhu JY (2020) Contrastive learning for unpaired image-to-image translation. In: Vedaldi A, Bischof H, Brox T, Frahm JM (eds) Computer vision - ECCV 2020. 16th European conference, Glasgow, UK, August 2020. Lecture notes in computer science, vol 12354. pp 319–345. Springer, Heidelberg. https://doi.org/10.1007/978-3-030-58545-7_19
87. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Proceedings of the 34th international conference on neural information processing systems, Curran Associates Inc., Vancouver, 6–12 December 2020. <https://doi.org/10.1109/powertech55446.2023.10202713>
88. Dhariwal P, Nichol AQ (2021) Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**:8780–8794. <https://doi.org/10.5555/3540261.3540933>
89. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the 9th international conference on learning representations, OpenReview.net, 3–7 May 2021.
90. Chen JY, He YF, Frey EC, Li Y, Du Y (2021) ViT-V-Net: vision transformer for unsupervised volumetric medical image registration. arXiv preprint [arXiv: 2104.06468](https://arxiv.org/abs/2104.06468)
91. Zhang YG, Pei YR, Zha HB (2021) Learning dual transformer network for diffeomorphic registration. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng YF et al (eds) Medical image computing and computer assisted intervention - MICCAI 2021. 24th international conference, Strasbourg, France, September–October 2021. Lecture notes in computer science, vol 12904. Springer, Heidelberg, pp 129–138. https://doi.org/10.1007/978-3-030-87202-1_13
92. Mok TCW, Chung ACS (2022) Affine medical image registration with coarse-to-fine vision transformer. In: Proceedings of 2022 IEEE/CVF conference on computer vision and pattern recognition, IEEE, New Orleans, 18–24 June 2022. <https://doi.org/10.1109/cvpr52688.2022.02017>
93. Chen JY, Frey EC, He YF, Segars WP, Li Y, Du Y (2022) TransMorph: transformer for unsupervised medical image registration. *Med Image Anal* **82**:102615. <https://doi.org/10.1016/j.media.2022.102615>
94. Song L, Liu GX, Ma MR (2022) TD-Net: unsupervised medical image registration network based on transformer and CNN. *Appl Intell* **52**(15):18201–18209. <https://doi.org/10.1007/s10489-022-03472-w>
95. Wang YB, Qian W, Li MQ, Zhang XM (2022) A transformer-based network for deformable medical image registration. In: Fang L, Povey D, Zhai GT, Mei T, Wang RP (eds) Artificial intelligence. Second CAAI international conference, CICA 2022, Beijing, China, August 2022. Lecture notes in computer science, vol 13604. Springer, Heidelberg, pp 502–513. https://doi.org/10.1007/978-3-031-20497-5_41
96. Shi JC, He YT, Kong YY, Coatrieux JL, Shu HZ, Yang GY, et al (2022) XMorpher: full transformer for deformable medical image registration via cross attention. In: Wang LW, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer assisted intervention - MICCAI 2022. 25th international conference, Singapore, September 2022. Lecture notes in computer science, vol 13436. Springer, Heidelberg, pp 217–226. https://doi.org/10.1007/978-3-031-16446-0_21
97. Zhu YP, Lu S (2022) Swin-VoxelMorph: a symmetric unsupervised learning model for deformable medical image registration using swin transformer. In: Wang LW, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer assisted intervention - MICCAI 2022. 25th international conference, Singapore, September 2022. Lecture notes in computer science, vol 13436. Springer, Heidelberg, pp 78–87. https://doi.org/10.1007/978-3-031-16446-0_8

98. Chen ZY, Zheng YJ, Gee JC (2023) TransMatch: a transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. *IEEE Trans Med Imaging* **43**(1):15–27. <https://doi.org/10.1109/tmi.2023.3288136>
99. Wang HQ, Ni D, Wang Y (2023) ModeT: learning deformable image registration via motion decomposition transformer. In: Greenspan H, Madabhushi A, Mousavi P, Salcudean S, Duncan J, Syeda-Mahmood T, et al (eds) Medical image computing and computer assisted intervention - MICCAI 2023. 26th international conference, Vancouver, BC, Canada, October 2023. Lecture notes in computer science, vol 14229. Springer, Heidelberg, pp 740–749. https://doi.org/10.1007/978-3-031-43999-5_70
100. Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z et al (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of 2021 IEEE/CVF international conference on computer vision, IEEE, Montreal, 10–17 October 2021, pp 10012–10022. <https://doi.org/10.1109/iccv48922.2021.00986>
101. Nazib A, Fookes C, Perrin D (2018) A comparative analysis of registration tools: traditional vs deep learning approach on high resolution tissue cleared data. arXiv preprint [arXiv: 1810.08315](https://arxiv.org/abs/1810.08315)
102. Legg PA, Rosin PL, Marshall D, Morgan JE (2013) Improving accuracy and efficiency of mutual information for multi-modal retinal image registration using adaptive probability density estimation. *Comput Med Imaging Graph* **37**(7–8):597–606. <https://doi.org/10.1016/j.compmedimag.2013.08.004>
103. Reel PS, Dooley LS, Wong KCP, Börner A (2013) Robust retinal image registration using expectation maximisation with mutual information. In: Proceedings of 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, Vancouver, 26–31 May 2013. <https://doi.org/10.1109/icassp.2013.6637824>
104. Reel PS, Dooley LS, Wong KCP, Börner A (2014) Enhanced retinal image registration accuracy using expectation maximisation and variable bin-sized mutual information. In: Proceedings of 2014 IEEE international conference on acoustics, speech and signal processing, IEEE, Florence, 4–9 May 2014. <https://doi.org/10.1109/icassp.2014.6854883>
105. Cideciyan AV (1995) Registration of ocular fundus images: an algorithm using cross-correlation of triple invariant image descriptors. *IEEE Eng Med Biol Mag* **14**(1):52–58. <https://doi.org/10.1109/51.340749>
106. Stewart CV, Tsai CL, Roysam B (2003) The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *IEEE Trans Med Imaging* **22**(11):1379–1394. <https://doi.org/10.1109/TMI.2003.819276>
107. Guo XY, Hsu W, Lee ML, Wong TY (2006) A tree matching approach for the temporal registration of retinal images. In: Proceedings of 2006 18th IEEE international conference on tools with artificial intelligence, IEEE, Arlington, 13–15 November 2006. <https://doi.org/10.1109/ictai.2006.22>
108. Zheng YJ, Hunter AA, Wu J, Wang HZ, Gao JB, Maguire MG et al (2011) Landmark matching based automatic retinal image registration with linear programming and self-similarities. In: Székely G, Hahn HK (eds) Information processing in medical imaging. 22nd international conference, IPMI 2011, Kloster Irsee, Germany, July 2011. Lecture notes in computer science, vol 6801. Springer, Heidelberg, pp 674–685. https://doi.org/10.1007/978-3-642-22092-0_55
109. Zheng YJ, Daniel E, Hunter III AA, Xiao R, Gao JB, Li HS, et al (2014) Landmark matching based retinal image alignment by enforcing sparsity in correspondence matrix. *Med Image Anal* **18**(6):903–913. <https://doi.org/10.1016/j.media.2013.09.009>
110. Hervella AS, Rouco J, Novo J, Ortega M (2018) Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. *Procedia Comput Sci* **126**:97–104. <https://doi.org/10.1016/j.procs.2018.07.213>
111. Koukounis D, Nicholson L, Bull DR, Achim A (2011) Retinal image registration based on multiscale products and optic disc detection. In: Proceedings of 2011 annual international conference of the IEEE engineering in medicine and biology society, IEEE, Boston, 30 August 2011–3 September 2011. <https://doi.org/10.1109/iembs.2011.6091541>
112. Ramli R, Idris MYI, Hasikin K, Karim NKA, Abdul Wahab AW, Ahmady I et al (2017) Feature-based retinal image registration using D-saddle feature. *J Healthc Eng* **2017**:1489524. <https://doi.org/10.1155/2017/1489524>
113. Yang GH, Stewart CV, Sofka M, Tsai CL (2007) Registration of challenging image pairs: Initialization, estimation, and decision. *IEEE Trans Pattern Anal Mach Intell* **29**(11):1973–1989. <https://doi.org/10.1109/TPAMI.2007.1116>
114. Chen J, Tian J, Lee N, Zheng J, Smith RT, Laine AF (2010) A partial intensity invariant feature descriptor for multimodal retinal image registration. *IEEE Trans Biomed Eng* **57**(7):1707–1718. <https://doi.org/10.1109/TBME.2010.2042169>
115. Gharabaghi S, Daneshvar S, Sedaaghi MH (2013) Retinal image registration using geometrical features. *J Digit Imaging* **26**(2):248–258. <https://doi.org/10.1007/s10278-012-9501-7>
116. Ghassabi Z, Shanbehzadeh J, Mohammadzadeh A, Ostadzadeh SS (2015) Colour retinal fundus image registration by selecting stable extremum points in the scale-invariant feature transform detector. *IET Image Process* **9**(10):889–900. <https://doi.org/10.1049/iet-ipr.2014.0907>
117. Saha SK, Xiao D, Frost S, Kanagasam Y (2016) A two-step approach for longitudinal registration of retinal images. *J Med Syst* **40**(12):277. <https://doi.org/10.1007/s10916-016-0640-0>
118. Li QL, Li SY, Wu YJ, Guo W, Qi SW, Huang G et al (2020) Orientation-independent feature matching (OIFM) for multimodal retinal image registration. *Biomed Signal Process Control* **60**:101957. <https://doi.org/10.1016/j.bspc.2020.101957>
119. Lee J, Liu P, Cheng J, Fu H (2019) A deep step pattern representation for multimodal retinal image registration. In: Proceedings of 2019 IEEE/CVF international conference on computer vision, IEEE, Seoul, 27 October 2019–2 November 2019. <https://doi.org/10.1109/iccv.2019.00518>
120. Zhang JK, An C, Dai J, Amador M, Bartsch DU, Borooah S, et al (2019) Joint vessel segmentation and deformable registration on multi-modal retinal images based on style transfer. In: Proceedings of 2019 IEEE international conference on image processing, IEEE, Taipei, China, 22–25 September 2019, pp 839–843. <https://doi.org/10.1109/icip.2019.8802932>
121. De Silva T, Chew EY, Hotaling N, Cukras CA (2021) Deep-learning based multi-modal retinal image registration for longitudinal analysis of patients with age-related macular degeneration. *Biomed Opt Express* **12**(1):619–636. <https://doi.org/10.1364/BOE.408573>
122. Wang YQ, Zhang JK, An C, Cavichini M, Jhingan M, Amador-Patarroyo MJ, et al (2020) A segmentation based robust deep learning framework for multimodal retinal image registration. In: Proceedings of 2020 IEEE international conference on acoustics, speech and signal processing, IEEE, Barcelona, 4–8 May 2020. <https://doi.org/10.1109/icassp40776.2020.9054077>
123. Tian YT, Hu Y, Ma YH, Hao HY, Mou L, Yang JL et al (2020) Multi-scale u-net with edge guidance for multimodal retinal image deformable registration. In: Proceedings of the 42nd annual international conference of the IEEE engineering in medicine & biology society, IEEE, Montreal, 20–24 July 2020. <https://doi.org/10.1109/embc44109.2020.9175613>
124. Zou BJ, He ZY, Zhao RC, Zhu CZ, Liao WM, Li S (2020) Non-rigid retinal image registration using an unsupervised structure-driven regression network. *Neurocomputing* **404**:14–25. <https://doi.org/10.1016/j.neucom.2020.04.122>
125. Wang YQ, Zhang JK, Cavichini M, Bartsch DUG, Freeman WR, Nguyen TQ et al (2021) Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework. *IEEE Trans Image Process* **30**:3167–3178. <https://doi.org/10.1109/tip.2021.3058570>
126. Zhang JK, Wang YQ, Dai J, Cavichini M, Bartsch DUG, Freeman WR et al (2021) Two-step registration on multi-modal retinal images via deep neural networks. *IEEE Trans Image Process* **31**:823–838. <https://doi.org/10.1109/tip.2021.3135708>
127. Sui XD, Zheng YJ, Jiang YY, Jiao WZ, Ding YH (2021) Deep multispectral image registration network. *Comput Med Imaging Graph* **87**:101815. <https://doi.org/10.1016/j.compmedimag.2020.101815>
128. An C, Wang YQ, Zhang JK, Nguyen TQ (2022) Self-supervised rigid registration for multimodal retinal images. *IEEE Trans Image Process* **31**:5733–5747. <https://doi.org/10.1109/tip.2022.3201476>
129. Benvenuto GA, Colnago M, Casaca W (2022) Unsupervised deep learning network for deformable fundus image registration. In: Proceedings of 2022 IEEE international conference on acoustics, speech and signal processing, IEEE, Singapore, 23–27 May 2022. <https://doi.org/10.1109/icassp43922.2022.9747686>
130. López-Varela E, Novo J, Fernández-Vigo JI, Moreno-Morillo FJ, Ortega M (2022) Unsupervised deformable image registration in a landmark scarcity scenario: choroid octa. In: Sclaroff S, Distante C, Leo M, Farinella GM, Tombari F (eds) Image analysis and processing - ICIAP 2022. 21st international conference, Lecce, Italy, May 2022. Lecture notes in computer science, vol 13231. Springer, Heidelberg, pp. 89–99. https://doi.org/10.1007/978-3-031-06427-2_8

131. Rivas-Villar D, Hervella ÁS, Rouco J, Novo J (2022) Color fundus image registration using a learning-based domain-specific landmark detection methodology. *Comput Biol Med* **140**:105101. <https://doi.org/10.1016/j.combiomed.2021.105101>
132. Kim GY, Kim JY, Lee SH, Kim SM (2022) Robust detection model of vascular landmarks for retinal image registration: A two-stage convolutional neural network. *Biomed Res Int* **2022**:1705338. <https://doi.org/10.1155/2022/1705338>
133. Santarossa M, Kilic A, von der Burchard C, Schmarje L, Zelenka C, Reinhold S et al (2022) Medregnet: unsupervised multimodal retinal-image registration with gans and ranking loss. In: *Proceedings of SPIE 12032, medical imaging 2022: image processing*, SPIE, San Diego, 4 April 2022. <https://doi.org/10.1117/1.2.2607653>
134. Liu JZ, Li XR, Wei QJ, Xu J, Ding DY (2022) Semi-supervised keypoint detector and descriptor for retinal image matching. In: Shai A, Gabriel B, Moustapha C, Giovanni MF, Tal H (eds) *Computer vision - ECCV 2022. 17th European conference on computer vision*, Tel Aviv, Israel, October 2022. Springer, Heidelberg, pp 593–609. https://doi.org/10.1007/978-3-031-19803-8_35
135. Rivas-Villar D, Motschi AR, Pircher M, Hitznberger CK, Schranz M, Roberts PK et al (2023) Automated inter-device 3d oct image registration using deep learning and retinal layer segmentation. *Biomed Opt Express* **14**(7):3726–3747. <https://doi.org/10.1364/boe.493047>
136. Liu JZ, Li XR (2023) Geometrized transformer for self-supervised homography estimation. In: *Proceedings of 2023 international conference on computer vision, ICCC, Paris, 2-6 October 2023*. <https://doi.org/10.1109/icc51070.2023.00876>
137. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* **24**(6):381–395. <https://doi.org/10.1016/b978-0-08-051581-6.50070-2>
138. DeTone D, Malisiewicz T, Rabinovich A (2018) SuperPoint: self-supervised interest point detection and description. In: *Proceedings of 2018 IEEE/CVF conference on computer vision and pattern recognition workshops, ICCVW, Salt Lake City, 18-22 June 2018*. <https://doi.org/10.1109/cvprw.2018.00060>
139. Kamran SA, Fariha Hossain K, Tavakkoli A, Zuckerbrod S, Baker SA, Sanders KM (2020) Fundus2Angio: a conditional gan architecture for generating fluorescein angiography images from retinal fundus photography. In: *Advances in visual computing. 15th international symposium, ISVC 2020, San Diego, CA, USA, October 2020. Lecture notes in computer science, vol 12510*. Springer, Heidelberg, pp 125–138. https://doi.org/10.1007/978-3-030-64559-5_10
140. Andreini P, Ciano G, Bonechi S, Graziani C, Lachi V, Mecocci A et al (2021) A two-stage gan for high-resolution retinal image generation and segmentation. *Electronics* **11**(1):60. <https://doi.org/10.3390/electronics11010060>
141. Hering A, Hansen L, Mok TCW, Chung ACS, Siebert H, Häger S et al (2022) Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Trans Med Imaging*, **42**(3):697–712. <https://doi.org/10.1109/TMI.2022.3213983>
142. Ding L, Kang TD, Kuriyan AE, Ramchandran RS, Wykoff CC, Sharma G (2023) Combining feature correspondence with parametric chamfer alignment: Hybrid two-stage registration for ultra-widefield retinal images. *IEEE Trans Biomed Eng* **70**(2):523–532. <https://doi.org/10.1109/TBME.2022.3196458>
143. Wykoff CC, Nittala MG, Zhou B, Fan WY, Velaga SB, Lampen SIR, et al (2019) Intravitreal aflibercept for retinal nonperfusion in proliferative diabetic retinopathy: Outcomes from the randomized RECOVERY trial. *Ophthalmol Retina* **3**(12):1076–1086. <https://doi.org/10.1016/j.oret.2019.07.011>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.