Original Article

Joint keypoint detection and description network for color fundus image registration

David Rivas-Villar[1,2]^, Álvaro S. Hervella[1,2]^, José Rouco[1,2]^, Jorge Novo[1,2]^

[1]VARPA Group, A Coruña Biomedical Research Institute (INIBIC), University of A Coruña, Xubias de Arriba, A Coruña, Spain; [2]CITIC Research Centre, University of A Coruña, Campus de Elviña, A Coruña, Spain

Contributions: (I) Conception and design: All authors; (II) Administrative support: J Rouco, J Novo; (III) Provision of study materials or patients: J Rouco, J Novo; (IV) Collection and assembly of data: D Rivas-Villar, ÁS Hervella; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: David Rivas-Villar, MSc. University of A Coruña, Campus de Elviña, s/n, 15071, A Coruña, Spain. Email: david.rivas.villar@udc.es.

Background: Retinal imaging is widely used to diagnose many diseases, both systemic and eye-specific. In these cases, image registration, which is the process of aligning images taken from different viewpoints or moments in time, is fundamental to compare different images and to assess changes in their appearance, commonly caused by disease progression. Currently, the field of color fundus registration is dominated by classical methods, as deep learning alternatives have not shown sufficient improvement over classic methods to justify the added computational cost. However, deep learning registration methods are still considered beneficial as they can be easily adapted to different modalities and devices following a data-driven learning approach.

Methods: In this work, we propose a novel methodology to register color fundus images using deep learning for the joint detection and description of keypoints. In particular, we use an unsupervised neural network trained to obtain repeatable keypoints and reliable descriptors. These keypoints and descriptors allow to produce an accurate registration using RANdom SAmple Consensus (RANSAC). We train the method using the Messidor dataset and test it with the Fundus Image Registration Dataset (FIRE) dataset, both of which are publicly accessible.

Results: Our work demonstrates a color fundus registration method that is robust to changes in imaging devices and capture conditions. Moreover, we conduct multiple experiments exploring several of the method's parameters to assess their impact on the registration performance. The method obtained an overall Registration Score of 0.695 for the whole FIRE dataset (0.925 for category S, 0.352 for P, and 0.726 for A).

Conclusions: Our proposal improves the results of previous deep learning methods in every category and surpasses the performance of classical approaches in category A which has disease progression and thus represents the most relevant scenario for clinical practice as registration is commonly used in patients with diseases for disease monitoring purposes.

Keywords: Medical image registration; deep learning; feature-based registration (FBR); retinal imaging; ophthalmology

Submitted Jan 02, 2023. Accepted for publication Apr 23, 2023. Published online May 26, 2023.
doi: 10.21037/qims-23-4
View this article at: https://dx.doi.org/10.21037/qims-23-4

^ ORCID: David Rivas-Villar, 0000-0001-7824-8098; Álvaro S. Hervella, 0000-0002-9080-9836; José Rouco, 0000-0003-4407-9091; Jorge Novo, 0000-0002-0125-3064.

Introduction

Currently, the registration of medical images is of upmost importance due to the numerous applications it has on clinical practice (1). Image registration is the process in which a pair of images taken under different imaging viewpoints are aligned. This process facilitates the simultaneous analysis of several images. Therefore, clinicians can draw better and more informed conclusions (2). Particularly, medical image registration facilitates the comparison of images taken at different times, which is useful to perform longitudinal studies and monitor disease progression (3). Additionally, image registration is also useful for aligning the images with a candidate disease model (2), which helps to monitor the disease and treatment outcomes. Similarly, automated image registration is often essential for computer-aided diagnosis (CAD) systems (4,5), which cannot be based on manual registration due to the complexity of this task and its time requirements. Therefore, retinal registration methods should desirably be robust against disease lesions so that they can properly help assessing pathological progression. Moreover, image registration is used in other fields of medical imaging as it can allow for image to image translation which is the process of learning a mapping between an input image and an output image. These methods allow to transform color fundus images into fundus angiographies (6,7), a more invasive type of image. Image-to-image translation can be approached using paired methods, such as pix2pix (8,9) requiring paired and registered images, or unpaired methods such as cycle-GAN (10,11).

Image registration for retinal image analysis is especially relevant, as the eyes are the only organs in the human body that allow non-invasive *in vivo* observation of the blood vessels and neuronal tissue (12). Particularly, due to the features of retinal images, multiple types of transformation models are used like similarity, affine or homographic, depending on the image modality. However, it should be noted that local deformations usually do not happen in the eye fundus. Color fundus imaging is very common in current clinical practice due to its wide availability and cost-effectiveness. Examples of color fundus images are shown in *Figure 1*. These images are used to diagnose multiple diseases, such as Diabetic Retinopathy or Age-Related Macular Degeneration (13). However, color fundus images also present several characteristics that complicate the image registration process. Due to the photographic nature of this imaging technique, the produced images are subject to several variations, including spatial shifts, including spatial shifts (patient movement, imperfect device placement...), color or illumination variations, changes in focus, etc. Furthermore, the presence of disease usually alters the appearance of the retina. Image registration is commonly used in patients with diseases to monitor their progression or remission so retinal image registration methods should be robust to these morphological changes. The usual presence of morphological changes due to disease progression or remission in combination with the variance introduced by the image capture process make color fundus image registration a challenging task.

Overall, image registration approaches can be classified according to the methodology employed to align the image pairs. In this sense, there are two different groups, classical methods and deep learning approaches. Furthermore, each of those two groups can be divided into intensity-based (IBR) and feature-based registration (FBR) (14), depending on the information used to register the images. While classical methods are still widely used, the novel deep learning approaches have several advantages over them. For instance, the end-to-end training, from raw data to the expected result, avoids the need for ad-hoc image pre-processing and feature engineering. These two tasks are labor intensive and limit the flexibility and adaptability of the classical methods. Additionally, variations in the images are common in retinal imaging, such as morphological changes caused by diseases or differences in the image characteristics due to different devices or capture conditions. Therefore, there is an increasing interest in the development of deep learning methods for automatic image registration.

In general, the image registration process starts by defining a fixed image, which is used as reference, and a moving image, which is transformed to match the fixed image accordingly. Classical IBR approaches are based on the optimization of a similarity metric between the intensity values of the fixed and the transformed moving images. There are several similarity metrics that can be used for this purpose, such as mutual information (MI) (15), normalized cross correlation (NCC) (16), mean squared difference (MSD) (17), etc. Similarly, some recent deep learning methods aim at computing a similarity score between images using Convolutional Neural Networks (CNNs) (18) or at speeding the computation of classical metrics using convolutions (19). However, in these cases, the registration procedure still remains an iterative process to search for the optimal transformation. In contrast, other deep learning methods aim to directly predict the transformation matrix

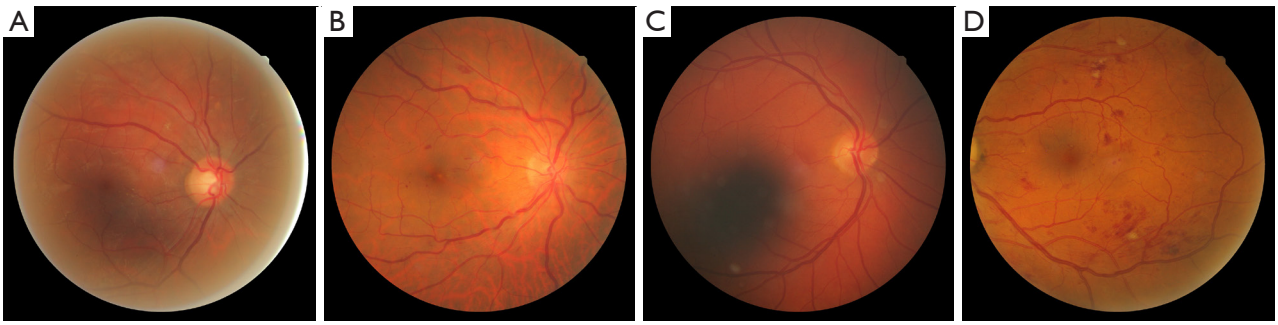

Figure 1 Representative examples of color fundus images. (A) An image with a light halo; (B) an image with a striped pattern; (C) an image with a dark spot; (D) a retina with pathological lesions, in this case hemorrhages among others.

using parameter regression (20,21).

On the other hand, FBR methods are based on keypoints (also known as landmarks). These keypoints are distinctive spatial locations present in both images of a registration pair, hence they can be matched and used to infer the transformation between the images. Therefore, these keypoints are the data on which the registration process is based. In particular, the goal is to find the sufficient amount of keypoint correspondences in the image pair to uniquely characterize a transformation model. These points can be generic or domain specific. Generic keypoints are obtained using broad-domain landmark detectors, like Harris corner detector (22), SIFT (23,24) or SURF (25). On the contrary, domain specific keypoints are often relevant only to a particular application or domain. In this regard, their higher specificity of the domain-specific keypoints allows for faster matching and computation at the cost of flexibility. Commonly, to simplify the finding of correspondences among keypoints, especially when using generic approaches, keypoint descriptors are also used. These descriptors are transformation-invariant feature representations unique for each landmark.

Classical FBR methods are commonly designed with two separate stages for keypoint extraction: keypoint detection and keypoint description. Thus, they are known as detect-then-describe approaches. These methods first find relevant points and then compute their descriptors in order to be able to match them (26). In this context, some novel deep learning methods can also be used to compute descriptors for the keypoints that are detected using classical methods. These hybrid approaches commonly outperform the fully classical methods (27,28). More recently, several deep learning methods capable of jointly detecting and describing keypoints have also been proposed, reaching the best state of the art results in natural image registration such as D2-net (29), R2D2 (30) or SuperPoint (31). These novel approaches combine the typical two-stage approach into a single step capable of detecting and describing the keypoints at the same time.

Contrary to the natural image setting, most classical medical image registration methods are usually IBR instead of FBR (keypoints) (32). Additionally, some recent deep learning medical image registration methods are focused on the novel paradigm of direct parameter regression (21,33,34). These methods are based on a deep neural network capable of directly predicting the transformation parameters or the deformation field that aligns the two input images which form the registration pair. Some of these novel deep learning-based methods have obtained performances superior to classical approaches in several domains, like brain MRI registration (21,35,36). However, these methods have not provided the same performance level in color fundus images, as they cannot improve the results of classical approaches or compete with feature-based deep learning methods.

IBR methods and direct parameter regression approaches are less suitable for color fundus registration than FBR methods (37), despite their success in other medical image modalities, as evidenced by the state of the art results (38,39). This is due to the specific characteristics of fundus images, such as their photographic nature, which differentiates them from other types of medical images. In that regard, fundus images may present large spatial displacements within an image pair, sub-optimal focus in one or both images of the pair, variable illumination, etc. Additionally, the local features that may be useful for image registration, such as those in the arterio-venous tree or the optic disc, occupy a relatively small portion of the image and are scattered over

a homogeneous background. Moreover, fundus images commonly show disease progression or regression, which is more detrimental to intensity or direct regression-based methods as the appearance of the retina can deeply change due to the disease. Furthermore, direct regression methods often overfit their output producing deformation fields that are not realistic and thus are unsuitable. These factors make it difficult to successfully adapt the most commonly used deep learning methods in medical image registration, such as Voxelmorph (21).

The field of retinal image registration is commonly dominated by FBR methods. Particularly, mono-modal color fundus image registration, the focus of this work, is completely dominated by classical FBR approaches. The best results in the public reference color fundus image dataset, Fundus Image Registration Dataset (FIRE) (40), are obtained by methods of this kind, like REMPE (38) and VOTUS (39). However, some recent works also tested deep learning-based FBR methods, showing accurate performance, although not reaching the results of the classical approaches (41).

In multi-modal retinal image registration, FBR deep learning methods have surpassed the performance of classical approaches (42). Retinal image registration methods can be based on either generic (42) or domain-specific keypoints (14,41). In this regard, using domain-specific keypoints, such as blood vessel tree crossovers and bifurcations, involves the use of supervised learning and manually labeled data (14,41). Similarly, the work of Lee *et al.* (43) learns to classify these intersection points in different classes and thus it is able to match them among images, in multimodal retinal image registration. On the other hand, generic keypoint detectors generally do not require labeled data as they are unsupervised (42), an advantage in domains with a lack of labeled images. Moreover, hybrid methods have also been successfully used. For instance, Li *et al.* (44) use classical descriptors [histogram of oriented gradients (HOG)] and then optimize the result with a deformable intensity registration. However, the main drawback of these methods is their execution time and the fact that the deformable methods are completely dependent on the output of the rigid registration and can often produce deformation fields which do not adjust to the real transformation.

Other works have explored the use of deep learning to create direct parameter regression networks (34,45). These directly predict the transformation matrix or deformation field parameters from directly from the input images. However, none of these methods have demonstrated to be competitive against classical approaches, neither in mono-modal color fundus imaging (34) nor in multi-modal registration (45).

Overall, FBR deep learning methods have proven to be competitive with classical approaches, surpassing them in multi-modal retinal imaging (42). However, in color fundus image registration, deep learning methods have yet to obtain this level of performance (41). Therefore, in this work, we propose to address the registration of color fundus images using an FBR deep learning method, which uses a single neural network that jointly detects and describes relevant keypoints in the images. For that purpose, we adapt the R2D2 methodology that was originally developed for natural image registration (30). This methodology allows us to train a neural network, in a single step, for both keypoint detection and description. Additionally, the training is performed without manually annotated ground truth, which is specifically beneficial in this domain due to the scarcity of labeled data. The original R2D2 methodology was developed for its use with natural images, which present very different characteristics to the retinal images in this work (such as the variety of textures, colors, and illumination as well as the size of the images and their contents). Therefore, we perform an exhaustive experimentation and study the effect of different elements in the methodology when it is applied to this new imaging domain. For that purpose, we train and test our method in separate datasets, particularly using the reference color fundus registration dataset FIRE (40) as test set. The cross-dataset validation ensures that our method is robust to changes in imaging devices, patients' characteristics and capture conditions. This is very desirable due to the variety of imaging equipment across ophthalmic services worldwide.

Related work

Currently, in the field of natural image registration the best performing approaches are deep learning FBR methods. While there are some methods that are based on the classical paradigm of detect-then-describe (46), employing two separate steps to detect keypoints and then calculate their descriptors, now the prevailing trend are detect-and-describe approaches. These methods detect and describe the keypoints in a single step and are able to obtain the best results reported in the literature. Some of the most relevant methods in this novel paradigm are D2-net (29), R2D2 (30) and SuperPoint (31). A brief summary of these methods is

provided below.

D2-net (29) introduced late keypoint detection, postponing the keypoint selection step until after the description stage. The output feature maps from the network represent the descriptors for each pixel in the image. The keypoints are selected performing non-local maximum suppression on a feature map followed by a non-maximum suppression across each descriptor. Effectively, this means that the pixels detected as keypoints have distinct descriptors that should provide accurate matching. Thus, both keypoint detection and description are based on high-level information as they are extracted from the deeper layers of a CNN.

R2D2 (30) proposes a similar approach to D2-net (29). However, this method includes dedicated reliability and repeatability maps. These maps are created to make the detected keypoints repeatable (present in the same locations in both images of the pair) and their descriptors reliable (allow to match the exact same keypoints between the two images). This successfully transforms the keypoint selection heuristic from D2-net into a learnable process that is performed by the network.

On the other hand, SuperPoint (31), solves the problem of keypoint detection in a completely different way. They create a pseudo-ground truth dataset of keypoint locations in real images. To do so, they train an initial keypoint detector on synthetic data composed of simple geometric shapes with no ambiguity in the keypoint locations. This initial keypoint detector is then used to create the pseudo-ground truth dataset in which SuperPoint is actually trained. The network used by this approach employs a shared encoder for two separate decoders, one for keypoint detection and other for keypoint description, thus completing the detect-and-describe pipeline.

More recently, these well-known baseline methods were improved by substituting classical geometrical matching methods like RANSAC (RAndom SAmple Consensus) (47) with deep learning methods like SuperGlue (48) among others (49). Other novel approaches consist of using a dense matching among descriptors, eliminating the need for an explicit keypoint detector (50,51).

In contrast to natural images, the field of monomodal color fundus image registration is still dominated by classical FBR approaches, namely REMPE (38) and VOTUS (39). REMPE combines broad-domain keypoints (SIFT) with domain-specific landmarks (blood vessel bifurcations). Then, it employs RANSAC to find an approximated registration, which is later refined with Particle Swarm Optimization and with a more complex transformation model. This two-step keypoint-matching registration pipeline is performed several times to find the optimal solution. On the other hand, VOTUS (39) obtains a graph from the whole blood vessel tree for each image in the pair, matching them to obtain the transformation. This process uses a novel algorithm and several classical image feature descriptors like Gabor filters. VOTUS matches the detected keypoints using DeSAC (Deterministic Sample and Consensus). Both of these methods provide the best overall results in the state of the art. However, they require ad-hoc manual parameter adjusting, which limits their flexibility.

Recently, several works have tested novel deep learning approaches (34,41) in monomodal color fundus image registration. For instance, the method of Rivas-Villar *et al.* (41) proposes a CNN to detect keypoints (blood vessel crossovers and bifurcations). This method is trained using heatmaps derived from human labeling. This way, it is able to produce accurate and repeatable keypoint localization. Due to the specificity of these keypoints, they can use RANSAC directly without the need to compute or match descriptors. This gives this method a significant advantage in execution times, but it cannot improve the results of the classical approaches. On the other hand, the work of Zou *et al.* (34) uses direct parameter regression to register color fundus images. This approach employs a Structure-Driven Regression Network (SDRN), capable of creating deformation fields at different scales. In the same way as (41), it cannot compete with the state of the art. Overall, these works (34,41) demonstrate that deep learning methods cannot yet compete with highly tuned classical methods in color fundus registration.

However, the use of deep learning provides numerous advantages over the classical methods and has shown to be successful in other areas of retinal imaging, such as multi-modal registration. For instance, in the multi-modal registration presented in the work of Wang *et al.* (42) used SuperPoint (31) over segmented vessel maps to register multi-modal retinal images, particularly from fundus images with infrared reflectance and fluorescein angiography. Recently, they improved their previous method (52) to avoid the need of labeled data. However, as it is common on multi-modal registration methods (53) they maintain the need for intermediate representations of the input data, in this case the blood vessel maps. It should be noted that none of these multi-modal registration methods report results on monomodal registration, despite FIRE being the only

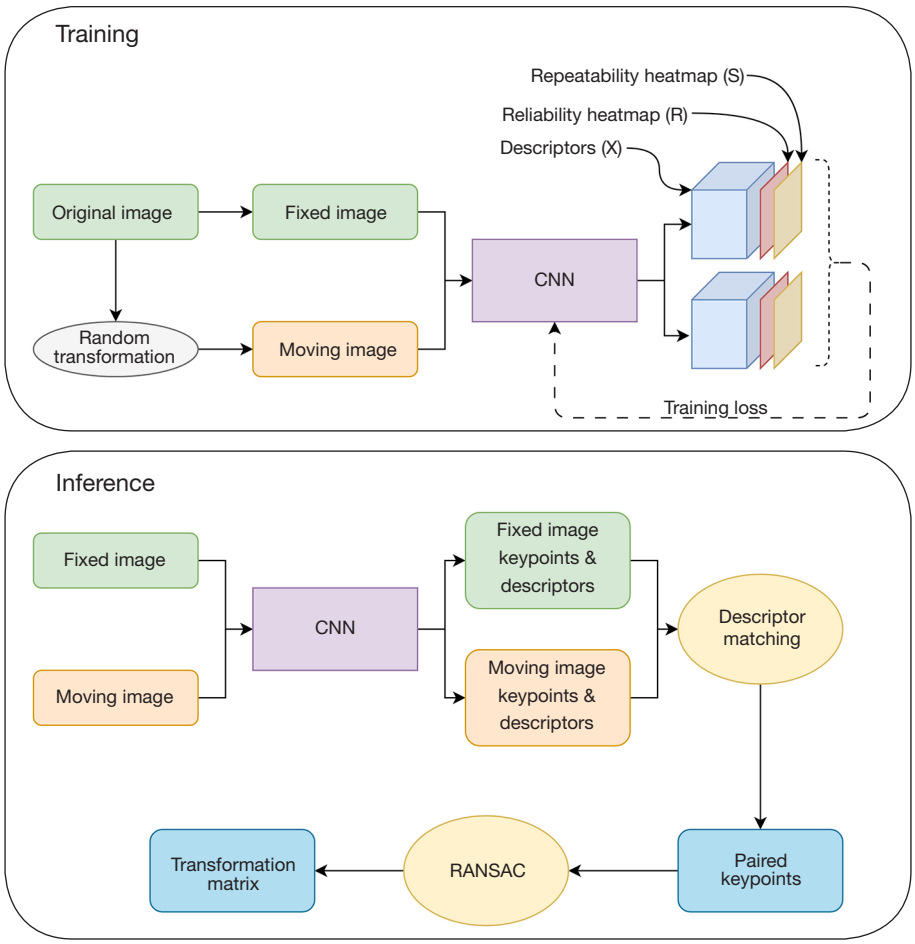

Figure 2 Overview of the methodology. CNN, convolutional neural network; RANSAC, RAndom SAmple Consensus.

dataset with registration ground truth for retinal images. Overall, the main advantage of deep learning approaches is to allow end-to-end learning, directly from the raw input data to the final solution. Moreover, they are more flexible while requiring no feature engineering or complex parameter tuning. This allows deep learning methods to adapt to the different imaging devices and modalities, along with potentially adjusting to challenging conditions, like those derived from the evolution of pathological lesions. The novel deep learning alternative that we propose for the registration of color fundus retinal images exploits these advantages while also making it possible to compete with the highly tuned classical methods. For that purpose, we adapt the R2D2 methodology, which does not require manually annotated data and can jointly detect and describe keypoints using a single deep neural network.

Methods

A complete overview of the proposed registration pipeline can be seen in *Figure 2*. In order to register two different color fundus images we propose using R2D2 (30) as reference, using a single deep neural network capable of jointly detecting and describing keypoints. From these data, we can use a standard registration pipeline with keypoint descriptor matching and RANSAC to infer the transformation. We train a neural network using synthetic image pairs, from a single image we create a registration pair by applying geometric transformations. For each input image, the network generates a dense descriptor block and a repeatability and reliability heatmap. These are used in the loss function so that the network learns to both detect and describe keypoints. Then, on inference, from

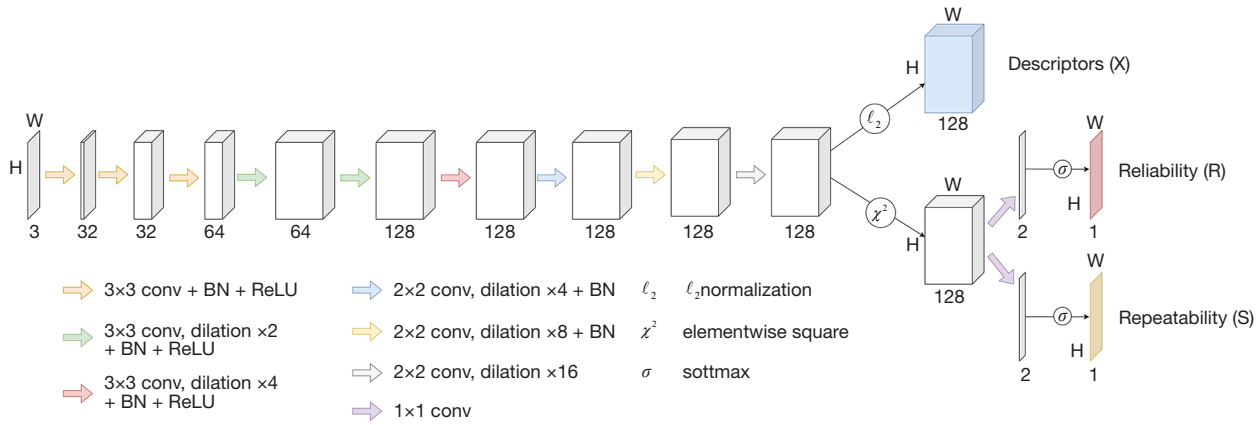

Figure 3 Fully convolutional L2-net used in this work.

two misaligned fundus images, which may be captured in different time-frames or with different viewpoints, we first obtain a set of keypoints and their descriptors using the trained network. Next, the computed keypoint descriptors are matched to find the possible corresponding pairs of keypoints among the two images. Finally, applying the RANSAC algorithm, the obtained set of paired keypoints is used to estimate the transformation matrix between the images.

Network architecture and training

As in R2D2 (30), the joint detection and description of keypoints is performed using a fully convolutional neural network. A representation of the network can be seen in *Figure 3*. This network presents three different outputs. The first output (X) is a dense set of descriptors, consisting of one descriptor for each pixel in the original input image. The second output is a reliability heatmap (R) that intends to estimate how informative and discriminative the descriptor is for each point in the image. In particular, the reliability map is expected to highlight keypoints for which a successful matching is more likely. Finally, the third output is a repeatability heatmap (S) which acts as the base keypoint detector. The keypoints are the local maxima of this heatmap, which is trained to produce maximal points that can be consistently detected on both images that form each pair.

Regarding the network architecture, we use the modified L2-Net (54) proposed in R2D2. The modifications are intended to produce output maps of the same size as the input image (30). The network is trained using unlabeled images and simulated geometric transformations. In particular, for each training image, we generate a pair (I,I') where I denotes the original image and I' the same image with a set of geometrical transformations applied (e.g. shearing). Simultaneously, from this set of transformations, we define the optical flow U, which indicates the correspondence between the points of one image and the other. Then, both images, I and I', are independently processed by the network, producing the corresponding descriptors (X and X') as well as the reliability (R and R') and repeatability (S and S') heatmaps. The losses used to train the network are described in detail below.

Training losses

As in R2D2 (30), we train the network using two separate loss terms, one intended for the repeatability heatmap and the other for the reliability heatmap and the descriptors. The end goal of the repeatability loss is to incentivize the network to generate local maxima (i.e., peaks) in the repeatability heatmaps while ensuring that all the local maxima present in the reliability heatmap S match with the ones in S'. These maxima can then be used as keypoints in the matching process. This is achieved by maximizing the cosine similarity between S and S' in conjunction with a peakiness loss that incentivizes the creation of a single peak per reliability window. In that regard, the peakiness loss for each heatmap is computed as:

$$L_{peaky}(S) = 1 - \frac{1}{|\mathrm{W}|}\sum_{w\in \mathrm{W}}\left(\max S[w] - \mathrm{mean}\, S[w]\right) \quad [1]$$

where W denotes the set of all possible overlapping windows of size N × N, $S[w]$ a particular window w extracted from S, max computes the maximum value of a set, and mean the average. Additionally, it must be noticed

that the window size controls the frequency of detected keypoints per image, as the peakiness loss incentivizes the creation of a single peak per reliability window. Therefore, the lower the window size the more keypoints that are detected and vice-versa.

This same set of windows is also used for the calculation of the cosine similarity loss, which is defined as:

$$L_{cosim}(S,S',U)=1-\frac{1}{|\mathbb{W}|}\sum_{w\in\mathbb{W}}cosim\left(S_f[w],S'_{fU}[w]\right) \quad [2]$$

where $S_f[w]$ represents the flattened version of the $N\times N$ window w extracted from the heatmap S, and $S'_{fU}[w]$ the flattened version of the corresponding window obtained from the transformed image S' using the optical flow U, and *cosim* represents the cosine similarity.

Accordingly, the final equation for the repeatability loss, used to train the repeatability map generation, combines both Lcosim and Lpeaky as:

$$L_{rep}(S,S',U)=L_{cosim}(S,S',U)+\frac{1}{2}\left(L_{peaky}(S)+L_{peaky}(S')\right) \quad [3]$$

In the case of reliability loss, the end goal of the training phase is two-fold. First, the goal is to train the network, so it produces descriptors as distinctive as possible, maximizing the distance among descriptors of different points while minimizing the distance among descriptor of corresponding points. The second target is to train the reliability heatmap R so that it is capable of predicting which points are able to produce descriptors that can be trusted to be accurate and unique for that particular point in the image. This means determining whether each descriptor can unequivocally match its corresponding from the other image.

As in R2D2 (30) the descriptors are trained using the AP loss (30,55), which is based on a differentiable approximation of the Average Precision (AP). In particular, for any pixel (i,j) in the original image I, its descriptor $X_{i,j}$ can be compared with the descriptors of the transformed image I', being these ranked by their distance to $X_{i,j}$. Then, taking into consideration the optical flow U, the AP could be computed by assessing how well the true corresponding descriptor $X'_{U,i,j}$ has been ranked. In this sense, a high AP indicates that the descriptors from the original image match accurately with their corresponding ones in the transformed image I'. The AP loss is computed as:

$$L_{AP}=\sum_{i,j}\left[1-AP\left(p_{i,j}\right)\right] \quad [4]$$

where $p_{i,j}$ denotes a particular patch for which the AP is computed.

This AP loss is modified with additional terms in order to learn the reliability heatmap and avoid optimizing the descriptors in regions with insufficient or non-describing patterns, for which a reliable descriptor cannot be successfully learned. Therefore, the final reliability loss, which is used to learn both the reliability heatmap and the descriptors, is defined as:

$$L_{AP,R}=\sum_{i,j}\left[1-AP\left(p_{i,j}\right)R_{i,j}-k\left(1-R_{i,j}\right)\right] \quad [5]$$

where $k\in[0,1]$ is the AP threshold above which a descriptor is considered reliable. In that regard, using this training loss, when a patch is reliable (i.e., $AP > k$) the loss incentivizes the maximization of the reliability (R). On the contrary, when a patch is unreliable (i.e., $AP < k$) the loss minimizes the reliability. The middle value of $k = 0.5$ was found to be adequate in (30) and thus this value is the one used in our work.

Finally, the full global loss is defined as:

$$L=L_{AP,R}+L_{rep} \quad [6]$$

Inference

Once the network is trained, it can be used to obtain the keypoints and descriptors for any pair of images. Using these keypoints and descriptors we can infer the transformation needed to align both images forming the pair.

Firstly, for each image, a set of candidate keypoints is computed as the local maxima of the repeatability heatmap. As in (30), we follow a multi-scale approach and compute the keypoints at different scales. This means that the trained network is ran multiple times, reducing the size of the input image progressively. Afterwards, we create a list of keypoints over all the scales, using their score in the repeatability heatmap ($S_{i,j}$) as ordering criteria. From this set of keypoints, the ones with low score in the reliability heatmap ($R_{i,j}$) are removed.

After the extraction and selection of the suitable keypoints, the matching pairs between images need to be found in order to estimate the transformation. To do so, we firstly match the keypoints of one image with the other using an algorithm that, for each descriptor, finds its closest counterpart in the other image in terms of the Euclidean distance between descriptors. It should be noted that, for a pair of descriptors to match, both need to have each other

as their closest descriptor. Thus, there cannot be descriptors with more than one match.

Next, the matched keypoint pairs are used as starting data for the estimation of the transformation model using the RANSAC algorithm. RANSAC is able to estimate a mathematical model from a set of observations by separating the data into inliers, which do explain the model, and outliers, which are noise and, thus, do not fit in the model and are discarded. The number of inliers needed to create a transformation model depends entirely on its complexity (56). Consequently, as we use a projective transformation with 8 degrees of freedom (DoF), our model needs at least 4 inlier keypoints (57). The projective transformation has the highest number of DoF of any rigid transformation. Therefore, its use can be beneficial as the extra degrees of freedom can help in certain images with high deformations while not inquiring in the additional complexity for optimization found in non-rigid transformations.

The proposed methodology is trained and tested with color fundus datasets captured with a 45º field of view (FOV), as described in Section "Datasets". Thus, during inference, it is enough to take into account the region of interest (RoI) size to adjust the resolution of the test set images to that of the training dataset. However, our trained networks can be used for inference with color fundus images captured with any FOV, provided that the images are properly scaled to match the resolution of the training images (measured in pixels per square millimeter). This can be easily done by knowing the resolutions of the RoI (i.e., the retinal fundus) of both sets of images. These resolutions can be calculated from the RoI sizes and their angular FOV degrees (58). This allows our method to be applied directly to other datasets without the need for re-training or fine-tuning.

Datasets

The registration performance is evaluated on the public FIRE (40) dataset, which currently is the only available registration dataset for fundus images. For that reason, this dataset is commonly used as benchmark in most of the state of the art works (38,39,41). FIRE contains 134 image pairs of different eyes obtained from 39 separate patients. This dataset has a registration ground truth consisting of control points that were manually annotated in blood vessel crossovers and bifurcations. These control points are scattered evenly on the overlapping region between both images in a pair. This allows to test the quality of the transformation obtained by registration methods.

The size of the images in the FIRE dataset is 2,912×2,912 pixels and it was captured with 45º of FOV. The dataset, which contains 134 image pairs, is divided into 3 separate categories with 71 pairs belonging to category S, 49 to P and 14 to A. Category S has a high degree of overlapping between the images of each pair and no morphological changes due to diseases. Similarly, category A has a high degree of overlapping but it also has morphological changes in its pairs caused by the progression of diseases. Finally, category P has the lowest amount of overlapping and no morphological changes. It should be noted that, as we use data from two public datasets, the study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

In our experiments, the FIRE dataset is used in full as hold-out test set. We use the public Messidor-2 dataset (59,60) to train the network. This dataset is composed of 1,748 images captured with 45º of FOV from multiple Diabetic Retinopathy examinations. The images present varying sizes, particularly 1,440×960, 2,240×1,488 and 2,304×1,536 pixels. In order to train the network with a consistent image size, the images are normalized to a common scale. Thus, the Messidor dataset is normalized using the bigger image size as reference, 2,304×1,536. In this regard, larger image sizes should have less error due to the lower upscaling factor when transforming the points to the full FIRE image size at which the evaluation is performed, as required by the standardized set of metrics used in the state of the art. In order to increase the consistency among FIRE and Messidor, the image sizes and scaling parameters for the different images must be controlled. As both datasets have the same FOV (45º) this can be done directly using the size of the RoI, without having to account for FOV variation. Particularly, the RoI for fundus images correspond to the circular eye fundus which is bordered by a black background (*Figures 4,5*). By resizing the images using the size of the RoI as a reference, the image features of both datasets will be on the same scale despite differences in size of the image border or aspect ratio between datasets. In that regard, it must be noted that we test the network using different input image sizes, all calculated in this manner, in order to check the influence of this factor on the registration performance, as described in depth in Section "Results".

Representative images from the Messidor and FIRE dataset are shown in *Figures 4,5*, respectively.

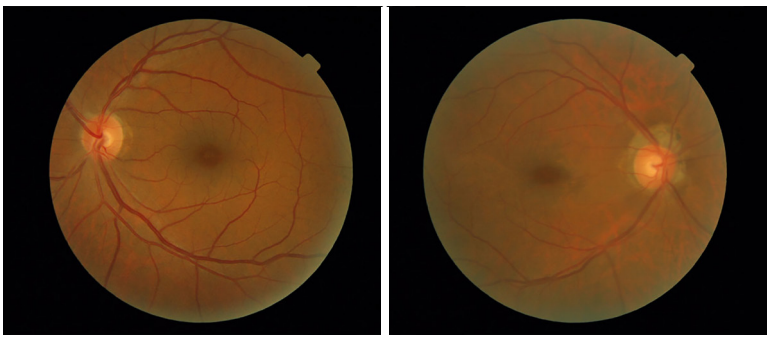

Figure 4 Representative images from the Messidor 2 dataset.

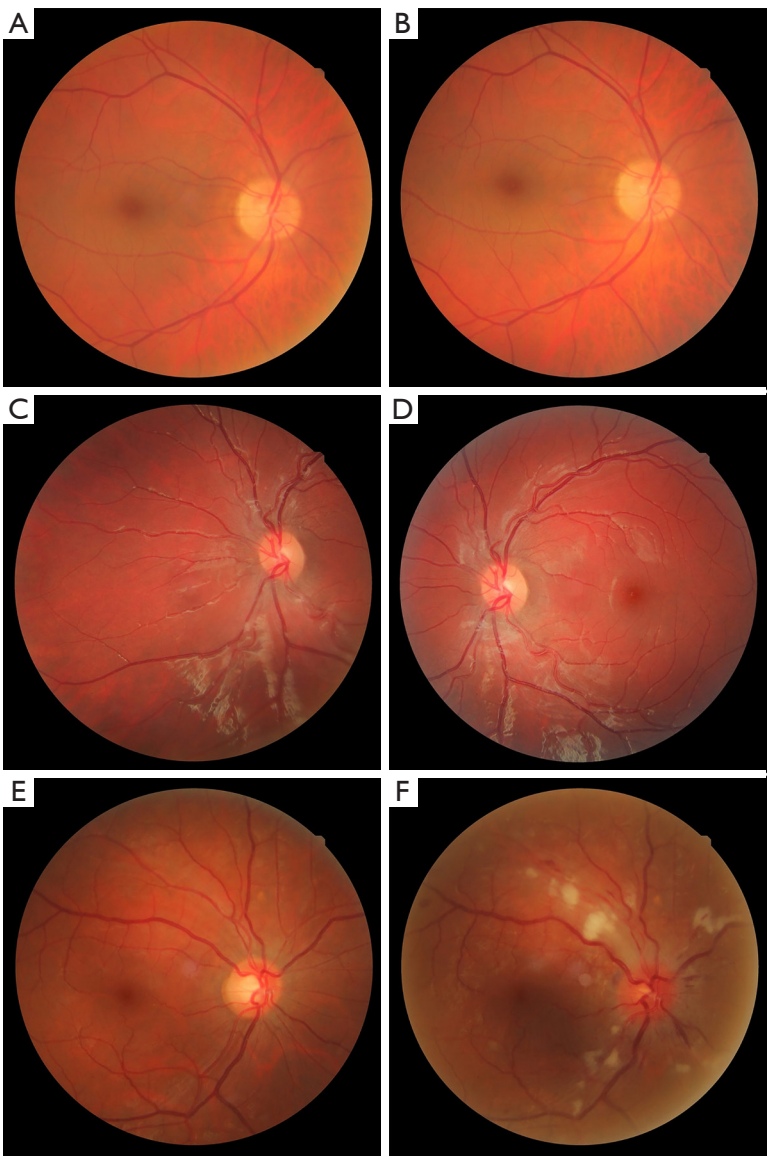

Figure 5 Examples of registration pairs belonging to the FIRE dataset. (A,B) A pair of images from category S. (C,D) A pair of images from category P. (E,F) A pair of images from category A.

Experimental details

The network parameters are initialized with a zero-centered normal distribution using the method proposed by He *et al.* (61). The network was trained from scratch and the optimization algorithm and hyperparameters were set according to the reference state of the art work (30). Particularly, the chosen optimization algorithm is Adam (62) with decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, the default values proposed by its authors. The learning rate is set to 0.0001 and the weight decay to 0.0005. The batch size is set to 8 pairs of images except unless otherwise specified. The network is trained until convergence of the training loss. In particular, the training is stopped after 100 epochs without a significant change in the training loss value.

The training is performed using unlabeled individual images. We create synthetic image pairs applying a series of transformations, following the approach proposed in (30). Particularly, from each individual image, we obtain two copies, one will act as the fixed image and the other as the moving image in the registration process. Image tilting and pixel noise are applied to the moving image as in (30). Both the image tilting and the pixel noise are randomly sampled and follow a uniform distribution. Finally, color augmentation is also performed on the moving image by varying the image components in HSV color space, following the proposal in (63), which is specific for retinal images.

The network training is performed at multiple scales by randomly changing the scale of the input images. In this regard, the same scaling is applied for both the fixed and moving image. For this step we will set the random scaling to be between 1× and 1/3× of the input image size.

Finally, the training is performed using patches of size 192×192 that are randomly cropped from the images (30). The cropping is adapted to the retinal domain, which is characterized by the circular RoI of the images. In that regard, we set the random cropping so that at least 70% of the cropped patch has to be inside of the RoI. From this cropped patch of the fixed image, the corresponding patch in the moving image is found. To account for parts of the moving image getting cut-off due to the spatial transformation, we set the additional condition that the final overlapping region among both patches is at least 50% inside of their RoI. This ensures that the overlapped part among both patches is actually descriptive. If the overlap inside the RoI is near 0, there would not be enough coincident points, which would prevent the network from training. Conversely, setting a restrictive overlap setting (i.e., high amount of overlapping), could also prevent the network from producing a satisfactory result. Therefore, a middle point value, not too nearing either 100% or 0% must be chosen. We arbitrarily set the RoI overlap of the two patches to 50%, as our testing reported that the specific value of this factor did not significantly impact the registration performance.

Regarding the keypoint process that is described in Section "Inference", the network is run several times, reducing the image size. In particular, starting from the maximum image size, *Size*, the image is downsampled by $2^{1/4}$ each time, until its size is smaller than 1/3 *Size*. The maximum number of keypoints extracted per image is set to 5,000. This number is used as it did not impact the performance notably in our testing so we used the one that was proposed by the original authors. Moreover, due to the difference in size between the training and test sets, the images from FIRE must be resized in order to match those of training. Therefore, depending on the resolution used for training, the test images are resized accordingly. It should be noted that, as previously mentioned, this is done using the RoI as reference, as is common practice in color fundus imaging. Furthermore, once the image is processed the keypoints are scaled back (if needed) to the original FIRE resolution in order to evaluate the registration, as detailed in subsection "Evaluation methodology".

The proposed methodology and the performed experiments were implemented in Python 3.8.10 using PyTorch 1.8.2 in combination with CUDA 11.1 and cuDNN 8.0.5. Training, testing and development was performed on a machine with Ubuntu 20.04.3 LTS and equipped with an AMD Ryzen Threadripper 3960X 24-Core Processor with 256 GB of RAM and two NVIDIA RTX A6000 with 48 GB of VRAM each.

Evaluation methodology

In this work we evaluate the performance of our method using the Registration Score that was proposed for the FIRE dataset (40). The first step to obtain the Registration Score is to calculate the registration error, in pixels, for each image pair in the dataset. This error is calculated as the mean of the Euclidean distances between the control points of the fixed and moving image. If the registration error of an image pair is below a certain threshold it is classified as a success, if it is not, it is deemed unsuccessful. Therefore, by altering this threshold and progressively

increasing it, we can obtain a curve. Then, by representing the error threshold in the X axis and the percentage of successes in the Y axis, we can compute an area under curve (AUC) that is used as the Registration Score. Particularly, the proposed AUC is computed between 0 and 25 pixels of error and between 0–100% of success. It should be noted that the error between control points used to calculate the Registration Score is always computed on the original image size of the FIRE dataset, which is 2,912×2,912, enabling direct comparison with state-of-the-art works. The size of the images in the FIRE dataset is significantly bigger than most available datasets (like Messidor 2) and thus some of our experiments require the upscaling of the detected keypoints back to the original FIRE resolution, after being detected by the network in the resolution used for training it. This upscaling can incur in some loss of precision which also means loss in performance in terms of the final score. Using a lower resolution than the original FIRE one means that the keypoint location will be less precise in location, due to the reduced output resolution. This could impact the Registration Score curves by displacing them towards higher error thresholds.

The Registration Score is computed for each of the categories separately as well as for the whole dataset in order to properly assess the performance of our model in each different case. There are several sources of stochasticity that may affect the results. First, despite the convergence in training, at each training iteration the network parameters are still slightly modified to better suit each particular batch of images. Second, the training task that the loss evaluates is not directly comparable to the final goal of registering images using RANSAC. Moreover, RANSAC itself produces non-deterministic behavior, even with high budgets. Additionally, the upscaling process for the detected keypoints can increase the slight differences produced by the factors previously described. These reasons motivate the use of aggregated metrics over the last epochs in order to produce a representative evaluation. In order to take into account the stochasticity of the training process as well as the inherent randomness produced by RANSAC, the results are presented as means and standard deviations computed from the last 50 training epochs, getting data each five epochs, for a total of then data samples. These metrics allows us to report reliable results, taking into account the stochasticity of the method. Furthermore, as we report means and standard deviations, we facilitate more complete comparisons with our method for future methods.

Results

We study the effects of several parameters in the registration performance of the proposed methodology. In particular we analyze: (I) the variation in the overall input image size, (II) the size of the patches used as input to train the network and calculate the reliability loss, and (III) the size of the windows used to calculate the repeatability loss.

Analysis of the image size

We study the effects of the image size in the proposed methodology by testing three separate target sizes: small, medium and large. In particular, we choose image sizes that are similar to those commonly used in the state of the art, hence facilitating the comparison among methods. First, the smallest input image size is chosen to be similar to the one of a previous work (41), but adapted to the aspect ratio of the training dataset (Messidor dataset) as well as the chosen methodology. This results in an image size 1,152×768 px. To test a medium size, we choose the Messidor image size. As Messidor is composed of images with different sizes, we normalize the dataset to a consistent size, choosing the biggest one available, 2,304×1,536 px. Finally, in order to test a large input size, we select the image size of the FIRE dataset (40), which adapted to Messidor's aspect ratio is 4,368×2,912 px. This large image size allows to directly detect and describe the keypoints over the original images of the FIRE dataset, hence avoiding the re-scaling of the detected keypoints to compute the Registration Score. This could provide precision advantages as there is inherent error in the scaling process. However, the large image size requires re-scaling all the Messidor training images up to double its dimensions. Therefore, as a counterpart, this could lead to artifacts that may affect the training of the network. The conducted experiments allow us to study this trade-off.

In these experiments, the window size for the computation of the repeatability loss, which controls the frequency of detected keypoints, is kept at a ratio proportional to the input size. Therefore, as we double the input image size we also double the window size, resulting in a window of 16 pixels for the small size input images, 32 for the medium ones and 64 for the large size. This way, keeping the size ratio for the window and the input image, creates a constant amount of keypoints per image for all the experiments, regardless of the image size.

On the other hand, for this experiment we keep the

Table 1 Results for the different training image sizes, measured in Registration Score in the FIRE image size

Image size	Category A	Category S	Category P	FIRE
Small	0.726±0.022	0.925±0.002	0.352±0.057	0.695±0.023
Medium	0.520±0.079	0.780±0.069	0.065±0.023	0.492±0.048
Large	0.267±0.046	0.374±0.153	0.009±0.013	0.229±0.090

Data are presented as mean±standard deviation of the registration score. FIRE, Fundus Image Registration Dataset.

input patch at a constant size of 192×192 pixels for every image size. Therefore, the bigger the input image is, the less it will fit in this default patch. In particular, this input patch size represents 25% of the height of the image in the small image size, 12.5% in medium and 6.6% in the large image size. Moreover, the patch size also limits the input image size, as the images are scaled randomly between their original size and 1/3 of it. Therefore, there exists a lower bound for image size, which is the input patch size. Conversely, the input patch size is limited by the input image size. This means that 1/3 of the input image size must be bigger than the input patch so that this patch does not encompass an area bigger than the image itself. This prevents us from halving the small image size again to test another lower input image size.

The results for these experiments, testing the different input image sizes, are shown in *Table 1*. Overall, we can see that the small image size produces the best results in every category of the FIRE dataset, as well as in the full dataset. Furthermore, we can see that increasing the image size notably degrades the performance with each size increase.

Analysis of the input patch size

We also study the effect that the size of the input patch, during training, has on the registration performance. By default, the input patch size is set to 192×192 pixels, as proposed by the authors in (30). This size affects the amount of information seen by the network as well as the amount of points included in the calculation of the AP component of the reliability loss. Thus, increasing the size may potentially provide an improved training feedback, leading to better convergence of the descriptors and better registration performance. However, in the previous experiment this patch size stayed the same, therefore reducing the amount of the image that the network sees. Therefore, in this experiment, we propose to use proportional patch sizes, increasing them in combination with the images. Thus, taking as baseline the patch of 192 pixels in the small size, we propose to use patches equivalent in size to this in both medium and large image sizes. The baseline patch of 192 pixels in the small image size is equivalent to 25% of the height of the image or 4.2% of the area of the image. Then, for the medium, we double the patch size following the doubling up in image size. Therefore, the patch of 384 is also equivalent to 25% of the height of the image or 4.2% of the area of the image in medium size. Finally, we double the patch size once again for the large image size. However, as this image size is made to fit the FIRE dataset, the proposed patch of 768 pixels represents 26.4% of the image size or 4.6% of the image area. Regarding the repeatability window size, we use the same ones as in the previous experiments, proportional. Thus, in this experiment all the input sizes are equivalent across image sizes.

However, there are some limitations regarding the maximum patch size for each of the input image sizes. The maximum possible size for the input patch is limited by the random downscaling applied as data augmentation. These augmentations are key to the performance of the network as, in inference, the network is run multiple times over the input image, downsampling it. In particular, each input image can be randomly scaled between 1× and 1/3× its size. Therefore, the input patch cannot be bigger than 1/3 or 33% of the height of the input image, as then it could be bigger than the whole image.

Additionally, in the case of the large patches the GPU memory could also be a limiting factor. In that regard, for the experiments with larger patches we had to lower the batch size from the default 8. However, in order to make the different experiments comparable to one another, we accumulate gradients in order to simulate the original batch of 8 images. This allows us to effectively train the networks with a batch of 8 in spite of the limited memory.

The results for these experiments are reported on *Table 2*. These results show that increasing the patch size provides improved results. However, despite the improvement in performance provided by the use of a proportional input patch size in the medium and large image sizes, the small

Table 2 Registration performance on the FIRE dataset for different configurations of input image size and input patch size

Image size	Patch	% of height	Category A	Category S	Category P	FIRE
Small	192 px	25%	0.726±0.022	0.925±0.002	0.352±0.057	0.695±0.023
Medium	192 px	12.5%	0.520±0.079	0.780±0.069	0.065±0.023	0.492±0.048
	384 px	25%	0.582±0.037	0.907±0.010	0.160±0.020	0.599±0.011
Large	192 px	6.6%	0.267±0.046	0.374±0.153	0.009±0.013	0.229±0.090
	768 px	26.4%	0.382±0.045	0.82±0.035	0.067±0.016	0.496±0.023

The performance is measured in terms of registration score. Data are presented as mean ± standard deviation of the Registration Score. FIRE, Fundus Image Registration Dataset.

Table 3 Registration performance on the FIRE dataset for different window sizes in the repeatability loss

Window size	Category A	Category S	Category P	FIRE
8 px	–	–	–	–
16 px	0.726±0.022	0.925±0.002	0.352±0.057	0.695±0.023
32 px	0.706±0.018	0.921±0.002	0.289±0.049	0.668±0.020
64 px	0.661±0.075	0.894±0.052	0.282±0.052	0.644±0.068

The performance is measured in terms of Registration Score. Data are presented as mean ± standard deviation of the Registration Score. "–" indicates network incapable of successful registration. FIRE, Fundus Image Registration Dataset.

image size still obtains the best results.

Analysis of the repeatability window size

Along with the changes in image size and input patch size, we also study how the size of the windows used in the computation of the repeatability loss impact the performance of the method. In that regard, it must be noticed that the window size directly affects the keypoint localization because it controls the frequency of the local maxima (peaks) in the repeatability heatmap. Lowering the window size results in more frequent keypoints, especially increasing the population of keypoints in certain areas with a high amount of information. Meanwhile, increasing the window size results in less frequent but more repeatable and distinctive keypoints. In that regard, the window size also limits the maximum amount of keypoints that can be selected from each image. As indicated in Section "Experimental details", up to 5,000 keypoints are selected per image (30). However, this value is not necessarily met in all the experiments, as enlarging the window may not always allow to create that many keypoints in the first place. These observations motivate the experimentation on the effect of the window size. For these experiments, the input image size and the input patch size are fixed to the ones that produce the best results in previous experiments. In these experiments, the window size for the computation of the repeatability loss, which controls the frequency of detected keypoints, is kept at a ratio proportional to the input size. Therefore, as we double the input image size, we also double the window size, resulting in a window of 16 pixels for the small size input images, 32 for the medium ones and 64 for the large size. This way, keeping the size ratio for the window and the input image, creates a constant amount of keypoints per image for all the experiments, regardless of the image size.

The results for these experiments, testing different repeatability window sizes, are shown in *Table 3*. The best results are obtained using the smaller window that allows the network to converge, in this case 16 pixels. Using a smaller window (8 pixels) impedes the network's training. Moreover, increasing the window size from 16 pixels causes the results to worsen in every category of FIRE.

State of the art comparison

In this section, we provide a comparison among the proposed methodology and the best methods in the state of the art. Every method is evaluated in the FIRE dataset using Registration Score as the metric. For the proposed methodology, we use the combination of image,

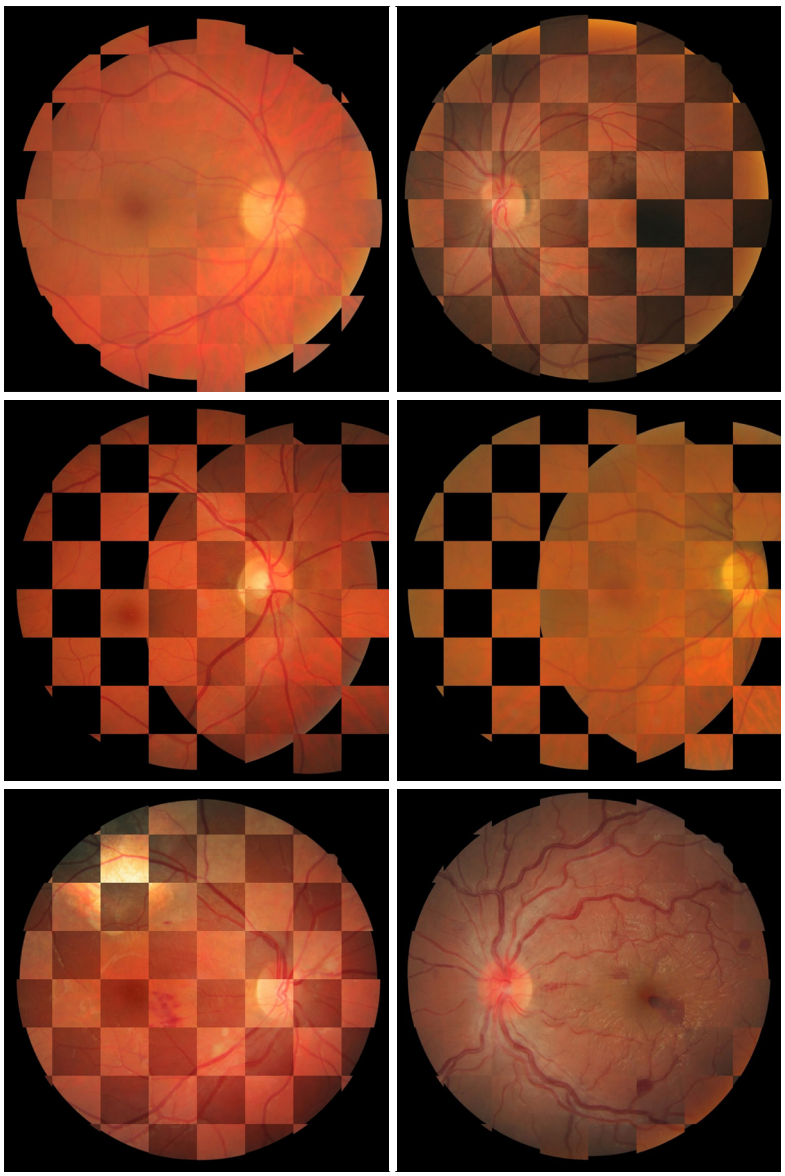

Figure 6 Representative registration examples from the FIRE dataset, the top row corresponds to category S, the middle row to category P and the bottom row to category A. FIRE, Fundus Image Registration Dataset.

patch and window sizes that produce the best results. This combination corresponds to the small image size, a patch size of 192 pixels and a window size of 16 pixels. Representative registered images for each class in the FIRE dataset obtained with the best presented approach are presented in *Figure 6*. The comparison with the state of the art is shown in *Table 4*. Moreover, a qualitative comparison of registration examples is shown in *Figure 7*, showing the results of our method and other state-of-the-art methods (38,41). Overall, our proposal obtains satisfactory performance. It produces the best results in category A and shows similar performance to the best methods in category S. However, in category P, the performance is not comparable to the best methods.

Discussion

Analysis of the image size

The results shown in *Table 1* reveal that the small image

Table 4 Comparison among the proposed methodology and different SOTA methods

Name	Registration score (AUC)				Transformation model
	S	P	A	FIRE	
VOTUS (39)	0.934	0.672	0.681	0.812	Quadratic
REMPE (38)	0.958	0.542	0.660	0.773	Ellipsoid eye model
GFEMR (64)	0.812	0.607	0.474	0.702	Elastic
Proposed work	0.925	0.352	0.726	0.695	Projective
SIFT+WGTM (26)	0.837	0.544	0.407	0.685	Quadratic
Deep CB (41)	0.908	0.293	0.660	0.657	Similarity
GDB-ICP (23)	0.814	0.303	0.303	0.576	Quadratic
Harris-PIIFD (22)	0.900	0.090	0.443	0.553	Polynomial
ED-DB-ICP (24)	0.604	0.441	0.497	0.553	Affine
SURF+WGTM (65)	0.835	0.061	0.069	0.472	Quadratic
RIR-BS (66)	0.772	0.0049	0.124	0.440	Projective
EyeSLAM (67)	0.308	0.224	0.269	0.273	Rigid
ATS-RGM (68)	0.369	0.000	0.147	0.211	Elastic

SOTA results extracted from (38) and (41). The performance is measured in terms of Registration Score and the methods are ranked by their overall performance in the FIRE dataset. SOTA, start-of-the-art; AUC, area under curve; FIRE, Fundus Image Registration Dataset.

size produces the best results and that increasing the image size degrades the performance. In that regard, it is especially relevant that the large image size, which does not require any re-scaling of the detected keypoints, obtains the lowest Registration Scores by a notable margin. However, it must be noticed that other factors may also affect the performance. For instance, keeping the same input patch size, the image size increase means a reduction in the proportion of the image that is processed by the network. Therefore, the network has less context to determine if the keypoints are relevant and to accurately describe them. This, in turn, may cause the network to focus on low level features producing less distinctive keypoints and less informative descriptors. These results motivate the study of the input patch size, which in addition also has an effect in the amount and variety of points used in the calculation of the reliability loss. Increasing the size of this patch gives the network more context, which may increase the performance.

Analysis of the input patch size

Overall, the results in *Table 2* show that increasing the patch size provides improved results. Increasing the patch size allows the AP loss to process a higher number of points, adding more meaningful points that are similar to the true correspondence but are in other locations of the image. The addition of these points helps improve the learning of the descriptors as they provide challenging negative examples. This explains the improved performance in the medium and large sizes when increasing the patch size. In this regard, it should be noted that the random augmentations applied during training limit the maximum input patch size, so that it cannot be meaningfully increased from the 25% used as baseline.

Despite the improvement in results in the larger image size provided by the use of a proportional input patch size, the small image size still produces the best results. This is due to the potential benefits discussed in the previous section. In that regard, although the bigger patch size has the potential to improve the context that is seen by the network, the amount of context that is used to detect the keypoints and create the descriptors is also limited by the receptive field of the network architecture. Using the current network, the lower image sizes are beneficial as more of the image can fit in the receptive field and thus the extracted features are of higher level. In that regard, it seems that current network architectures are limited with

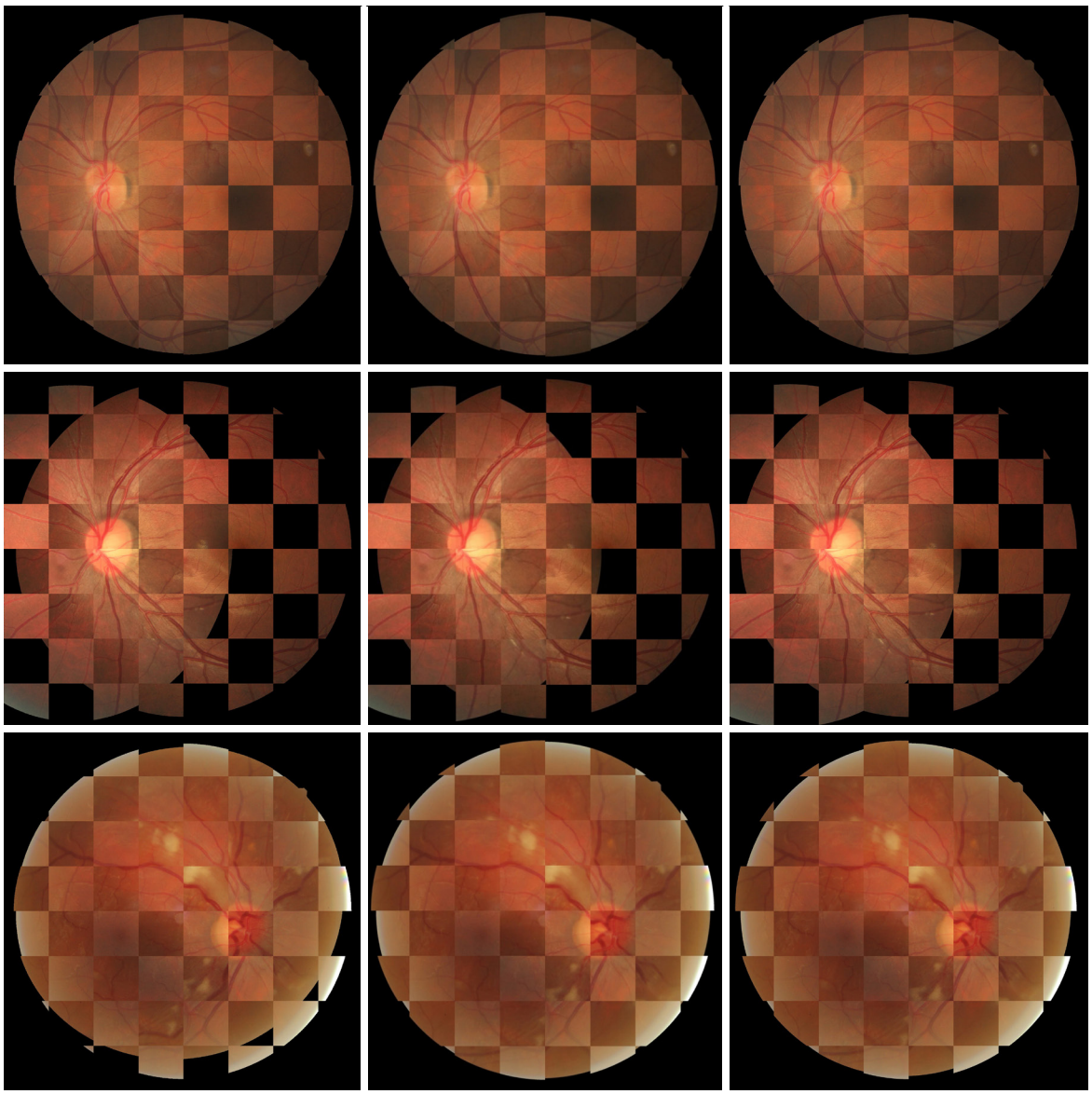

Figure 7 Registration examples from the FIRE dataset for different state-of-the-art methods. Left column is REMPE (38), the middle column is our method, and the left column results are from (41). The top row images correspond to category S, the middle row to category P and the bottom row to category A. FIRE, Fundus Image Registration Dataset.

regards to current image capture devices, which are able to produce images with 4× the size of what current state of the art retinal image and color fundus registration methods are usually able to use. For instance, the work of Wang *et al.* (42,52) operates in the same image size as our best approach (768 px) while the work of Zou *et al.* (34) operates in an even lower input image size (256 px). These image sizes are much smaller than FIRE's original size (2,912×2,912 px) and, therefore, there is an inherent loss of details as the network cannot take advantage of the full image resolution.

Analysis of the repeatability window size

In the results for these experiments, show *Table 3*, the best performance is achieved using the smallest window size that allows the network to converge to a successful solution (16 pixels). In practice, this is the case that produces the most frequent keypoints. In the case of the smallest tested

window (8 pixels), the network is forced to detect keypoints very close together, which impedes its successful training as the images do not have enough information to support such high frequency of repeatable keypoints.

Beyond 16 pixels, as we double the window size, the performance degrades in every category of the test dataset. This is especially notable in the Category P, where increasing the window size from the original 16 pixels to 32, causes a 6% drop in Registration Score. Category P is particularly sensitive to a reduction in the amount and frequency of keypoints because the images in this group have a low degree of overlapping. This scenario requires a higher frequency of detected keypoints in order to have a sufficient number of them to successfully register the images.

In the case of category A, there is also a significant decrease in performance when increasing the window size, though in this case the larger reduction happens between 32 and 64 pixels. Again, this can be explained due to the frequency of keypoints, as in category A the images are affected by pathological progression. Therefore, enough keypoints must be present over all the eye fundus such that, if there are new lesions in one of the images, nearby zones can still provide enough keypoints to successfully register the image pair.

Finally, in category S, due to the lack of pathologies as well as the high overlapping, the decrease in performance when increasing the window size is much less noticeable, decreasing around only 3% going from the best to the worst result.

Overall, this window size presents a trade-off between frequent albeit more un-repeatable keypoints and less frequent, more scattered, and highly repeatable keypoints. Generally, more frequent keypoints improve the registration results. Although there is a limit on how close together the keypoints can be detected by the network. This is evidenced during training as, if the network is forced to detect very frequent keypoints, it cannot learn to do it and thus never converges to a successful solution. As the window size increases, the chances that the network produces the exact same keypoints in each image of the pair also increases. However, due to the particularities of retinal images, the higher keypoint frequency proves to be more important than the increased repeatability provided by the bigger window sizes. As mentioned, this is due to the presence of pathological lesions and the low overlapping among images in some pairs. Furthermore, while the lower window sizes can produce keypoints in non-informative regions, this issue is nullified by the selection of the most salient keypoints (based on their score in the repeatability map) as well as the filtering of keypoints whose descriptor is not discriminative enough (based on their score in reliability heatmap).

State of the art comparison

As per the results shown in *Table 4*, our method obtains the best results in the FIRE category A and shows similar performance to the best methods in category S. However, in category P, the performance is not comparable to the best methods. Regarding these results, it must be noticed that category A is arguably the most relevant for the clinical practice. This category shows disease progression between the two images that form each pair. Therefore, in order to obtain reliable results in this category, the approaches must be robust against pathological changes in the retina. This characteristic is particularly relevant as image registration is commonly used either in CAD systems or directly by the clinicians in different pathological scenarios, including longitudinal studies. Considering this context, it is remarkable that our method outperforms all the previous alternatives in category A, which is the only category with disease progression or remission. In that regard, the specific characteristics of this category usually make the registration particularly challenging.

Regarding categories S and P, our method outperforms the previous best Deep Learning method (41). Additionally, in category S our proposal shows results rivaling the best overall methods. However, due to the lower complexity of the image pairs, the differences among methods are inherently very small. In the case of P, even if our method outperforms the deep learning state of the art, the results are still far from those obtained by classical methods like VOTUS (39) or REMPE (38). This is caused by the low overlapping between the images in category P, which highlights the oval or spherical shape of the retina. To successfully estimate the spatial transformation in this case, it would be necessary a custom geometrical model adapted to the specific shape of the retina [which can change from person to person (69) and also within the same person due to disease progression (70)] as well as the FOV of the capturing device. Using a projective transformation model has the advantage of requiring few keypoints to successfully register the images, which can be useful when registering images with severe pathological progression. However, its main drawback is the lack of degrees of freedom to correctly align very disjointed images with low amount of expected overlapping.

Concerning the results for the overall dataset, our method improves the deep learning state of the art. However, when comparing it to the overall best methods, our approach is hampered by its performance in the P category. In that regard, it must be noticed that category P represents around 37% of the image pairs in the test set. On the contrary, category A, the one with disease progression and where our method obtains the best overall results, only represents around 10% of the test set. Therefore, the overall comparison is skewed due to the low number of pathological samples in the dataset.

Besides the improvements in category A, another relevant novelty of our proposal is that it is a single stage method, capable of simultaneously detecting and describing keypoints. In this regard, an ablation study of the two parts of the network output could be considered. However, this would only be possible by replacing either one of them with an additional algorithm and effectively transformed the method into a two-stage pipeline. Additionally, given that the network is jointly trained for both tasks, the keypoints and descriptors are optimized for each other. Therefore, changing either the detector or descriptor would lead to suboptimal performance. Similarly, removing training loss terms prevents the network from functioning as it prevents the network from doing both tasks (i.e., keypoint detection and description). Moreover, testing the original approach without our modifications is not adequate as it is affected by the particularities of the color fundus images, like the region outside of the RoI. Lastly, another experiment that could be considered would be to use the original R2D2 without modifications. However, this would be inadequate due to the particularities of the color fundus images, like the black background outside the RoI.

Conclusions

Retinal image registration is very important for clinical practice as well as CAD systems in order to help in longitudinal studies and disease monitoring. In that regard, there is a great interest in the development of novel registration methods, especially if they allow to successfully work in scenarios of disease progression. Currently, the field of color fundus registration is dominated by ad-hoc classical methods. Deep learning methods are desirable due to their increased adaptability and flexibility. However, previous deep learning methods do not achieve competitive results in the state of the art.

In this work, we propose a deep learning-based FBR methodology for the registration of retinal images. Our method is based on a proven state-of-the-art approach created for natural images. In particular, we use a deep neural network to jointly detect and describe representative keypoints in the retinal images. First, our approach uses a repeatability map to ensure that the detected keypoints are repeatable in the images being registered. Second, a reliability heatmap is also used in order to detect which keypoints in the image provide sufficiently discriminative descriptors. Then, the detected and matched keypoints are used in the RANSAC algorithm to estimate a projective transformation model. The training of the network is unsupervised and does not require a manually annotated ground truth.

In order to validate the proposed methodology, we conducted multiple experiments and studied how various factors of the methodology affect the performance. The training of the network was performed using images from the public Messidor-2 dataset. Meanwhile, the evaluation is conducted on the public FIRE dataset. This cross-dataset validation ensures that our method is robust to changes in imaging device and capture conditions.

The obtained results show a satisfactory performance. Particularly, our approach improves the results obtained by previous deep learning-based methods in all of the categories of the FIRE dataset. Furthermore, our method gets the best results in category A of the FIRE dataset, which is the one containing images with disease progression. This is the category with the most relevance for clinical practice, as registering these kinds of images would facilitate longitudinal studies and disease progression monitoring. However, the results in the category with low degree of overlapping, although better than the rest of deep learning approaches, were not competitive with the best ad-hoc classical methods. This is an area that should be considered for improvement in future works. In that regard, as future research directions, we consider including the learning of domain specific keypoints, like blood vessel crossovers and bifurcations. This could prove beneficial as it has been previously demonstrated in successful classical approaches. Similarly, our experiments also showed evidence that state-of-the-art networks are limited by their design when processing current-day high resolution images. Our proposal, like those in the state of the art, works best on medium or small images and therefore cannot take advantage of all the detail provided by current capture devices that can produce images with very high resolutions. Therefore, another possibility to explore in future work

is to design novel network architectures that allows to efficiently and accurately take advantage of high resolution images.

Acknowledgments

Funding: This work was supported by Ministerio de Ciencia e Innovación, Government of Spain, through the RTI2018-095894-B-I00, PID2019-108435RB-I00, TED2021-131201B-I00, and PDC2022-133132-I00 research projects; Consellería de Cultura, Educación e Universidade Xunta de Galicia through the Grupos de Referencia Competitiva grant (Ref. ED431C 2020/24), the predoctoral fellowship (Ref. ED481A 2021/147) and the postdoctoral fellowship (Ref. ED481B-2022-025); CITIC, Centro de Investigación de Galicia (Ref. ED431G 2019/01), itself received financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-4/coif). All authors report that their institution received funding from Ministerio de Ciencia e Innovación, Government of Spain; Consellería de Cultura, Educación e Universidade, Xunta de Galicia and the European Regional Development Fund. DRV receives funding from Consellería de Cultura, Educación e Universidade, Xunta de Galicia through a predoctoral fellowship. ÁSH receives funding from Consellería de Cultura, Educación e Universidade, Xunta de Galicia through a postdoctoral fellowship. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Viergever MA, Maintz JBA, Klein S, Murphy K, Staring M, Pluim JPW. A survey of medical image registration - under review. Med Image Anal 2016;33:140-4.
2. Hajnal J, Hill D, Hawkes DJ. Medical image registration. Biomedical engineering series. Boca Raton, FL: CRC Press, 2001.
3. Narasimha-Iyer H, Can A, Roysam B, Tanenbaum HL, Majerovics A. Integrated analysis of vascular and nonvascular changes from color retinal fundus image sequences. IEEE Trans Biomed Eng 2007;54:1436-45.
4. Morita K. Computer-aided Diagnosis Systems Based on Medical Image Registration [Ph.D. Thesis]. Japan: University of Hyogo Kobe, 2019.
5. Yanase J, Triantaphyllou E. The seven key challenges for the future of computer-aided diagnosis in medicine. Int J Med Inform 2019;129:413-22.
6. Huang K, Li M, Yu J, Miao J, Hu Z, Yuan S, Chen Q. Lesion-aware generative adversarial networks for color fundus image to fundus fluorescein angiography translation. Comput Methods Programs Biomed 2023;229:107306.
7. Li P, He Y, Wang P, Wang J, Shi G, Chen Y. Synthesizing multi-frame high-resolution fluorescein angiography images from retinal fundus images using generative adversarial networks. Biomed Eng Online 2023;22:16.
8. Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017:1125-34.
9. Hervella ÁS, Rouco J, Novo J, Ortega M. Retinal microaneurysms detection using adversarial pre-training with unlabeled multimodal images. Inf Fusion 2022;79:146-61.
10. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, 2017:2223-32.
11. Hervella AS, Rouco J, Novo J, Ortega M. Deep Multimodal Reconstruction of Retinal Images Using Paired or Unpaired Data. In: 2019 International Joint Conference on Neural Networks (IJCNN), 2019.

12. Forrester JV, Dick AD, McMenamin PG, Roberts F, Pearlman E. The eye e-book: basic sciences in practice. Elsevier Health Sciences, 2020.
13. Salmon JF. Kanski's Clinical Ophthalmology: A Systematic Approach. Elsevier, 2020.
14. Hervella ÁS, Rouco J, Novo J, Ortega M. Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. Procedia Comput Sci 2018;126:97-104.
15. Pluim JPW, Maintz JBA, Viergever MA. Image Registration by Maximization of Combined Mutual Information and Gradient Information. In: Delp SL, DiGoia AM, Jaramaz B, editors. Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2000. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000:452-61.
16. Balakrishnan G, Zhao A, Sabuncu MR, Dalca AV, Guttag J. An Unsupervised Learning Model for Deformable Medical Image Registration. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018:9252-60.
17. Kybic J, Unser M. Fast parametric elastic image registration. IEEE Trans Image Process 2003;12:1427-42.
18. Cheng X, Zhang L, Zheng Y. Deep similarity learning for multimodal medical images. Computer Comput Methods Biomech Biomed Eng Imaging & Vis 2018;6:248-52.
19. Menchón-Lara R-M, Simmross-Wattenberg F, Rodríguez-Cayetano M, Casaseca-de-la-Higuera P, Á. Martín-Fernández M, Alberola-López C. Efficient convolution-based pairwise elastic image registration on three multimodal similarity metrics. Signal Process 2023;202:108771.
20. Haskins G, Kruger U, Yan: Deep learning in medical image registration: a survey. Mach Vis Appl 2020;31:8.
21. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Trans Med Imaging 2019. [Epub ahead of print]. doi: 10.1109/TMI.2019.2897538.
22. Chen J, Tian J, Lee N, Zheng J, Smith RT, Laine AF. A partial intensity invariant feature descriptor for multimodal retinal image registration. IEEE Trans Biomed Eng 2010;57:1707-18.
23. Yang G, Stewart CV, Sofka M, Tsai CL. Registration of challenging image pairs: initialization, estimation, and decision. IEEE Trans Pattern Anal Mach Intell 2007;29:1973-89.
24. Tsai CL, Li CY, Yang G, Lin KS. The edge-driven dual-bootstrap iterative closest point algorithm for registration of multimodal fluorescein angiogram sequence. IEEE Trans Med Imaging 2010;29:636-49.
25. Wang G, Wang Z, Chen Y, Zhao W. Robust point matching method for multimodal retinal image registration. Biomed Signal Process Control 2015;19:68-76.
26. Lowe DG. Distinctive Image Features from Scale-Invariant Keypoints. Int J Comput Vis 2004;60:91-110.
27. Noh H, Araujo A, Sim J, Weyand T, Han B. Large-Scale Image Retrieval With Attentive Deep Local Features. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
28. Luo Z, Shen T, Zhou L, Zhang J, Yao Y, Li S, Fang T, Quan L. ContextDesc: Local Descriptor Augmentation With Cross-Modality Context. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
29. Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A, Sattler T. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
30. Revaud J, De Souza C, Humenberger M, Weinzaepfel: R2D2: Reliable and Repeatable Detector and Descriptor. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems, 2019.
31. DeTone D, Malisiewicz T, Rabinovich A. SuperPoint: Self-Supervised Interest Point Detection and Description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.
32. Chen X, Diaz-Pinto A, Ravikumar N, Frangi A. Deep learning in medical image registration. Prog Biomed Eng 2021. Available online: https://iopscience.iop.org/article/10.1088/2516-1091/abd37c
33. Hu R, Yan H, Nian F, Mao R, Li T. Unsupervised computed tomography and cone-beam computed tomography image registration using a dual attention network. Quant Imaging Med Surg 2022;12:3705-16.
34. Zou B, He Z, Zhao R, Zhu C, Liao W, Li S. Non-rigid retinal image registration using an unsupervised structure-driven regression network. Neurocomputing 2020;404:14-25.
35. de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Išgum I. A deep learning framework for unsupervised affine and deformable image registration. Med Image Anal 2019;52:128-43.
36. Xiao H, Teng X, Liu C, Li T, Ren G, Yang R, Shen D, Cai J. A review of deep learning-based three-dimensional

medical image registration methods. Quant Imaging Med Surg 2021;11:4895-916.
37. Saha SK, Xiao D, Bhuiyan A, Wong TY, Kanagasingam Y. Color fundus image registration techniques and applications for automated analysis of diabetic retinopathy progression: A review. Biomed Signal Process Control 2019;47:288-302.
38. Hernandez-Matas C, Zabulis X, Argyros AA. REMPE: Registration of Retinal Images Through Eye Modelling and Pose Estimation. IEEE J Biomed Health Inform 2020;24:3362-73.
39. Motta D, Casaca W, Paiva A. Vessel Optimal Transport for Automated Alignment of Retinal Fundus Images. IEEE Trans Image Process 2019;28:6154-68.
40. Hernandez-Matas C, Zabulis X, Triantafyllou A, Anyfanti P, Douma S, Argyros A. FIRE: Fundus Image Registration Dataset. Available online: https://carlos.hernandez.im/papers/2017_07_JMO.pdf
41. Rivas-Villar D, Hervella ÁS, Rouco J, Novo J. Color fundus image registration using a learning-based domain-specific landmark detection methodology. Comput Biol Med 2021. [Epub ahead of print]. doi: 10.1016/j.compbiomed.2021.105101.
42. Wang Y, Zhang J, Cavichini M, Bartsch DG, Freeman WR, Nguyen TQ, An C. Robust Content-Adaptive Global Registration for Multimodal Retinal Images Using Weakly Supervised Deep-Learning Framework. IEEE Trans Image Process 2021;30:3167-78.
43. Lee J, Liu P, Cheng J, Fu H. A Deep Step Pattern Representation for Multimodal Retinal Image Registration. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019:5076-85.
44. Li Z, Huang F, Zhang J, Dashtbozorg B, Abbasi-Sureshjani S, Sun Y, Long X, Yu Q, Romeny BTH, Tan T. Multi-modal and multi-vendor retina image registration. Biomed Opt Express 2018;9:410-22.
45. Luo G, Chen X, Shi F, Peng Y, Xiang D, Chen Q, Xu X, Zhu W, Fan Y. Multimodal affine registration for ICGA and MCSL fundus images of high myopia. Biomed Opt Express 2020;11:4443-57.
46. Mishchuk A, Mishkin D, Radenović F, Matas J. Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. In: Advances in Neural Information Processing Systems, 2017. Red Hook, NY, USA: Curran Associates Inc., 2017:4829-40.
47. Fischler M, Bolles R. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 1981;24:381-95.
48. Sarlin P-E, DeTone D, Malisiewicz T, Rabinovich A. Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020:4938-47.
49. Kuang Z, Li J, He M, Wang T, Zhao Y. DenseGAP: Graph-Structured Dense Correspondence Learning with Anchor Points. In: 2022 26th International Conference on Pattern Recognition (ICPR), 2022:542-9.
50. Sun J, Shen Z, Wang Y, Bao H, Zhou X. LoFTR: Detector-Free Local Feature Matching With Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021:8922-31.
51. Chen H, Luo Z, Zhou L, Tian Y, Zhen M, Fang T, McKinnon D, Tsin Y, Quan L. Aspanformer: Detector-free image matching with adaptive span transformer. In: Computer Vision--ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23--27, 2022, Proceedings, Part XXXII, 2022:20-36.
52. An C, Wang Y, Zhang J, Nguyen TQ. Self-Supervised Rigid Registration for Multimodal Retinal Images. IEEE Trans Image Process 2022;31:5733-47.
53. Arikan M, Sadeghipour A, Gerendas B, Told R, Schmidt-Erfurt U. Deep learning based multi-modal registration for retinal imaging. In: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9, 2019:75-82.
54. Tian Y, Fan B, Wu F. L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:6128-36.
55. He K, Lu Y, Sclaroff S. Local Descriptors Optimized for Average Precision. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018:596-605.
56. Forsyth D, Ponce J. Computer vision: a modern approach. Pearson, 2012.
57. Brannan DA, Esplen MF, Gray JJ. Geometry. 2 ed. Cambridge University Press, 2012.
58. Molina-Casado JM, Carmona EJ, García-Feijoó J. Fast detection of the main anatomical structures in digital retinal images based on intra- and inter-structure relational knowledge. Comput Methods Programs Biomed 2017;149:55-68.

59. Abràmoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, Massin P, Cochener B, Gain P, Tang L, Lamard M, Moga DC, Quellec G, Niemeijer M. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol 2013;131:351-7.
60. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, Gain P, Ordonez R, Massin P, Erginay A, Charton B, Klein J-C. Feedback on a publicly distributed image database: the messidor database. Image Anal & Stereol 2014;33:231-4.
61. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: 2015 IEEE International Conference on Computer Vision (ICCV), 2015:1026-34.
62. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR), 2015.
63. Hervella ÁS, Rouco J, Novo J, Ortega M. Self-supervised multimodal reconstruction of retinal images over paired datasets. Expert Syst with Appl 2020;161:113674.
64. Wang J, Chen J, Xu H, Zhang S, Mei X, Huang J, Ma J. Gaussian field estimator with manifold regularization for retinal image registration. Signal Process 2019;157:225-35.
65. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-Up Robust Features (SURF). Comput Vis Image Underst 2008;110:346-59.
66. Chen L, Xiang Y, Chen Y, Zhang X. Retinal image registration using bifurcation structures. In: 2011 18th IEEE International Conference on Image Processing, 2011:2169-72.
67. Braun D, Yang S, Martel JN, Riviere CN, Becker BC. EyeSLAM: Real-time simultaneous localization and mapping of retinal vessels during intraocular microsurgery. Int J Med Robot 2018.
68. Serradell E, Pinheiro MA, Sznitman R, Kybic J, Moreno-Noguer F, Fua P. Non-Rigid Graph Registration Using Active Testing Search. IEEE Trans Pattern Anal Mach Intell 2015;37:625-38.
69. Verkicharla PK, Suheimat M, Schmid KL, Atchison DA. Differences in retinal shape between East Asian and Caucasian eyes. Ophthalmic Physiol Opt 2017;37:275-83.
70. Verkicharla PK, Mathur A, Mallen EA, Pope JM, Atchison DA. Eye shape and retinal shape, and their relation to peripheral refraction. Ophthalmic Physiol Opt 2012;32:184-99.

Cite this article as: Rivas-Villar D, Hervella ÁS, Rouco J, Novo J. Joint keypoint detection and description network for color fundus image registration. Quant Imaging Med Surg 2023;13(7):4540-4562. doi: 10.21037/qims-23-4