Neural Computing and Applications (2023) 35:14779–14797
https://doi.org/10.1007/s00521-023-08454-8

ORIGINAL ARTICLE

Deformable registration of multimodal retinal images using a weakly supervised deep learning approach

Javier Martínez-Río[1] · Enrique J. Carmona[1] · Daniel Cancelas[1] · Jorge Novo[2,3] · Marcos Ortega[2,3]

Received: 6 June 2022 / Accepted: 3 March 2023 / Published online: 28 March 2023

Abstract

There are different retinal vascular imaging modalities widely used in clinical practice to diagnose different retinal pathologies. The joint analysis of these multimodal images is of increasing interest since each of them provides common and complementary visual information. However, if we want to facilitate the comparison of two images, obtained with different techniques and containing the same retinal region of interest, it will be necessary to make a previous registration of both images. Here, we present a weakly supervised deep learning methodology for robust deformable registration of multimodal retinal images, which is applied to implement a method for the registration of fluorescein angiography (FA) and optical coherence tomography angiography (OCTA) images. This methodology is strongly inspired by VoxelMorph, a general unsupervised deep learning framework of the state of the art for deformable registration of unimodal medical images. The method was evaluated in a public dataset with 172 pairs of FA and superficial plexus OCTA images. The degree of alignment of the common information (blood vessels) and preservation of the non-common information (image background) in the transformed image were measured using the Dice coefficient (DC) and zero-normalized cross-correlation (ZNCC), respectively. The average values of the mentioned metrics, including the standard deviations, were DC = 0.72 $\pm$ 0.10 and ZNCC = 0.82 $\pm$ 0.04. The time required to obtain each pair of registered images was 0.12 s. These results outperform rigid and deformable registration methods with which our method was compared.

Keywords Multimodal image registration · Diffeomorphic transformation · Deep learning · VoxelMorph · OCT angiography · Fluorescein angiography

✉ Enrique J. Carmona
ecarmona@dia.uned.es

Javier Martínez-Río
javimdr@dia.uned.es

Daniel Cancelas
dcberlinchez@dia.uned.es

Jorge Novo
j.novo@udc.es

Marcos Ortega
m.ortega@udc.es

[1] Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, 28040 Madrid, Spain

[2] Department of Computer Science and Information Technologies, University of A Coruña, Campus de Elviña s/n, 15008 A Coruña, Spain

[3] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

1 Introduction

Retinal imaging has developed rapidly and is nowadays a fundamental tool for the detection and monitoring of different retinal diseases. Examples of these image techniques are color fundus photography (CFP) [1], hyperspectral retinal imaging (HRI) [2], fundus autofluorescence (FAF) [3], optical coherence tomography (OCT) [4], fluorescein angiography (FA) [5], indocyanine green angiography (ICGA) [6], and optical coherence tomography angiography (OCTA) [7]. In clinical practice, there is a growing interest in using multimodal approaches. The combination of these imaging modalities helps to improve the identification, evolution, or grade of different retinal pathologies, given that some anatomical or pathological structures are better recognized in a particular modality than they are in others.

The joint analysis of two different multimodal medical images requires of a previous step consisting in registering both images. However, the problem of registering pairs of multimodal medical images in general, and pairs of retinal images in particular, is more challenging than the one using unimodal medical images. In the multimodal problem, a part of information which is associated with one of the two images is not contained in the other, and vice versa. We will assume here that the pair of multimodal retinal images to be registered always presents a common structural information that is represented by blood vessels. Otherwise, the registration process would not make sense since there would be no information to guide that process. Specifically, all the retinal imaging techniques mentioned above fulfill this assumption.

As a case study, we chose the FA and OCTA image registration problem. It is an example of multimodal image registration problem that is interesting from a clinical point of view. FA is an invasive imaging modality in which dye is injected into the bloodstream to highlight those vessels belonging to the outermost retina layer. On the other hand, OCTA is a noninvasive imaging modality that produces a retinal blood flow image by comparing the differences between repeated OCT cross-sections of a given location in the retina. FA images have a wider field of view, but a smaller level of detail than OCTA images. Typical fields of view used in clinical practice for OCTA are 3×3 mm and 6×6 mm. Both techniques are widely used to study the functioning of the retinal microcirculation and, consequently, to diagnose and grade relevant ophthalmological diseases such as diabetic retinopathy and age-related macular degeneration, among others. The information provided by both techniques is complementary, and each one of them has advantages and disadvantages [8–10]. Although OCTA can display information at different depths of the retina, registration between the two imaging modalities considered is only possible when superficial plexus OCTA images are used, since FA only provides information of the superficial retinal plexus.

Regarding multimodal image registration techniques, those based on deep learning have been gaining more and more importance in recent years. When deep learning techniques are selected, registration ground truth is typically used in supervised end-to-end registration (SE2ER) approaches. However, this type of ground-truth is difficult to obtain, especially if the transformation to learn is deformable, since, in that case, the required ground-truth quality has to be very high. For this reason, the called weakly supervised end-to-end registration (WSE2ER) approaches have recently gained importance over SE2ER [11]. The term *weakly supervised* (or *semi-supervised*) refers to the fact that the network learns to obtain the registration transformation, but the used ground truth corresponds to partial or approximated segmentation knowledge. This rough knowledge is used by the network during training to detect landmarks between each pair of input images and thus guide the registration process. The WSE2ER paradigm is well suited to the medical image registration problem because, although in this type of domain, the segmentation ground truth is difficult to obtain manually, it can be approximately obtained using automatic segmentation methods.

It is also interesting to note that deformable image registration is more powerful and flexible than rigid one, given that the former is able to learn a correspondence (deformation field) between every pair of pixels to register, while the latter is limited by the type of transformation learned, generally linear.

Considering the advantages associated with deformable transformations and weakly supervised learning, in this work, we present a deep learning-based general methodology for deformable registration of multimodal retinal images based on WSE2ER. Our methodology is strongly inspirated by VoxelMorph [12, 13], a state-of-the-art general learning framework for deformable unsupervised end-to-end registration (UE2ER). The main contribution of our work is twofold. First, we show that is possible to adapt VoxelMorph to multimodal image domains. And second, we describe how to implement a deformable multimodal WSE2ER approach that proves to be competitive when applied to the registration problem of FA and superficial plexus OCTA images. Note that, to the best of our knowledge, VoxelMorph has only been used with unimodal medical images, and furthermore, there are no studies in the related literature devoted to registration of FA and OCTA images using deformable transformations. In any case, the methodology presented here is also valid for registering pairs of images belonging to any of the retinal imaging modalities mentioned at the beginning of this section.

The rest of the manuscript is organized as follows. Section 2 describes those approaches from the literature related with our proposal. Section 3 explains the features of the image dataset used in different experiments and the proposed methodology. In Sect. 4, the obtained results are shown, analyzed, and discussed. Finally, Sect. 5 summarizes the main conclusions of this work.

2 Related works

There are many approaches related to medical image registration based on deep learning. A comprehensive review is presented in [11]. Although these techniques require a hard training process, once done, image registration can be quickly and accurately calculated. Like our proposal, there

are other interesting WSE2ER approaches in the related literature that apply to multimodal medical imaging [14–16], but in practice they have been applied to other medical domains or imaging modalities than those used here. This prevents us from making a fair comparison between our proposal and these methods.

More specifically, considering the case of deep learning-based multimodal retinal image registration, there are works where rigid or deformable transformations are obtained using supervised, weakly supervised, or unsupervised approaches [17–23]. While deformable methods [19, 22, 23] are more competitive than rigid ones [17, 18, 20, 21], the former generally require that the input image pair is already approximately registered (usually via an affine transformation). To do the latter, there are two options: incorporate this stage into the methodology itself or assume that this step was previously performed with an external method. The first option increases the complexity of the proposed methodology but avoids having to use additional methods. On the other hand, the second option is simpler since it makes both types of registrations (approximate and deformable) independent, allowing us to use any of the numerous methods already existing in the literature to obtain pairs of approximately registered images.

In relation to the unsupervised methods [18, 19, 23], they have the advantage of not requiring segmentation knowledge, but in the context of multimodal images, the absence of this type of knowledge tends to reduce the accuracy of the obtained registration. As for the purely supervised methods [17, 20], the requirement of a high-quality ground-truth can hinder their applicability, especially in a context associated with medical images and their daily use in clinical practice. Finally, regarding the weakly supervised methods [21, 22], such as the one proposed here, they allow us to solve the shortcomings of the previous two (supervised and unsupervised), given that the segmentation knowledge required may be partial and imprecise (it may even contain noise), but it is still useful enough to guide and improve the registration process.

Some of the studies mentioned above are of special interest because they are very related to the characteristics of the method we propose or to the case study addressed here. For example, in [22], a weakly supervised approach is used for registration of HRI image pairs and, simultaneously, an estimation of the blood vessel segmentation is also done. Each of these two tasks is carried out by a different neural network: The blood vessel ground truth of the pair of multispectral images is used for fully supervised training of the segmentation network and weakly supervised learning of the registration network. Additionally, the prediction of the registration network, when is applied to the pair of vessel images (associated with the pair of images to be registered), is also used as the ground-truth to train the segmentation network, making use of the unsupervised adversarial training. Although the idea is interesting and, as our proposal, a weakly supervised mechanism is used to learn the registration network, two networks have to be learned and the used ground truth (blood vessel maps) have to be manually labeled. In our case, as will be seen in Sec. 3, the process is simpler because a single neural network has to be tuned, and the vessel masks used in the training process can be approximately obtained (in fact, we use a simple segmentation method to obtain them).

On the other hand, a two-step unsupervised learning framework based on deep convolutional networks is proposed in [23], which is used to register CFP and FA images. In the first step, three sequentially connected networks are used to carry out a coarse alignment based on vessel segmentation information. In the second step, the alignment obtained from the previous step is refined, using a deformable registration network which is aided by two modality transformers to guide registration. Unlike the weakly supervised approach of our proposal, the unsupervised nature of the approach described in [23] is interesting because it does not require ground-truth knowledge to carry out the training. However, it requires the training of four neural networks: three of them to obtain the parameters of an affine transformation that performs an approximate registration, and one more network, aided by two modal transformers, that learns the displacement field necessary to improve the final alignment accuracy.

Finally, in relation to our case study, the FA and OCTA image registration problem has been addressed manually or semi-automatically [24–27], but it has recently started to be solved automatically [17, 28, 29]. However, all existing automated approaches in the related medical literature are based on rigid transformations. As mentioned above, deformable transformations allow us to obtain more exact registrations than the rigid ones, especially when the pair of multimodal images to register cover fields of view with very different sizes, as occurs, for example, in the case of FA and OCTA images.

3 Materials and methods

In this section, we describe the methodology proposed to address the multimodal retinal image registration problem. This methodology is instantiated into a method for registering FA and OCTA images. The dataset containing the FA and OCTA images is also presented.

3.1 Dataset

The method's performance was evaluated using the same dataset used in [29], which is publicly available in [30]. Basically, the dataset contains a total of 86 cases that were previously anonymized: 31 healthy and 55 pathological. Each case corresponds to a patient's eye and consists of one FA image (1536×1536 pixels) and two superficial capillary plexus OCTA images which correspond to two different zoom levels: 3×3mm (320×320 pixels) and 6×6mm (320×320 pixels). Therefore, a total of 172 registrations are possible (86 pairs of FA and $\text{OCTA}_{3\times3}$ images plus 86 pairs of FA and $\text{OCTA}_{6\times6}$ images). In all the cases, the area encompassed by each OCTA image is contained in its respective FA image. Additionally, some of the OCTA images present typical motion artifacts that are produced by involuntary eye motions that occur during the capture time [31]. These artifacts are characterized as horizontal or vertical white lines that disrupt the continuity of the vessels. It is important to highlight that the images contained in this dataset were obtained under real conditions, that is, in the daily scenario of clinical practice. This makes the requirements for the registration method more challenging but, on the other hand, more realistic. Table 1 summarizes the distribution of the FA and OCTA image pairs of the dataset according to the different features mentioned above, and Fig. 1 shows two examples of healthy and pathological eye cases, illustrating different scenarios. More information about this dataset can be found in [29].

3.2 Method

We start by describing the generic methodology proposed to address the multimodal retinal image registration problem, and then, we show how to instantiate it to be applied to the particular problem of registering FA and OCTA images.

Table 1 Distribution of the pairs of FA and OCTA images based on different characteristics of these images: healthy or pathological eye, OCTA image zoom level (3×3 or 6×6), and presence or absence of motion artifacts in the OCTA images

	3×3 OCTA		6×6 OCTA	
	No-artifacts	Artifacts	No-artifacts	Artifacts
Healthy	23	8	25	6
Pathological	48	7	48	7

3.2.1 Description of the methodology

Usually, *deformable* registration strategies, also called *nonlinear* or *non-rigid*, involve two stages. First, an initial affine transformation is made to obtain an approximate registration. Second, a dense deformable transformation is applied, taking the output of the first stage as the input. In this work, we will only focus on the second stage, and therefore, we assume that the images to be non-rigidly registered are already affinely aligned. The consideration of these two stages separately is especially important when the alignment of the two images to be registered implies a strong component of rotation, scale, translation, or a mixture of two or more of these basic transformations. For example, this situation occurs when the field of view covered by one of the images in the pair is much larger than the one covered by the other (see Fig. 1).

Let F and M be the fixed and moving multimodal retinal images, respectively, and let ψ be the registration field that maps coordinates of M to coordinates of F. According to the VoxelMorph unsupervised framework [13], the problem of deformable registration may be seen as an optimization problem:

$$\arg\min_{\psi} \{ f_{\text{sim}}(F, \psi(M)) + \lambda f_{\text{reg}}(\psi) \}, \tag{1}$$

where $\psi(*)$ is the searched deformable transformation, $\psi(M)$ represents the result of transforming M using ψ, $f_{\text{sim}}(*,*)$ measures image similarity between its two inputs, and $f_{\text{reg}}(*)$ adds a regularization mechanism weighted by the parameter λ.

In addition, due to the particular characteristics of the registration problem addressed here (common information corresponds to blood vessel information), we propose to include, into the function to minimize, a similarity measure of the vascular network segmentation between F and M. Here, we will assume that a rough segmentation of the vessels will suffice to guarantee the weakly supervised nature of the proposed methodology. In this way, the new optimization problem is given by:

$$\arg\min_{\psi} \{ \gamma f_{sim}(F_{\text{seg}}, \psi(M_{\text{seg}})) + (1-\gamma) f_{\text{sim}} (F, \psi(M)) + \lambda f_{\text{reg}}(\psi) \}, \tag{2}$$

where F_{seg} and M_{seg} represent a rough segmentation of the vascular network in F and M, respectively, and γ is a parameter that allows us to weight the results of comparing the similarity between the pairs of images (F, M) and $(F_{\text{seg}}, M_{\text{seg}})$. Note that, from a general point of view, the two similarity functions used in Eq. (2) do not have to be the same.

In order to learn ψ, it is proposed to use a convolutional neural network (CNN). In a first approximation, as is done

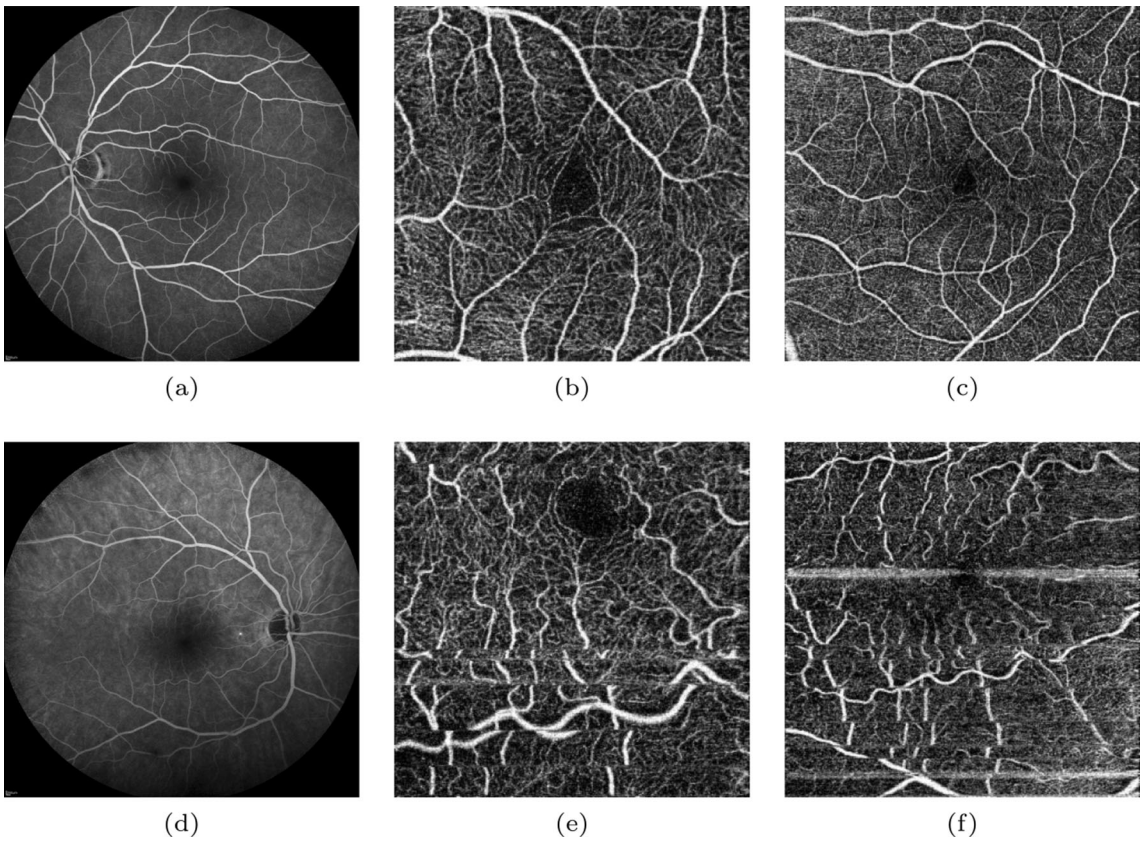

Fig. 1 Examples of healthy and pathological eye cases (first and second row, respectively): **a** FA image; **b** and **c** 3 × 3mm and 6 × 6mm superficial plexus OCTA image without artifacts, respectively; **d** FA image; **e** and **f** 3 × 3mm and 6 × 6mm superficial plexus OCTA image with artifacts, respectively

in [13], the CNN can be used to directly learn a function $h_\kappa(F, M) = \mathbf{r}$, where κ represents the network parameters and, $\mathbf{r}$, a displacement field (DF). In this context, let Ω be a n-dimensional spatial domain in which M, F, M_{seg} and F_{seg} are defined, then, for each pixel $\mathbf{p} \in \Omega \subset \mathbb{R}^n$, $\mathbf{r}(\mathbf{p})$ represents a displacement such that $F(\mathbf{p})$ and $\psi(M(\mathbf{p}))$ should correspond to similar locations, being $\psi = I + \mathbf{r}$ and, I, the identity transformation. From now on, the module dedicated to calculating $\psi(M(\mathbf{p}))$ will be called the *transformation layer*. However, there is no guarantee of getting a smooth ψ in this learning process and, for example, two or more pixels could collapse into only one transformed pixel and vice versa. Therefore, to avoid this inconvenience and obtain a more realistic smooth ψ, we decided to work with diffeomorphic transformations. Note that every diffeomorphic deformation is bijective, and itself and its inverse are differentiable. Consequently, this type of transformation always preserves the topology and guarantees that the DF will be smooth. Here, the idea is to parameterize the deformation field with a stationary velocity field (SVF), $\mathbf{v}$, and integrate it within the network to obtain a diffeomorphism [12, 32]. This integration operation will be done in a new module denominated the *integration layer*.

A graphical overview of the proposed general method is shown in Fig. 2. As indicated in the figure, the network can be used in two different modes: *training* and *inference mode*. In the former, the network weights are learned using the following loss function:

$$F_{\text{loss}} = \gamma f_{\text{sim}}(F_{\text{seg}}, \psi(M_{\text{seg}})) + (1-\gamma) f_{\text{sim}}(F, \psi(M)) + \lambda f_{\text{reg}}(\psi). \tag{3}$$

Note that this function is the same as the one to be minimized in our optimization problem (see Eq. (2)), but it is different to the one proposes in VoxelMorph [13]. Here, the idea is to study the influence of each similarity functions in the method performance, including the more extreme cases ($\gamma = 0$ and $\gamma = 1$). In the training mode, the network learns how to obtain a SVF by each pair of F and M images. In turn, each SVF thus obtained is integrated to obtain its respective DF and is also used to further encourage the achievement of a smoother DF by mean of the regularization mechanism. Finally, each DF is used to calculate the transformed images $\psi(M)$ and $\psi(M_{\text{seg}})$, which are used to calculate the similarity with F and F_{seg}, respectively.

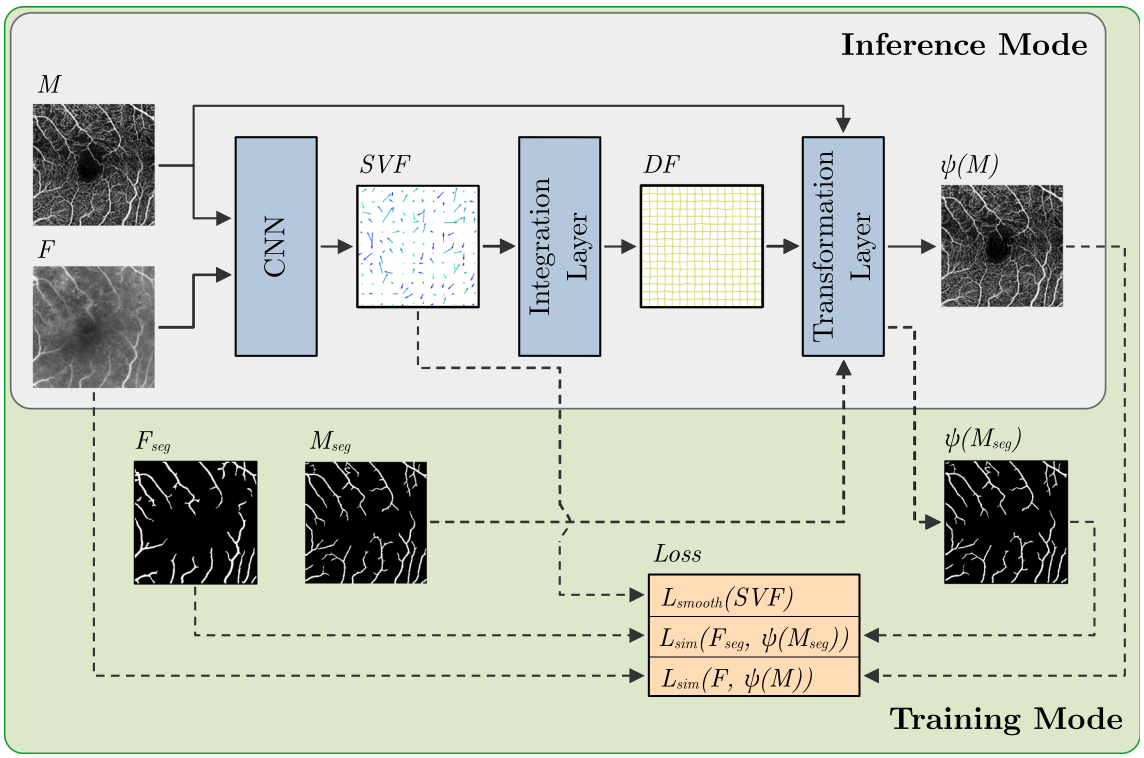

Fig. 2 Overview of the proposed multimodal image registration method. The *training mode* allows the convolutional neural network (CNN) to learn its weights. The *inference mode* allows the CNN to register a given pair of input images. See Sect. 3.2 for more details

On the other hand, in the inference mode, the trained network only uses the pair of images to be registered, F and M, to produce a specific SVF that, once integrated, will produce a specific DF. Finally, this DF will be used to obtain $\psi(M)$, which represent the image aligned with F, thus ending the registration process. Note that, in this mode, the segmented images, F_{seg} and M_{seg}, are not utilized, and therefore, vessel segmentation is no longer necessary to register new pairs of images.

3.2.2 Instantiating the architecture

We now describe how the general architecture described above is instantiated to solve the multimodal registration problem of FA and superficial plexus OCTA images.

On the one hand, we associate F and M with the FA and OCTA images, respectively. They were previously registered by means of an affine transformation using the three-stage method described in [29]. There, the first stage is used to segment the blood vessels in both types of images. Then, the segmentation information is employed to obtain an approximate registration based on template matching. Note that, in the original dataset, the FA images have a wider field of view than the one corresponding to the OCTA images (see Fig. 1). Lastly, an evolutionary algorithm is applied to refine the previous rough alignment and obtain an affine registration. On the other hand, F_{seg} and M_{seg} correspond to the FA and OCTA vessel segmentation images, respectively. These segmentation images are obtained from the output of the first stage of the method described above (see [29] for more information), being M_{seg} the result of applying the corresponding affine transformation to the OCTA vessel segmentation image. Due to the weakly unsupervised nature of the proposed method, we assume that the convolutional neural network will be able to learn to extract landmarks from the vessel information, even when this segmentation is noisy or does not contain all the vessel pixels of the input image pair. In fact, the masks obtained by the utilized segmentation method mainly contain information of the main vessels and, in addition, may include noise.

Inspired by the VoxelMorph architecture, the CNN used here is an auto-encoder. The model is based on a U-Net [33], which consists of encoding and decoding layers with connections between different levels (skip connections). The network receives a single input formed by the concatenation of the FA and OCTA image pair, generating a 2-channel input image of size $H \times W \times 2$, where H and W are the height and width of the images, respectively. The encoder and decoder are made up of convolutions that allow the network to capture the common characteristics to the pair of input images, which will be used to estimate the

SVF. In turn, the integration of SVF will allow the method to obtain the final DF. The convolution layers are made up of a kernel size of 3×3, stride equal to 1, and padding is used so that the output has the same size as the input. The number of filters used in each layer varies, as can be seen in Fig. 3. After each convolution stage, the nonlinear activation function LeakyReLU is used with a coefficient value of 0.2. In the encoding stage, max pooling layers, with a kernel size of 2×2 and stride equal to 2, are applied to reduce the image size by half. At the time of minimum size, we work with image sizes of 1/16 of the original size, that is, 20×20 pixels. Likewise, during the decoding stage, this process is done in reverse, increasing the size of the images in each layer (upsampling) until the original size is restored. Note that, during every step of the decoding path, a concatenation with the correspondingly cropped feature map from the encoding path is made (see dashed arrows in Fig. 3).

Although different representations could be applied for the DF, as mentioned in Sect. 3.2.1, we choose to work with a diffeomorphic transformation, using a previous SVF representation. For the implementation of the diffeomorphic integration layer, we used the same approximation used in [12]. Specifically, the DF is defined through the following ordinary differential equation:

$$\frac{\partial \psi^{(t)}}{\partial t} = \mathbf{v}(\psi^{(t)}), \tag{4}$$

where $\psi(0) = I$ is the identity transformation and t is the time. Then, the SVF (represented by $\mathbf{v}$) is conveniently integrated over $t = [0, 1]$, using *scaling and squaring* as the integration method [34], to obtain the final DF, that is, $\psi(1)$. The only parameter required in this process is the so-called *number of integration steps* that represents the number of scaling-squaring steps to do in the integration process.

The purpose of the spatial transformation layer is to apply the DF, which was estimated by the integration layer,

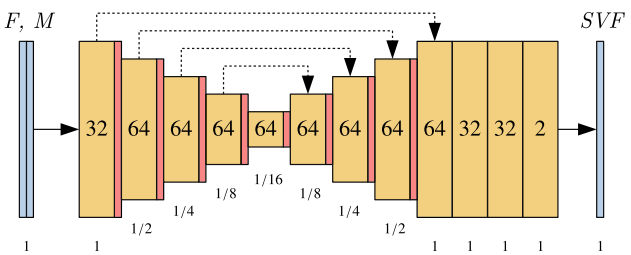

Fig. 3 Overview of the U-Net architecture. The orange boxes represent the convolutional layers with the number of filters used in each case. The red boxes represent the max pooling (encoder) and upsampling (decoder) layers. Each dashed arrow represents a skip connection, that is, a concatenation with the correspondingly cropped feature map from the encoding path. The number under the boxes indicates the size of the images with respect to the original input size

to M and M_{seg} in order to evaluate the differences between the pairs $(F, \psi(M))$ and $(F_{\text{seg}}, \psi(M_{\text{seg}}))$. We implement the same methodology followed in [13] to backpropagate errors during optimization, which uses a differentiable operation based on spatial transformer networks [35] to compute $\psi(M)$ and $\psi(M_{\text{seg}})$. In addition, since for each pixel $\mathbf{p}$ it is not guaranteed that the transformed pixel $\mathbf{p}' = \mathbf{p} + \mathbf{r}(\mathbf{p})$ corresponds to an integer position, a linear interpolation of the image values is done using the neighborhood of $\mathbf{p}'$:

$$\psi(M_i(\mathbf{p})) = \sum_{\mathbf{q} \in N(\mathbf{p}')} \left(M_i(\mathbf{q}) \prod_{d \in \{x,y\}} \left(1 - |\mathbf{p}'_d - \mathbf{q}_d|\right) \right), \tag{5}$$

where $M_i \in \{M, M_{\text{seg}}\}$, $N(\mathbf{p}')$ represents the four neighboring pixels of $\mathbf{p}'$ and the product subscript d iterates over the two directions of Ω.

In relation to the loss function, the metrics used to measure the similarity between the pairs (F, M) and $(F_{\text{seg}}, M_{\text{seg}})$ were the local-normalized cross-correlation and the Dice coefficient, respectively. Although other similarity measures are frequently used in the related literature to compare grayscale images, such as the mean absolute error, root mean square error, mutual information, and correlation, among others, we have selected correlation because, in addition to its accuracy and reliability [36], it has two important properties: it is normalized in the interval $[-1, 1]$ and it is independent of any offset or linear transformation in the set of pixel values to match. Specifically, the local-normalized cross-correlation (LNCC) is calculated by:

$$\text{LNCC}(A, B) = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \text{ZNCC}(W_A(\mathbf{p}), W_B(\mathbf{p})), \tag{6}$$

where $|*|$ expresses the cardinality of a set, A and B are images of equal size, $\Omega \subset \mathbb{R}^2$ is the domain of both images, $W_A(\mathbf{p})$ and $W_B(\mathbf{p})$ are, respectively, subwindows of A and B around the pixel $\mathbf{p}$ and size $n \times n$ (we used $n = 15$), and $ZNCC(*, *)$ is the zero-normalized cross-correlation function that is defined as:

$$\text{ZNCC}(W_A, W_B) = \frac{\sum_{p_i} \left(W_A(p_i) - \overline{W}_A\right)\left(W_B(p_i) - \overline{W}_B\right)}{\left[\sum_{p_i} \left(W_A(p_i) - \overline{W}_A\right)^2\right]^{1/2} \left[\sum_{p_i} \left(W_B(p_i) - \overline{W}_B\right)^2\right]^{1/2}}, \tag{7}$$

being $\overline{W}_A$ and $\overline{W}_B$ the average intensity of all pixels in subwindows A and B, respectively, $W_A(p_i)$ and $W_B(p_i)$ are the value of the i-th pixel in the subwindows A and B, respectively, and p_i is the index of summation, which is defined for each pixel contained in W_A or W_B, as appropriate. Note that Eq. (7) is different from the one used in

the VoxelMorph original version [13], where the numerator and denominator were squared. We decided to use the original definition of ZNCC to penalize values corresponding to negative correlations.

On the other hand, the Dice coefficient (DC) can be viewed as a similarity measure over sets. It ranges between 0 and 1 (the higher the DC value, the greater the similarity), and it is used a lot in medical domains to compare binary segmentation algorithm outputs against ground-truth masks. Alternatively, Jaccard index (JI) could have been used instead, but the performances of both indexes are similar because they are positively correlated: given a DC value, the respective JI value can be calculated and vice versa. Concretely, let A and B be two binary images, then DC is defined as:

$$\mathrm{DC}(A,B) = \frac{2TP}{2TP + FP + FN}, \tag{8}$$

being TP, FP, and FN the true positives, false positives, and false negatives, respectively.

Finally, as is done in [12, 32], we use the output of the integration module, $\mathbf{v}$, to apply a regularization mechanism oriented to obtain a smooth ψ, anticipating that the DF obtained by the integration layer may not be purely diffeomorphic. Specifically, a diffusion regularizer on the 2-norm of gradient of $\mathbf{v}$ is used:

$$f_{\mathrm{reg}}(\psi) = \sum_{\mathbf{p}\in\Omega} \|\nabla \mathbf{v}(\mathbf{p})\|^2, \tag{9}$$

where $\mathbf{p}$ represents the coordinates of each pixel in Ω. Therefore, considering the above definitions for f_{sim} and f_{reg}, Eq. (3) is finally instantiated as:

$$F_{\mathrm{loss}} = -\left[\gamma\,\mathrm{DC}(F_{\mathrm{seg}}, \psi(M_{\mathrm{seg}})) + (1-\gamma)\,\mathrm{LNCC}(F, \psi(M))\right] + \lambda \sum_{\mathbf{p}\in\Omega} \|\nabla \mathbf{v}(\mathbf{p})\|^2. \tag{10}$$

Note that high values of DC and correlation mean a better registration. Therefore, the signs for the terms associated with the DC and LNCC are negative because we are considering a minimization problem.

4 Experiments and results

We present below the different experiments made to evaluate the performance of the proposed architecture when it is applied to the FA and OCTA image registration problem. Before the presentation of results, we also describe some implementation details and the metrics used to evaluate our method.

4.1 Implementation details

As mentioned in Sect. 3.2.2, the previous affine registration of each pair of FA and OCTA images and the FA an OCTA vessel segmented images were obtained using the method proposed in [29]. However, this method fails in the affine alignment of two pairs of images (both corresponding to pathological eyes, containing 3×3 OCTA images with and without artifacts). Therefore, the available set of affinely registered image pairs contains 170 of the 172 pairs belonging to the original dataset. All these images were finally rescaled to the size of the input of the network, that is, $H \times W$ was set to 320×320 pixels. Additionally, in order not to provide artifact noisy information in the CNN training process, we left out the pairs of images containing artifacts (27/170). The performance of the trained model in this type of image is discussed in Sect. 4.3.

Due to the not very high number of training image pairs without artifacts ($170 - 27 = 143$), 5-fold cross-validation and data augmentation were applied. Specifically, in each CNN training epoch, a random rigid or affine transformation was applied to each pair of original training images, including pairs of grayscale and vessel segmentation images. In this way, the original training images were never seen by the network at any time, since it is very unlikely that any of the randomly generated transformations corresponds to the identity matrix. The batch size used in each training and the total number of epochs were set equal to 10 and 200, respectively. An Adam optimizer was used with the following configuration: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate equal to 10^{-4}. In relation to the number of integration steps used in the integration layer, we observed that the method performance increased when the value of this parameter also did; finally, a value equal to 10 was used since we did not notice improvements above this threshold.

The implemented Python code was an adaptation of the original VoxelMorph code [37], where the particularities of our approach were included. All the experiments were run on a computer with a Nvidia Tesla V100 16GB GPU.

4.2 Evaluation metrics

When working with multimodal image registration problems, it is non-trivial to evaluate the quality of the obtained registration. In a first approximation, we should measure the degree of alignment of the common information (blood vessels in our case) between F_{seg} (FA vessel image) and $\psi(M_{\mathrm{seg}})$ (transformed OCTA vessel image). Here, we propose to use the DC (see Eq. (8)) to evaluate that degree of alignment. It is important to highlight that, even if the alignment obtained were perfect, it would be practically

Table 2 Tuning hyperparameters: mean and standard deviation of the Dice coefficient, DC(FA_{seg}, $\psi(OCTA_{seg})$), zero-normalized cross-correlation, ZNCC(OCTA,ψ(OCTA)), and number of pixel (percentage in parentheses) where $|\mathbf{J}_\psi|_{\leq 0}$, using 5-fold cross-validation for different configurations of the hyperparameters γ (compromise between segmentation and grayscale information) and λ (weight of the regularization)

γ	λ	DC	ZNCC	$\lvert\mathbf{J}_\psi\rvert_{\leq 0}$ (%)
0.00	1	0.6724 ± 0.0151	0.8209 ± 0.0121	0.00 ± 0.00 (0.00)
	0.1	0.6666 ± 0.0150	0.7381 ± 0.0121	0.15 ± 0.06 (0.00)
	0.01	0.6615 ± 0.0140	0.6919 ± 0.0136	20.06 ± 7.90 (0.01)
	0.001	0.6582 ± 0.0139	0.6712 ± 0.0115	88.85 ± 24.83 (0.08)
	0	0.6569 ± 0.0144	0.6681 ± 0.0106	114.50 ± 30.07 (0.11)
0.25	1	0.6856 ± 0.0158	0.8228 ± 0.0079	0.00 ± 0.00 (0.00)
	0.1	0.6957 ± 0.0151	0.7587 ± 0.0082	0.05 ± 0.05 (0.00)
	0.01	0.6982 ± 0.0146	0.7022 ± 0.0077	11.39 ± 2.32 (0.01)
	0.001	0.6970 ± 0.0147	0.6825 ± 0.0052	68.66 ± 11.08 (0.06)
	0	0.6970 ± 0.0146	0.6778 ± 0.0048	99.17 ± 16.39 (0.09)
0.50	1	0.6962 ± 0.0174	0.8291 ± 0.0070	0.00 ± 0.00 (0.00)
	0.1	0.7113 ± 0.0162	0.7525 ± 0.0076	0.04 ± 0.06 (0.00)
	0.01	0.7178 ± 0.0154	0.6951 ± 0.0075	7.13 ± 2.06 (0.00)
	0.001	0.7182 ± 0.0151	0.6695 ± 0.0104	80.69 ± 17.17 (0.07)
	0	0.7178 ± 0.0151	0.6658 ± 0.0108	126.41 ± 25.07 (0.12)
0.75	1	0.7026 ± 0.0165	0.8322 ± 0.0068	0.00 ± 0.00 (0.00)
	0.1	0.7275 ± 0.0175	0.7484 ± 0.0026	0.00 ± 0.00 (0.00)
	0.01	0.7359 ± 0.0169	0.6792 ± 0.0064	4.42 ± 1.35 (0.00)
	0.001	0.7358 ± 0.0157	0.6509 ± 0.0081	127.06 ± 23.84 (0.12)
	0	0.7350 ± 0.0155	0.6437 ± 0.0085	280.78 ± 45.05 (0.27)
1.00	1	0.7094 ± 0.0172	0.8381 ± 0.0039	0.00 ± 0.00 (0.00)
	0.1	0.7368 ± 0.0179	0.7120 ± 0.0070	0.00 ± 0.00 (0.00)
	0.01	0.7450 ± 0.0170	0.6303 ± 0.0076	3.64 ± 1.52 (0.00)
	0.001	0.7463 ± 0.0162	0.6090 ± 0.0051	294.00 ± 41.17 (0.28)
	0	0.7454 ± 0.0157	0.6007 ± 0.0059	1145.04 ± 174.63 (1.11)

impossible to obtain a DC = 1. This is because the vessel mask pair, automatically obtained from each pair of images, rarely contain the same information. For example, the presence of noise in the grayscale images can be transmitted to the masks or the amount of vessel information displayed by each grayscale image pair can be different. In fact, the OCTA image usually shows more vessels than the FA image because the resolution of the former is greater than the one of the latter.

However, the metric described above is not enough since we should also check that the non-common information of the OCTA image is not altered after carrying out the transformation. Otherwise, we could lose the background information of the OCTA image modality. In order to measure the invariance degree of this type of information, we propose to evaluate $\mathrm{ZNCC}(M, \psi(M))$, being M the OCTA image (see Eq. (7)). Note that, in our case, the quantity of common information (vessel pixels) is much less than the one associated with the complementary information (image background). Therefore, a large value of ZNCC will imply, approximately, a strong similarity between the background information of the OCTA image and its transformed image. In the end, since we want to achieve both goals (alignment of the common information and conservation of the non-common information), we will have to simultaneously handle these two metrics during the evaluation of the method.

The smoothness of the deformation field is also evaluated, making use of its Jacobian matrix, $\mathbf{J}_{\psi(\mathbf{p})} = \nabla\psi(\mathbf{p}) \in \mathbb{R}^{2\times 2}$, which allows us to analyze important information about the behavior of ψ around each pixel $\mathbf{p}$. From a general point of view, a continuously differentiable function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is invertible near a point $\mathbf{p}$ if the Jacobian determinant at $\mathbf{p}$, $|\mathbf{J}_{\mathbf{f}(\mathbf{p})}|$, is nonzero. Furthermore, if $|\mathbf{J}_{\mathbf{f}(\mathbf{p})}| > 0$, then $\mathbf{f}$ preserves orientation near $\mathbf{p}$ ($\mathbf{f}$ expands volumes). Conversely, if $|\mathbf{J}_{\mathbf{f}(\mathbf{p})}| < 0$, then $\mathbf{f}$ reverses orientation ($\mathbf{f}$ shrinks volumes). Therefore, in our case, counting the number of pixels whose transformation matrix Jacobian determinant is non-positive, $|\mathbf{J}_\psi|_{\leq 0}$, we will get a measure about how diffeomorphic the obtained transformation is: the transformation will be diffeomorphic if all the pixels satisfy that $|\mathbf{J}_\psi| > 0$). Otherwise, the higher the number of pixels with $|\mathbf{J}_\psi|_{\leq 0}$, the less smooth the DF will be.

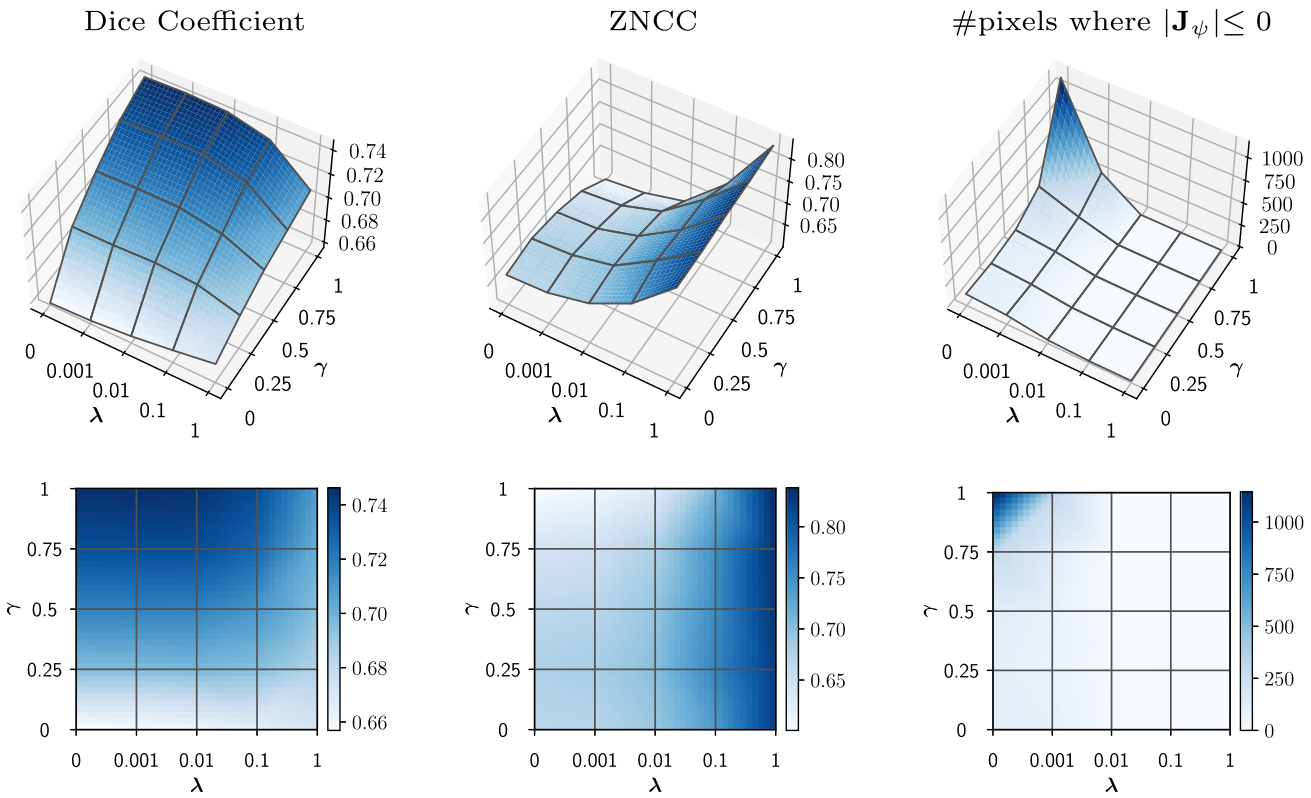

Fig. 4 Tuning hyperparameters: first and second rows show 3D and 2D color plots, respectively, which extrapolate the data shown in Table 2 for the Dice coefficient (first column), zero-normalized cross-correlation (second column), and number of pixels whose Jacobian determinant is non-positive (third column) (color figure online)

4.3 Results and discussion

In this section, we show and discuss the results obtained in different experiments. The first of them analyzes the best configuration of the hyperparameters associated with our method. In the second and third experiments, we study how the registration method performs when images with and without artifacts are used, respectively. Finally, in the last experiment, a comparison with other registration methods is made.

4.3.1 Tuning hyperparameters

First of all, we investigate the dependence of our method on its main hyperparameters (λ and γ). The idea is to analyze the influence of the regularization term and the compromise using segmentation and grayscale similarity information during the training process. Concretely, when $\gamma = 1$, the similarity between vessel mask pairs is only used in the loss function, that is, the similarity between grayscale image pairs is not considered; contrarily, when $\gamma = 0$, just the opposite happens. The rest of intermediate values represent different cases between these two extreme scenarios. In a similar way, a variation in the λ value from 0 to 1 allows us to study the influence of the regularization term in the learning process. Table 2 shows the results of this experiment. Each row of the table contains the result of evaluating the method for each hyperparameter configuration (γ and λ), using data augmentation and 5-fold cross-validation, as mentioned in Sect. 4.1. Given that we are using 5-fold cross-validation, we provide the mean and standard deviation of each evaluation metric considering the results of the five cross-validation sub-models. The results of Table 2, without considering standard deviation values, are also graphically represented in Fig. 4. Specifically, 3D- and 2D-color plots are used for extrapolating the rest of the DC, ZNCC, and $|\mathbf{J}_\psi|_{\leq 0}$ values when γ and λ vary in the interval [0, 1].

We can start by looking at the global behavior of the λ parameter. From Table 2 and Fig. 4, it is easy to see that the DC and ZNCC values are negatively and positively correlated, respectively, with the λ values. This behavior reveals that when the regularization mechanism becomes more important, it interferes with the alignment of the common information, but favors the conservation of the background information in the transformed image. On the other hand, the number of pixels where $|\mathbf{J}_\psi|_{\leq 0}$ is negatively correlated with the λ values. This result was

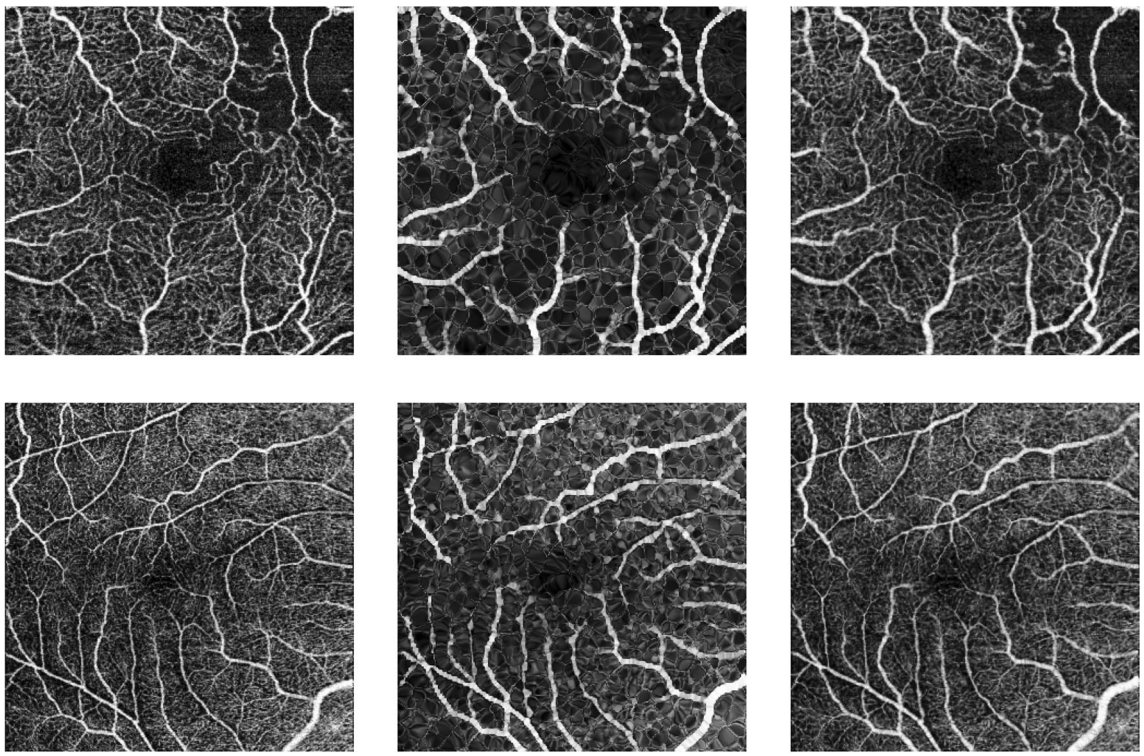

Fig. 5 Different examples showing the effect of the value λ (assuming $\gamma = 1$) in the transformed OCTA image background information. In each row, the first column shows the input OCTA image, and the second and third columns display the result of applying our deformation map to its respective OCTA image using a configuration (λ, γ) equal to (0.001, 1) and (1, 1), respectively

expected since the regularization term becomes more important when λ increases.

Having analyzed the behavior of λ, we now include the parameter γ in the discussion. Now, the idea is to establish the most promising region for our registration method. In principle, we will bias the choice of the best configuration by analyzing the value of DC, since this is the metric that best reflects the degree of alignment between the common information of each pair of images. Thus, considering any value of λ and holding it fixed, the best result is always obtained for $\gamma = 1$. Therefore, we can conclude that, during the training process, it is necessary to include in the loss function the similarity between the vessel mask pairs but not the explicit similarity between the grayscale image pairs. More specifically, the configuration $\gamma = 1$ and $\lambda = 0.001$ is the one that provides the best DC result (0.7463). However, this configuration has an undesirable effect, which is observed in Fig. 5: comparing the first and second columns, we can see that the background information in the transformed OCTA image is considerably altered (many artifacts are visible). This behavior is explained by the low ZNCC value (0.609) associated with the mentioned configuration. However, given that the ZNCC value increases as λ does, the disappearance of noise and artifacts in $\psi(\text{OCTA})$ should run parallel to this trend. Therefore, we must find a compromise between the DC and ZNCC. By looking at the respective 2D color plots (see Fig. 4), we obtain the solution: the region where the intensity is high in both plots corresponds to $\gamma \geq 0.75$ and $\lambda \geq 0.1$. This result is also consistent with the 2D Jacobian determinant color plot (see Fig. 4), where the transformation is guaranteed to be diffeomorphic in the mentioned region. The experimental evidence that reinforces this argument is shown in the third column of Fig. 5, where the configuration $\gamma = 1$ and $\lambda = 1$ is used. We can see how the noise and artifacts disappear in the transformed OCTA image when compared with the second column and, equally, how the background information is conserved when compared with the first column.

It is also interesting to analyze the region where $\gamma = 0$ and $\lambda \in [0, 1]$, which corresponds to an unsupervised training of the CNN. This case corresponds to an approach based on UE2ER, that is, the original framework of VoxelMorph. Now, we should consider $\lambda > 0.1$ because, otherwise, the transformation starts to be non-diffeomorphic, that is, the number of pixels with $|\mathbf{J}_\psi|_{\leq 0}$ becomes greater than zero (see Table 2). Therefore, focusing on the region of interest, $\gamma = 0$ and $\lambda > 0.1$ (see Fig. 4), a high preservation of the background information in the transformed image is expected (high ZNCC values), but at the cost of a moderate degree of vessel alignment (intermediate DC values).

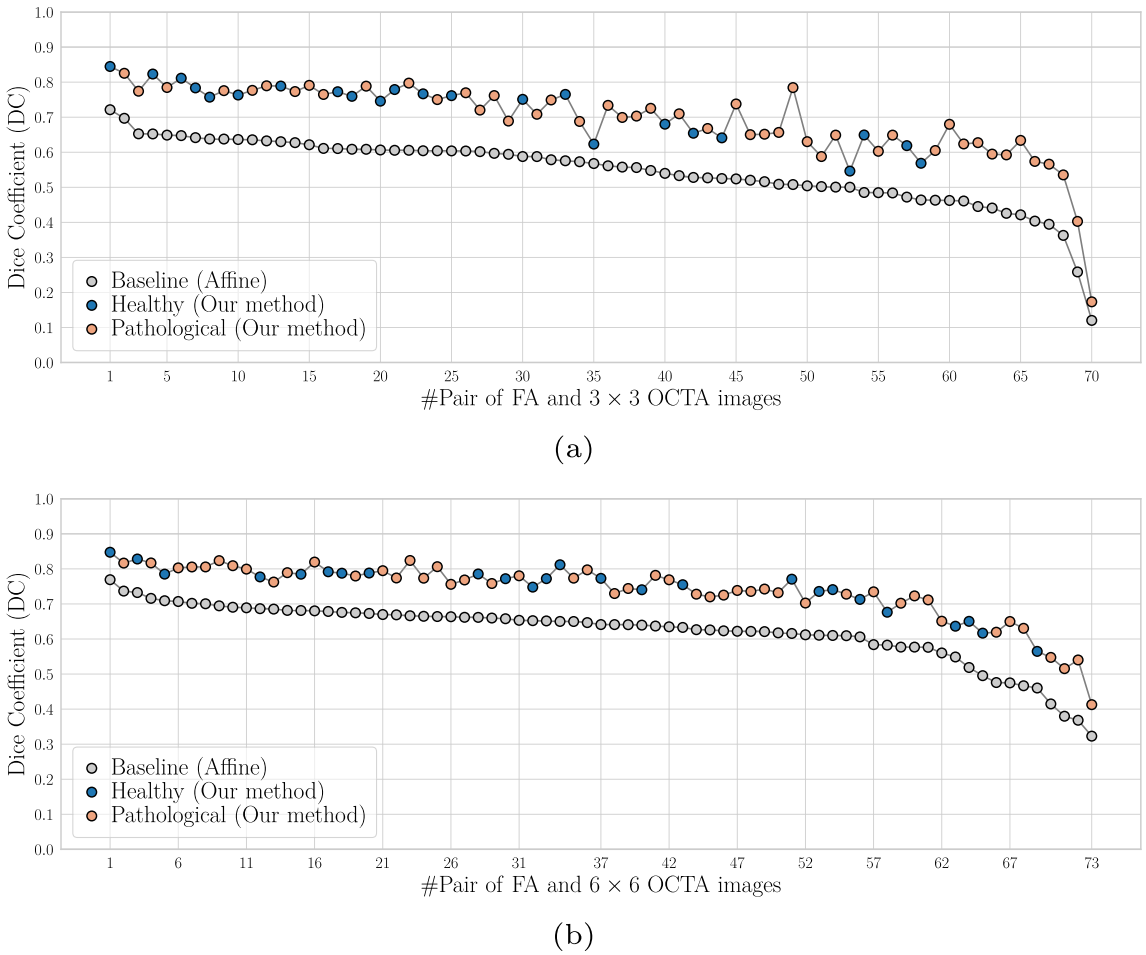

Fig. 6 Comparison of registration results using the affine and our deformable transformation in each pair of images without artifacts. The degree of alignment is expressed in values of the Dice coefficient: **a** pairs of FA and 3×3 OCTA images and **b** pairs of FA and 6×6 OCTA images. The information about healthy or pathological eyes is also shown (color figure online)

The worst scenario corresponds to the region in which $\gamma \geq 0.75$ and $\lambda \leq 0.001$. In this case, the low weight of the regularization term produces a transformation with many points where $|\mathbf{J}_{\psi}|_{\leq 0}$, that is, a transformation that is neither diffeomorphic nor smooth. Additionally, although the DC values in this region are relatively high (anticipating good vessel alignment), the presence of a noisy background in the transformed OCTA image is predicted by the low ZNCC values.

Considering the previous discussion, we can conclude that the best configuration of our registration method is the one corresponding to $\gamma = 1$ and $\lambda = 1$, where a good compromise exists between common information alignment and background information preservation in the transformed OCTA image, also guaranteeing a diffeomorphic transformation. Therefore, from here on, this will be the configuration used in the remaining experiments.

4.3.2 Using OCTA images without artefacts

In a second experiment, we select the best hyperparameter configuration, train the model with the entire augmented dataset used in the previous experiment (no cross-validation is used here), and finally, the learned model is evaluated using the set of original image pairs without artifacts. Note that, according to the data augmentation process described in Sect. 4.1, this last set of image pairs is not explicitly seen by the CNN during the training stage. Here, the idea is to use the DC value as a reference to compare the registrations obtained with our method with the affine registrations for each pair of images. Figure 6 shows the results of this comparison. We can see that, in all the cases, our deformable registration improves all the affine registration results. Concretely, 97.1% (68/70) and 98.6% (72/73) of cases containing 3×3 and 6×6 OCTA images, respectively, have a $DC \geq 0.5$. Contrarily, these percentages drop to 75.7% (53/70) and 87.7% (64/73) for the 3×3 and 6×6 affine registration results, respectively. The

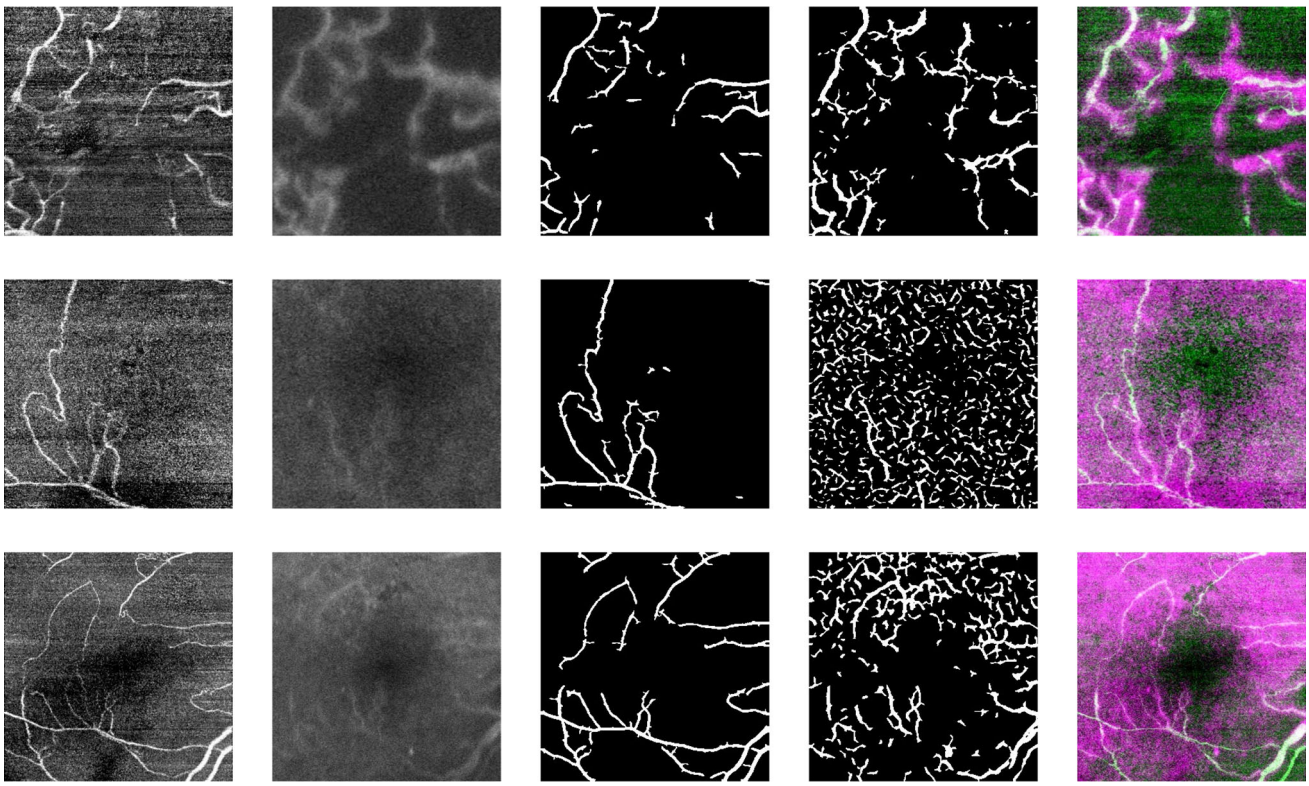

Fig. 7 Registration results obtained by our method in three extreme scenarios. The columns, from left to right, contain the pair of OCTA and FA images, their respective pair of vessel masks, and the final registration. The first and second rows correspond to the pairs of images with the worst and second worst DC values in Fig. 6a, respectively, and, the last row, to the pair of images with the worst DC value in Fig. 6b. The magenta and green colors correspond to FA and OCTA image information, respectively, and the white color appears when common information overlaps (color figure online)

means and standard deviations of the DC values, considering the 3×3 and 6×6 zoom levels, were 0.69 ± 0.11 and 0.74 ± 0.08, respectively, which shows that the behavior of our method is slightly better for the 6×6 zoom level. Likewise, for the baseline case, these figures were 0.55 ± 0.10 and 0.62 ± 0.09, respectively, showing that our method obtains an average increase of 0.14 and 0.12, respectively. That is, the improvement is slightly higher for the 3×3 zoom level.

The three pairs of images with DC values inferior to 0.5 are associated with pathological cases. However, they did not obtain a bad registration, as shown in Fig. 7. The problem in these three extreme scenarios lies in the presence of a lot of noise and the low number of vessels that appear in each pair of images, especially in the FA images, where the vessels are barely visible or blurred. In this context, the number of landmarks inferred by the network will be reduced and, in addition, some of them will be noisy, thus hindering the registration process. Despite obtaining DC values below 0.5 in all three cases, the registration results are acceptable and, in any case, each of them has a DC value which is higher than the one measured for the affine case (see Fig. 6a and 6b). Finally, Figs. 8 and 9 show some registration examples in less extreme scenarios, visually comparing the output obtained by our method and the baseline registration (affine case). Specifically, Fig. 8 highlights the visualization of the vessel overlap degree and Fig. 9 the vessel continuity degree.

4.3.3 Using OCTA images with artefacts

In a last experiment, we apply the subset of image pairs containing OCTA images with artifacts (see Fig. 1) to the model learned in the second experiment. Note that this subset neither took part in the CNN training nor in the data augmentation process. Although, before the registration process, it would be desirable to have a preprocessing stage to eliminate this kind of artifacts, we directly test our registration method in this hard scenario. Figure 10 shows a comparison between the degree of alignment obtained with our method and the affine registration for each pair of images. We can see that, in all the cases, our deformable registration improves the results obtained by the rigid transformation. Specifically, 57.1% (8/14) and 92.3% (12/13) of the pairs containing 3×3 and 6×6 OCTA images, respectively, have a DC ≥ 0.5 with our registration method.

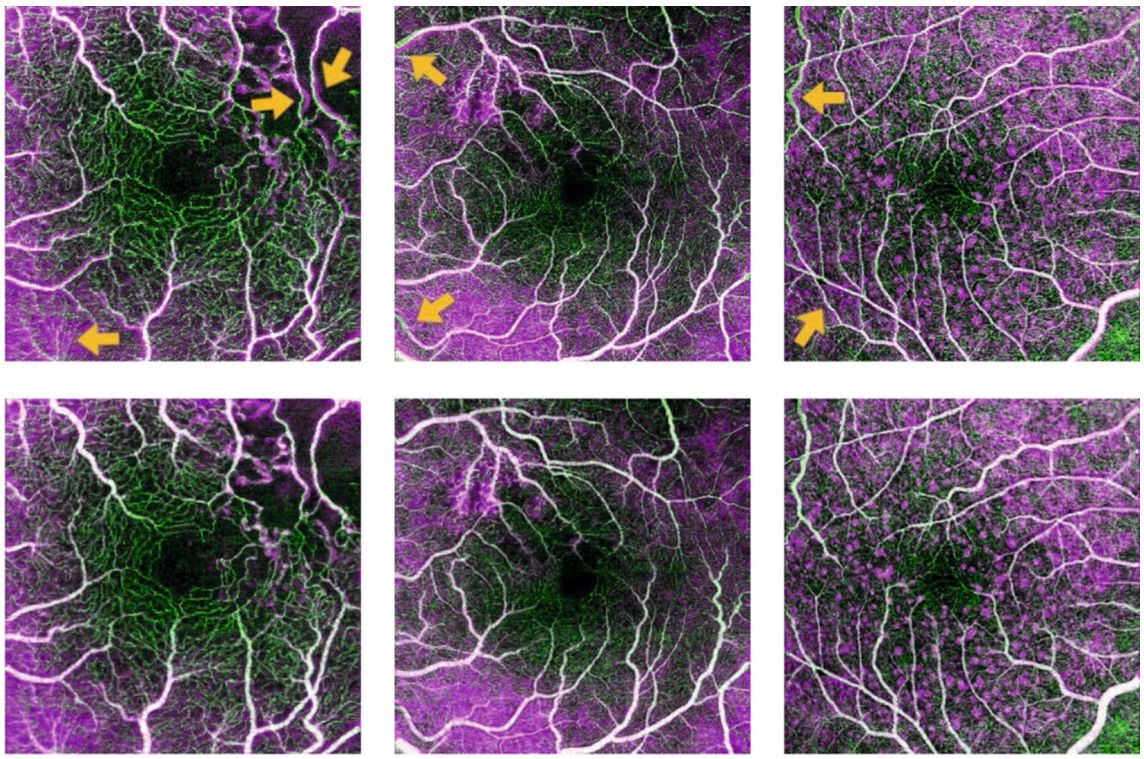

Fig. 8 Registration examples in pairs of images without artifacts using a visualization based on colors. First and second rows show the vessel overlap degree in the affine and our deformable registration, respectively. The magenta and green colors correspond to FA and OCTA image information, respectively, and the white color appears when common information overlaps. The yellow arrows indicate some (but not all) areas of the affine registration where the degree of alignment is improved in its respective deformable registration (color figure online)

Nevertheless, with the affine transformation, only 7.1% (1/14) and 15.4% (2/13) of the pairs reach the mentioned threshold when 3×3 and 6×6 OCTA images are involved, respectively. In addition, Fig. 11 shows two registration examples comparing both types of transformations. It is easy to observe how our deformable deformation outperforms the affine transformation. This improvement also includes a correction of the discontinuity of all those vessels that were "broken" due to the presence of artifacts.

4.3.4 Comparison with other methods

Finally, Table 3 summarizes the comparison of our method with the unsupervised VoxelMorph version and other classical deformable registration approaches. In relation to the latter, we have tried two methods denominated *B-spline* and *diffeomorphic demons* [38] from the popular and well-known SimpleITK registration framework [39]. The former is a deformable registration technique based on deformable B-splines [40], and the latter is a non-parametric diffeomorphic image registration algorithm based on the Thirion's demons algorithm [41]. Specifically, Table 3 shows the average values and respective standard deviation obtained for the DC, ZNCC, number of pixels where $|\mathbf{J}_\psi|_{\leq 0}$, and registration time per pair of images, using the subset of image pairs without artifacts.

In order to establish whether the differences between our method and each of the methods shown in Table 3 are statistically significant, we performed two different hypothesis tests: the parametric z-test and the nonparametric *Wilcoxon Rank Sum* (WRS) test. The former assumes normality in the pair of distributions being compared, while the latter does not. The p-values obtained with both tests, when comparing our method with each of those shown in Table 3, were always less than 0.05 (for both DC and ZNCC), indicating that the null hypothesis (*means are equal* for z-test, and *medians are equal* for WRS-test) can be rejected with a significance level of 5%. Additionally, the notched box plots shown in Fig. 12 allow us to verify how the width of the notch around the median and associated with our method does not overlap with any of the notches of the other methods, providing evidence of a statistically significant difference between the medians [42] and supporting the results obtained with the WRS-test. Therefore, we can conclude that the differences shown in favor of our method have a high statistical significance.

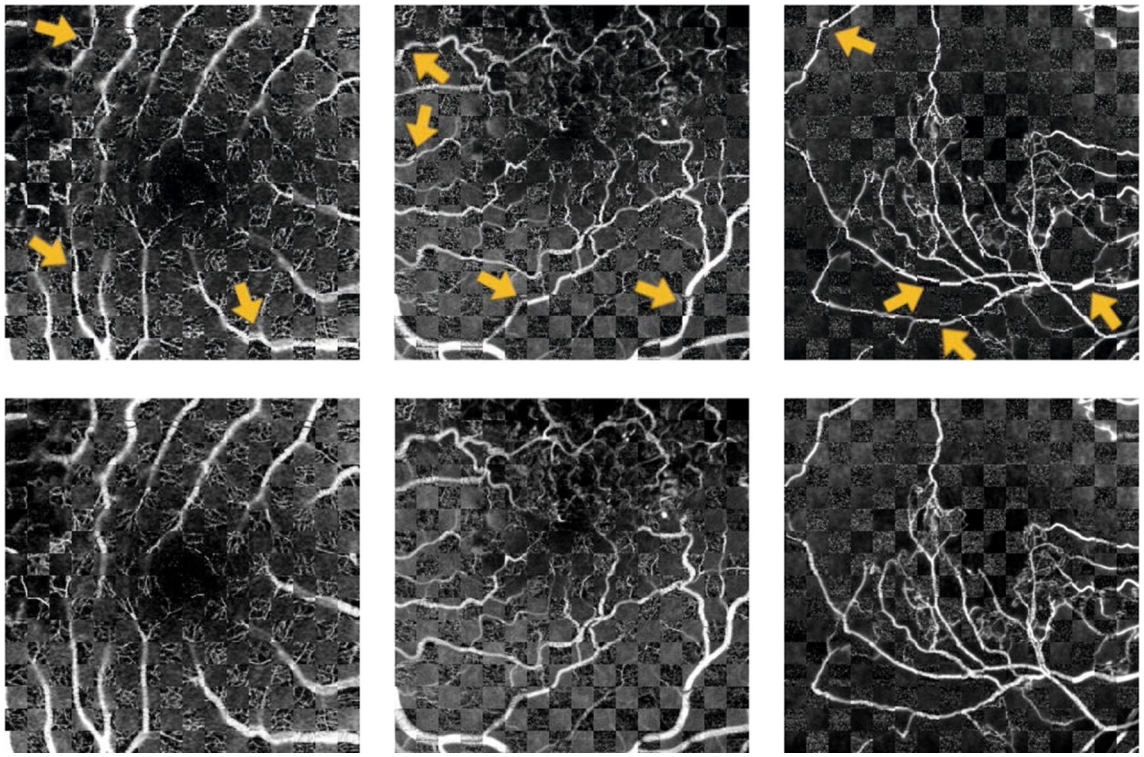

Fig. 9 Registration examples in pairs of images without artifacts using a checkerboard visualization. We can compare the vessel continuity degree obtained with the affine (first row) and our deformable (second row) transformation. The yellow arrows indicate some (but not all) areas of the affine registration where the degree of alignment is improved in their respective deformable registrations (color figure online)

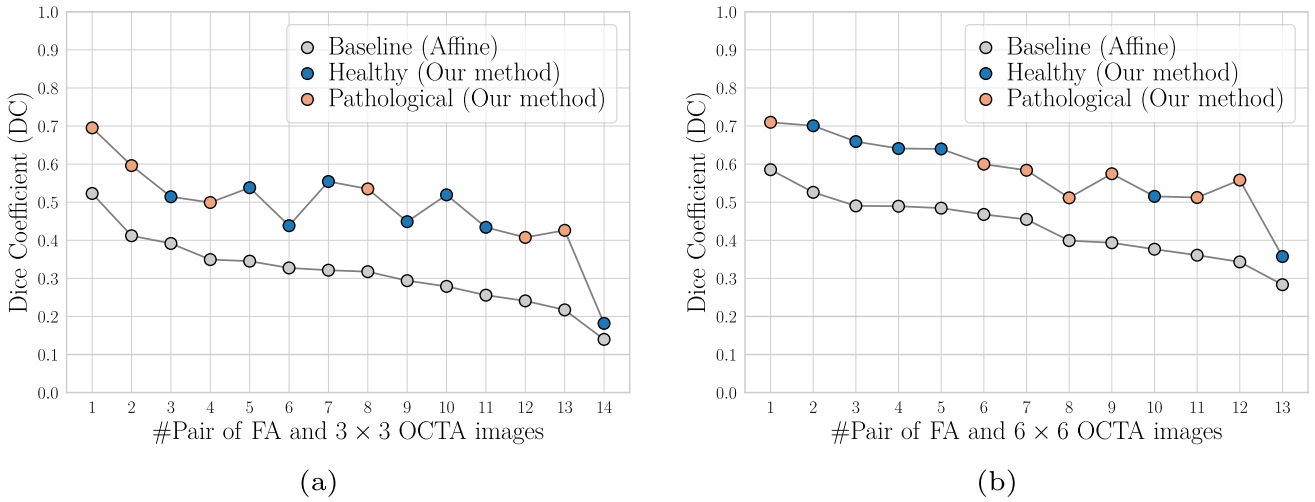

Fig. 10 Comparison of registration results using the affine and our deformable transformation in each pair of images with artifacts. The degree of alignment is expressed in values of the Dice coefficient: **a** pairs of FA and 3 × 3 OCTA images and **b** pairs of FA and 6 × 6 OCTA images. The information about healthy or pathological eyes is also shown (color figure online)

5 Conclusions

In this work, we have proposed a deformable registration methodology that is applied to register FA and superficial plexus OCTA images. Our architecture is strongly inspired by VoxelMorph, a state-of-the-art unsupervised deep learning framework for deformable registration of unimodal images. However, unlike VoxelMorph, our methodology is oriented to multimodal registration, and it is based on weakly supervised deep learning. In comparison with the unsupervised version, we provide evidence on how the use of common information in the CNN training

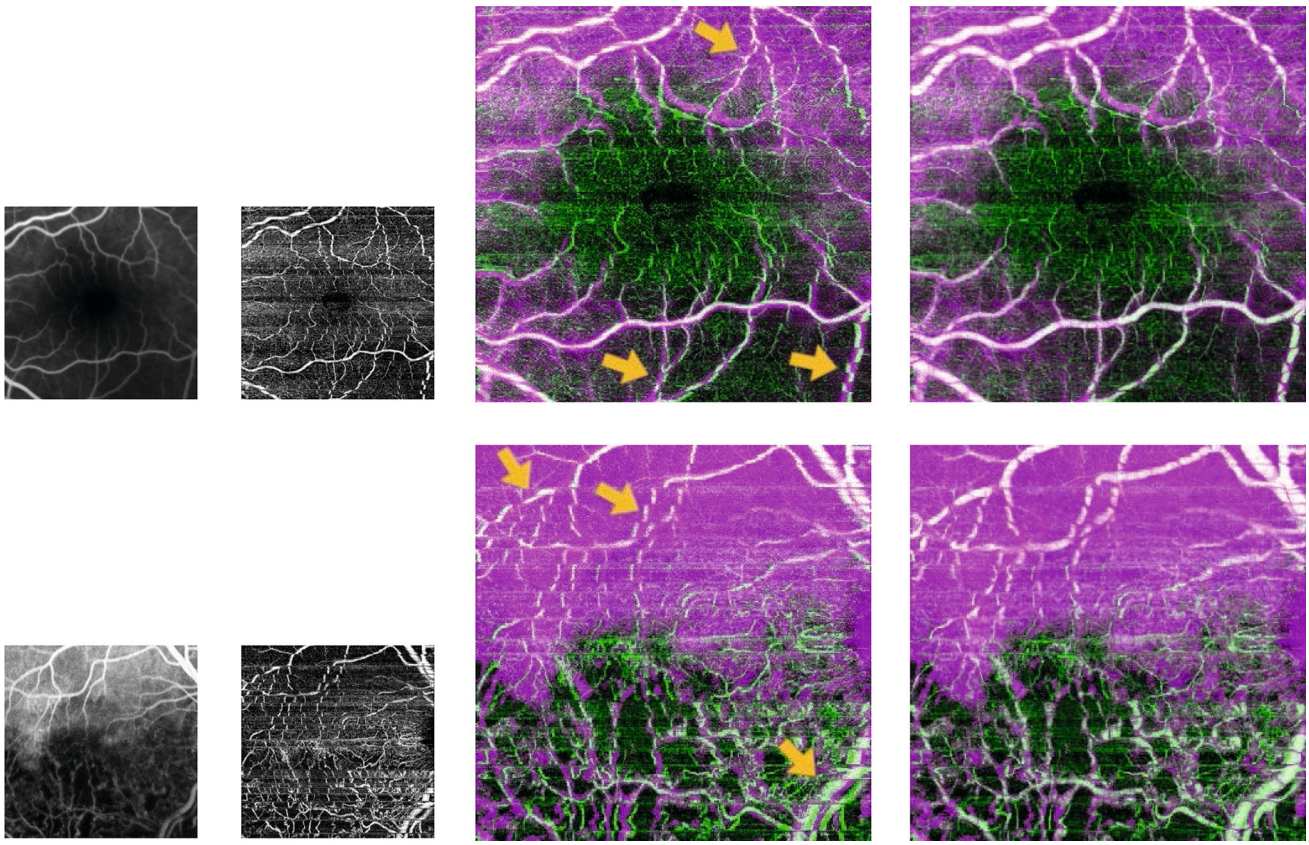

Fig. 11 Registration examples with pairs of images with artifacts using a visualization based on colors. In each row, the first and second columns show the pairs of FA and OCTA input images, respectively; the third and fourth columns display the affine and deformable registration results, respectively. The magenta and green colors correspond to FA and OCTA image information, respectively, and the white color appears when common information overlaps. In order to expand the alignment details, the scale in the last two columns is double the size of the one used in the first two. The yellow arrows indicate some (but not all) areas of the affine registration where the degree of alignment is improved in its respective deformable registration (color figure online)

Table 3 Comparison of the performance of our method with other classical deformable registration methods. The evaluation metrics used are the average values and respective standard deviations obtained for the Dice coefficient, DC(FA_{seg},$\psi(OCTA_{seg})$), zero-normalized cross-correlation, ZNCC(OCTA,ψ(OCTA)), number of pixels (percentage in parentheses) whose Jacobian determinant is non-positive ($|\mathbf{J}_\psi|_{\leq 0}$), and time taken for registering a pair of images. The boldfaced label denotes the method with the best results

Method	DC	ZNCC	$\lvert J_\psi\rvert_{\leq 0}$ (%)	Time (s)
Baseline (Affine)	0.5842 ± 0.0999	–	–	–
B-spline	0.6136 ± 0.1031	0.7631 ± 0.1150	4.58 ± 54.66 (0.00)	4.08 ± 2.79
Diff-Demons	0.6568 ± 0.0986	0.7872 ± 0.0353	40.93 ± 79.46 (0.04)	1.03 ± 0.04
VoxelMorph[a]	0.6726 ± 0.0942	0.7974 ± 0.0472	0.00 ± 0.00 (0.00)	0.12 ± 0.00
Our method	0.7166 ± 0.0964	0.8211 ± 0.0403	0.00 ± 0.00 (0.00)	0.12 ± 0.00

[a]Unsupervised VoxelMorph version

process is key to improve the alignment. Specifically, this common information corresponds to an approximation to the segmentation of the main vessel network, which might include noise or not contain all vessel pixels. Note that, in daily clinical practice, it is difficult, if not impossible, to obtain accurate ground truth, hence the importance of using a weakly supervised approach. We also conclude that the regularization mechanism has an important role in avoiding the alteration of the background information (non-common information) of the transformed image.

The experiments carried out have been evaluated in a set of 143 pairs of OCTA and FA images, including healthy

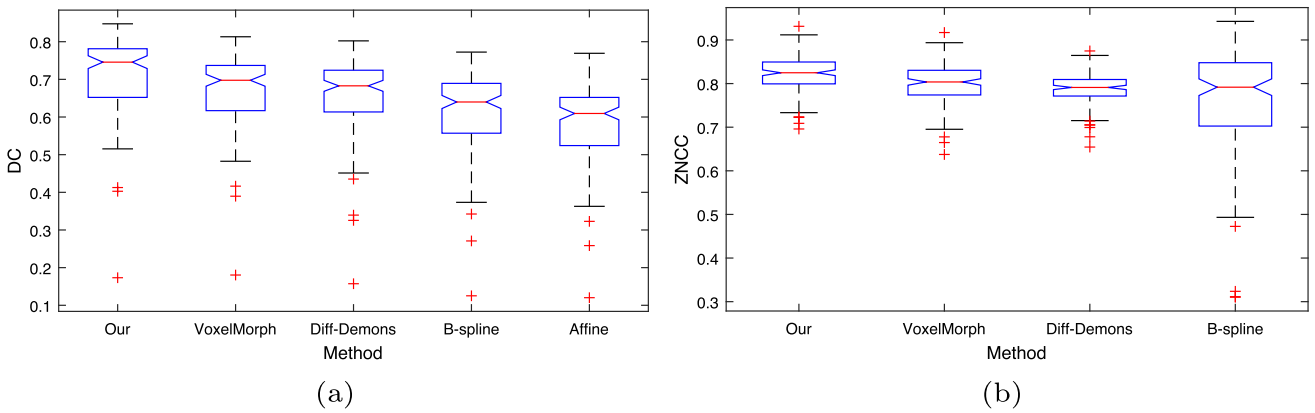

Fig. 12 Notched box plots showing the distributions of values around their medians for two of the metrics evaluated in the subset of image pairs without artifacts and considering different methods used to compare our proposal: **a** DC; **b** ZNCC

and pathological patients. In all cases, our registration method outperformed the unsupervised VoxelMorph version and other classical methods, both rigid (affine) and deformable, considering different evaluation metrics: degree of alignment of the common information, invariance of the non-common information after applying the transformation, smoothness of the deformation, and computational cost. The learned model was also used in a more severe scenario where the OCTA images, involved in the 27 pairs of images to register, present artifacts. This is a type of noise that appears frequently in OCTA images obtained in daily clinical practice. Here, the results obtained with our method continue to improve those obtained by the affine transformation, and it is even observed that, in the transformed OCTA image, the vessel discontinuity, produced by the presence of the artifacts, tends to be corrected. This behavior adds evidence of the robustness of our method.

In future lines of work, our method could be applied to pairs of images belonging to other retinal imaging techniques, such as those already mentioned in the introduction and different from the FA and OCTA modalities. As another option, new CNN architectures could be studied and tested in our approach, with the aim of further improving registration or facilitating the initial required alignment of regions of interest that are highly misaligned. For example, when faced with the problem of registering pairs of OCTA and FA images, the initial approximate registration required in our method was not trivial: the area covered by the FA image is significantly larger than the one corresponding to the OCTA image, and so we do not start from an ideal situation in which the pair of images is quasi-aligned.

Acknowledgements This work was supported by the Ministerio de Ciencia, Innovación y Universidades, Government of Spain, through the RTI2018-095894-B-I00 research project. Some of the authors of this work also receive financial support from the European Social Fund through the predoctoral contract ref. PEJD-2019-PRE/TIC-17030 and research assistant contract ref. PEJ-2019-AI/TIC-13771.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Availability of data and material The dataset used in this article is available on http://www.varpa.org/research/ophtalmology.html.

Declaration

Conflict of interest There is no conflict of interests regarding the publication of this paper.

References

1. Panwar N, Huang P, Lee J, Keane PA, Chuan TS, Richhariya A, Teoh S, Lim TH, Agrawal R (2016) Fundus photography in the 21st century: a review of recent technological advances and their implications for worldwide healthcare. Telemed e-Health 22(3):198–208
2. Reshef ER, Miller JB, Vavvas DG (2020) Hyperspectral imaging of the retina: a review. Int Ophthalmol Clin 60(1):85–96
3. Sparrow JR, Duncker T, Schuerch K, Paavo M, de Carvalho JRL (2020) Lessons learned from quantitative fundus autofluorescence. Prog Retinal Eye Res 74:100774
4. Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, Hee MR, Flotte T, Gregory K, Puliafito CA et al (1991) Optical coherence tomography. Science 254(5035):1178–1181

5. Olson JL, Mandava N (2006) Fluorescein angiography. In: Huang D, Kaiser PK, Lowder CY, Traboulsi EI (eds) Retinal Imaging. Mosby, Philadelphia, pp 3–21
6. Owens SL (1996) Indocyanine green angiography. Br J Ophthalmol 80(3):263–266
7. Kashani AH, Chen C-L, Gahm JK, Zheng F, Richter GM, Rosenfeld PJ, Shi Y, Wang RK (2017) Optical coherence tomography angiography: a comprehensive review of current methods and clinical applications. Prog Retinal Eye Res 60:66–100
8. de Carlo TE, Romano A, Waheed NK, Duker JS (2015) A review of optical coherence tomography angiography (OCTA). Int Journal of Retina Vitreous 1:1–15
9. Schwartz DM, Fingler J, Kim DY, Zawadzki RJ, Morse LS, Park SS, Fraser SE, Werner JS (2014) Phase-variance optical coherence tomography: a technique for noninvasive angiography. Ophthalmology 121(1):180–187
10. Matsunaga D, Yi J, Puliafito CA, Kashani AH (2014) OCT angiography in healthy human subjects. Ophthal Surg Lasers Imag Retina 45(6):510–515
11. Boveiri HR, Khayami R, Javidan R, Mehdizadeh A (2020) Medical image registration using deep neural networks: a comprehensive review. Comput Electrical Eng 87:106767
12. 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
13. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV (2019) VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans Med Imag 38(8):1788–1800
14. Hu Y, Modat M, Gibson E, Li W, Ghavami N, Bonmati E, Wang G, Bandula S, Moore CM, Emberton M et al (2018) Weakly-supervised convolutional neural networks for multimodal image registration. Med Image Anal 49:1–13
15. Hering A, Kuckertz S, Heldmann S, Heinrich M (2019) Memory-efficient 2.5D convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans. Int J Comput Assist Radiol Surg 14(11):1901–1912
16. Blendowski M, Bouteldja N, Heinrich MP (2020) Multimodal 3D medical image registration guided by shape encoder-decoder networks. International journal of computer assisted radiology and surgery 15(2):269–276
17. Arikan M, Sadeghipour A, Gerendas B, Told R, Schmidt-Erfurt U (2019) Deep learning based multi-modal registration for retinal imaging. In: Suzuki K, Reyes M, Syeda-Mahmood T, Konukoglu E, Glocker B, Wiest R, Gur Y, Greenspan H, Madabhushi A (eds) Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer, Cham, pp 75–82
18. Lee J, Liu P, Cheng J, Fu H (2019) A deep step pattern representation for multimodal retinal image registration. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 5076–5085
19. Tian Y, Hu Y, Ma Y, Hao H, Mou L, Yang J, Zhao Y, Liu J (2020) Multi-scale U-net with edge guidance for multimodal retinal image deformable registration. In: 42nd Annual international conference of the IEEE engineering in medicine biology society, pp 1360–1363
20. Silva TD, Chew EY, Hotaling N, Cukras CA (2021) Deep-learning based multi-modal retinal image registration for the longitudinal analysis of patients with age-related macular degeneration. Biomed Opt Exp 12(1):619–636
21. Wang Y, Zhang J, Cavichini M, Bartsch D-UG, Freeman WR, Nguyen TQ, An C (2021) Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework. IEEE Trans Image Process 30:3167–3178
22. Jiang Y, Zheng Y, Sui X, Jiao W, He Y, Jia W (2021) ASRNet: adversarial segmentation and registration networks for multi-spectral fundus images. Comput Syst Sci Eng 36(3):537–549
23. Zhang J, Wang Y, Dai J, Cavichini M, Bartsch D-UG, Freeman WR, Nguyen TQ, An C (2022) Two-step registration on multi-modal retinal images via deep neural networks. IEEE Trans Image Process 31:823–838
24. Jia Y, Bailey ST, Wilson DJ, Tan O, Klein ML, Flaxel CJ, Potsaid B, Liu JJ, Lu CD, Kraus MF, Fujimoto JG, Huang D (2014) Quantitative optical coherence tomography angiography of choroidal neovascularization in age-related macular degeneration. Ophthalmology 121(7):1435–1444
25. Teussink MM, Breukink MB, van Grinsven MJ, Hoyng CB, Klevering BJ, Boon CJ, de Jong EK, Theelen T (2015) OCT angiography compared to fluorescein and indocyanine green angiography in chronic central serous chorioretinopathy. Investigat Ophthalmol Visual Sci 56(9):5229–5237
26. Peres M, Kato R, Kniggendorf V, Cole E, Onal S, Torres E, Louzada R, Belfort R, Duker J, Novais E, Regatieri C (2016) Comparison of optical coherence tomography angiography and fluorescein angiography for the identification of retinal vascular changes in eyes with diabetic macular edema. Ophthal Surg Lasers Imag Retina 47:1013–1019
27. Stattin M, Haas A-M, Ahmed D, Stolba U, Graf A, Krepler K, Ansari-Shahrezaei S (2020) Detection rate of diabetic macular microaneurysms comparing dye-based angiography and optical coherence tomography angiography. Sci Rep 10:1–8
28. Told R, Reiter GS, Orsolya A, Mittermüller TJ, Eibenberger K, Schlanitz FG, Arikan M, Pollreisz A, Sacu S, Schmidt-Erfurth U (2020) Swept source optical coherence tomography angiography, fluorescein angiography, and indocyanine green angiography comparisons revisited: Using a novel deep-learning-assisted approach for image registration. Retina 40:2010–2017
29. Martínez-Río J, Carmona EJ, Cancelas D, Novo J, Ortega M (2021) Robust multimodal registration of fluorescein angiography and optical coherence tomography angiography images using evolutionary algorithms. Comput Biol Med 134:104529
30. University of A Coruña: FOCTAIR: Fluorescein and Optical Coherence Tomography Angiography Image Registration dataset. http://www.varpa.org/research/ophtalmology.html. [last access 2022/05/20] (2022)
31. Zang P, Liu G, Zhang M, Dongye C, Wang J, Pechauer AD, Hwang TS, Wilson DJ, Huang D, Li D, Jia Y (2016) Automated motion correction using parallel-strip registration for wide-field en face OCT angiogram. Biomedical Optics Express 7(7):2823–2836
32. Hoopes A, Hoffmann M, Fischl B, Guttag J, Dalca AV (2021) Hypermorph: Amortized hyperparameter learning for image registration. In: International conference on information processing in medical imaging, pp 3–17 . Springer
33. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, pp 234–241 . Springer
34. Arsigny V, Commowick O, Pennec X, Ayache N (2006) A log-Euclidean framework for statistics on diffeomorphisms. In: Larsen R, Nielsen M, Sporring J (eds) Medical image computing and computer-assisted intervention—MICCAI 2006. Springer, Berlin, pp 924–931
35. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. arXiv preprint arXiv:1506.02025
36. Sun S, Park HW, Haynor DR, Kim Y (2003) Fast template matching using correlation-based adaptive predictive search. Int J Imag Syst Technol 13:169–178

37. Dalca AV, Hoopes A, Hoffmann M, Fischl B (2022) VoxelMorph: Learning-based image registration. https://github.com/voxelmorph/voxelmorph. [last access 2022/01/15]
38. Vercauteren T, Pennec X, Perchant A, Ayache N (2009) Diffeomorphic demons: efficient non-parametric image registration. NeuroImage 45(1, Supplement 1):61–72
39. Lowekamp B, Chen D, Ibáñez L, Blezek D (2013) The design of simpleITK. Front Neuroinf 7:1–14
40. Rueckert D, Sonoda LI, Hayes C, Hill DLG, Leach MO, Hawkes DJ (1999) Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imag 18(8):712–721
41. Thirion J-P (1998) Image matching as a diffusion process: an analogy with Maxwell's demons. Med Image Anal 2(3):243–260
42. McGill R, Tukey JW, Larsen WA (1978) Variations of box plots. Am Stat 32(1):12–16