Applied Soft Computing Journal 91 (2020) 106210

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction

Álvaro S. Hervella *, José Rouco, Jorge Novo, Marcos Ortega

CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña, A Coruña, Spain
Department of Computer Science, Universidade da Coruña, A Coruña, Spain

ARTICLE INFO

Article history:
Received 10 September 2019
Received in revised form 21 January 2020
Accepted 3 March 2020
Available online 13 March 2020

Keywords:
Deep learning
Eye fundus
Self-supervised learning
Optic disc
Blood vessels
Fovea
Medical imaging
Transfer learning

ABSTRACT

Deep learning is becoming the reference paradigm for approaching many computer vision problems. Nevertheless, the training of deep neural networks typically requires a significantly large amount of annotated data, which is not always available. A proven approach to alleviate the scarcity of annotated data is transfer learning. However, in practice, the use of this technique typically relies on the availability of additional annotations, either from the same or natural domain. We propose a novel alternative that allows to apply transfer learning from unlabelled data of the same domain, which consists in the use of a multimodal reconstruction task. A neural network trained to generate one image modality from another must learn relevant patterns from the images to successfully solve the task. These learned patterns can then be used to solve additional tasks in the same domain, reducing the necessity of a large amount of annotated data.

In this work, we apply the described idea to the localization and segmentation of the most important anatomical structures of the eye fundus in retinography. The objective is to reduce the amount of annotated data that is required to solve the different tasks using deep neural networks. For that purpose, a neural network is pre-trained using the self-supervised multimodal reconstruction of fluorescein angiography from retinography. Then, the network is fine-tuned on the different target tasks performed on the retinography. The obtained results demonstrate that the proposed self-supervised transfer learning strategy leads to state-of-the-art performance in all the studied tasks with a significant reduction of the required annotations.

1. Introduction

The analysis of the anatomical structures in the retina represents an essential step for the diagnosis and screening of important ocular and systemic diseases. The morphology of the anatomical structures, such as blood vessels, fovea, or optic disc, can in itself provide evidence of the presence of certain diseases. Additionally, they can be used as reference for the localization of lesions as well as for the assessment of their severity [1].

The retinal anatomy can be studied using eye fundus photography, or retinography, which is a non-invasive and affordable imaging technique. These reasons motivate its widespread use in many clinical services, and make it an interesting target for the development of image analysis algorithms [2]. In this regard, several works have approached the automatic analysis of eye fundus images, including the localization or segmentation of the different anatomical structures [1]. Similarly to other medical fields, the number of methods based on neural networks has grown significantly in the last few years, which carried an improvement of the obtained results [3–5]. Currently, the use of deep neural networks (DNNs) is the standard approach in many computer vision applications when the required annotated data is available. DNNs have not only improved the results obtained with traditional methods, but have also brought a new simplified paradigm where no feature design is needed [6]. Instead, the focus has shifted to the design or selection of the most suitable network architectures, training losses and training strategies [7].

Regarding the automatic analysis of representative anatomical structures in retinography, the main limitation for the early use of DNNs was the scarcity of annotated data [3]. In that sense, the available datasets typically present a small number of annotated samples due to the difficulty of hand-labelling the retinal images in detail. Moreover, despite that some large datasets have been gathered, in practice, the annotated data usually present a meagre representation of pathological cases [8], given that those images are typically of higher variability and complexity.

The scarcity of annotated data is not specific to retinal imaging. Instead, this is a broadly relevant issue in medical imaging,

* Corresponding author at: CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña, A Coruña, Spain.
E-mail address: a.suarezh@udc.es (Á.S. Hervella).

https://doi.org/10.1016/j.asoc.2020.106210

where a high level of expertise is required for the reliable labelling of the medical data [9]. Conversely, a large amount of medical images is produced everyday in the different medical services due to the widespread use of imaging techniques in modern clinical practice [7]. This directly produces the availability of large unlabelled datasets, which may be used for the training of neural networks in unsupervised or semisupervised settings. Moreover, the medical images are typically accompanied by clinical reports describing the patient's conditions, which may be used for distilling image-level labels [7]. In contrast, pixel-level labels require to be annotated on purpose by, at least, one clinical expert. Moreover, the manual annotation of pixel-level labels represents a difficult task, being more tedious and time-consuming than the manual annotation of image-level labels. This is reflected in the number of annotated samples that are provided in common medical imaging datasets [1], being significantly smaller when the required annotations are more detailed [10].

The limited annotated data in medical imaging is typically alleviated using extensive data augmentation and transfer learning [7]. The use of data augmentation techniques including, e.g., rigid transformations, elastic deformations or colour transformations, has become a key component of successful deep learning methods [7]. Transfer learning, on the other hand, has been applied since the earliest deep learning approaches on medical imaging. Early works used the first layers of pre-trained classification networks as feature extractors [11]. These networks are trained in a broad domain application with extensive available data, such as ImageNet classification [12]. Posterior works additionally performed fine-tuning of the pre-trained layers together with additional layers that are specialized for the target task [3]. Multi-task learning techniques have also been recently explored for their ability to combine complementary tasks over data of the same domain [13]. Multi-task settings can be seen as a special case of transfer learning where the transference of knowledge is bidirectional and simultaneous between the involved tasks. In this case, the amount of labelled data is increased by using heterogeneous labels (for each task) over data of the same domain. However, these additional heterogeneous labels from the same domain can be also exploited using a regular pre-training and fine-tuning approach, to achieve improved results on the later tasks [14]. In the case of training multiple tasks of this kind, with varying difficulty, the training order may have an impact in the final performance. In this sense, some works have also proposed to optimize the sampling order or the different tasks to improve the final outcome [15].

Self-supervised methods are a recent alternative that allows the use of unlabelled data for transfer learning [16]. These approaches rely on the use of innovative complementary tasks which labels can be automatically computed from the unlabelled datasets and, thus, can be trained without the need of manual annotations. The purpose of training these self-supervised tasks is to learn relevant patterns of the domain from the data, and then use the learned patterns to improve the desired tasks through transfer of multitask learning. Existent proposals in medical imaging have exploited, as reference, the colour information in images using a colourization task [17] or the relation among longitudinal data by learning patient embeddings [18].

A rich source of information that has still not been exploited for self-supervised transfer learning is the unlabelled multimodal data in medical imaging. In modern clinical practice, it is common to analyse and diagnose the patients using multiple imaging techniques. This results in the availability of multimodal sets in which samples from complementary image modalities are available for the same patient. The availability of these multimodal data can be exploited using a self-supervised multimodal reconstruction task where a neural network is trained to generate one image modality from other. If the two involved modalities are different enough, the network has necessarily to learn the recognition of relevant domain-related patterns to successfully solve the task. Then, the learned models can be further adjusted to solve additional target tasks over the same input modality.

In particular, in this work, we experiment with these ideas in the context of the localization and segmentation of anatomical structures of the eye fundus in retinography. The objective is to reduce the amount of annotated data that is required to solve these tasks with a DNN, and to that end we propose to use the self-supervised multimodal reconstruction for transfer learning. Specifically, we pre-train the networks to generate fluorescein angiography from retinography. The retinography and angiography are complementary image modalities, both providing visualizations of the eye fundus. However, the angiography is an invasive modality that requires the injection of a contrast dye to the patients, providing additional information about the retinal vasculature and related lesions. In the proposed paradigm, both unlabelled image modalities are used to pre-train the networks. However, the target tasks are performed using a single image modality, which in this case is the retinography. Moreover, the unlabelled multimodal data for pre-training and the task-specific data for fine-tuning do not need to belong to the same patients. This allows the use of any multimodal dataset available in the same domain, independently of the target tasks.

With regards to the multimodal reconstruction, Hervella et al. [19] demonstrated that a pseudo-angiography representation can be generated from a given retinography using a DNN. Moreover, the vascular enhancement in the angiography can also be directly exploited to produce an approximate representation of the vascular tree in retinography [20], requiring an additional pre-processing of the target angiographies. None of the previous works, however, have taken advantage of the domain-specific patterns that a neural network must learn in order to perform the multimodal reconstruction. The idea proposed in this work exploits those patterns learned from the unlabelled multimodal data for transfer learning purposes. This represents a novel alternative to complement the training of a DNN and reduce the amount of annotated data that is required. As reference, an illustrative example of retinography, fluorescein angiography, and generated pseudo-angiography for the same eye is depicted in Fig. 1.

In order to demonstrate the advantages of the proposed self-supervised transfer learning strategy, we use the multimodal reconstruction as a common self-supervised pre-training for: (1) the localization of the fovea, (2) the localization and (3) segmentation of the optic disc, and (4) the segmentation of the retinal vasculature. Additionally, we aim at solving all these target tasks with the same standard methodology, including the network architecture and training strategy. In order to study the efficient use of annotated data with our proposal, we conducted an extensive experimentation with progressive amounts of annotated training data. The objective is to demonstrate that the self-supervised multimodal reconstruction successfully reduces the amount of annotations required to solve the considered target tasks.

1.1. State-of-the-art

In the literature, several works have approached the automatic analysis of the most important anatomical structures in retinography [1]. Previous works typically focus on the localization or segmentation of a single anatomical structure. However, the localization of the fovea has been traditionally approached together with the localization of the optic disc. This is motivated by the use of the optic disc location as reference to detect the fovea [21,22]. Additionally, the retinal vascular tree has also been

(a) (b) (c)

Fig. 1. Example of (a) retinography, (b) fluorescein angiography, and (c) pseudo-angiography for the same eye. The pseudo-angiography (c) is generated from (a) using the method proposed in Hervella et al. [19].

used as reference for the localization of both the optic disc and the fovea [22,23].

Regarding the optic disc, some proposals exploit its characteristic circular shape. For instance, edge detection filters can be used to obtain optic disc boundary candidates [23,24]. Then, these boundaries allow to derive both the segmented area and the centre coordinates after a refinement step, e.g., using a hough transformation [23,24] or measuring the distance to some pre-computed templates [25]. In this context, Dashtbozorg et al. [26] proposes specific filters in order to better match the optic disc shape. Alternatively, the characteristic colour patterns of the optic disc are also exploited by applying histogram matching [27]. Additionally, Qureshi et al. [25] explores the use of an ensemble of previously proposed algorithms to improve the results. In contrast, the most recent proposals use deep learning for both the localization [4,5] and the segmentation [3] of the optic disc. In the localization task, a convolutional network with fully-connected output layers can be used to predict the fovea coordinates [4]. However, instead, Meyer et al. [5] reformulates the problem as a heatmap regression task, which can be performed using fully-convolutional networks. The latter approach is the one that we have adopted in this work for the localization of both optic disc and fovea.

With regards to the fovea localization, traditional approaches typically rely on the previous detection of the optic disc to reduce the search area [21,22,28]. Additionally, Gegundez-Arias et al. [22] also makes use of the extracted retinal vascular tree to perform a better initial estimate of the foveal region. The final localization is usually performed exploiting the characteristic shape and colour of the foveal region. For instance, Niemeijer et al. [21] uses a k-NN regressor and features extracted from both the retinal image and the segmented blood vessels, whereas Gegundez-Arias et al. [22] uses thresholding techniques and features from the original image. In addition, the fovea and the optic disc can be detected using template matching with the same template filter but of opposite responses [28]. Similarly to the optic disc, the most recent proposals use DNNs for the regression of the fovea coordinates [4] or the prediction of a full-image size distance map [5].

In the case of the retinal vasculature segmentation, traditional approaches have typically relied on the characteristic tubular shape of blood vessels. This characteristic can be exploited using the gradients of the image or Gabor filter responses, among other techniques [29]. However, recent works have successfully solve this task using DNNs, either fully convolutional [3], fully connected [30] or convolutional with fully-connected output layers [31]. In this regard, the novelty of recent works is related to the use of specific network designs or training objectives, including, as reference, the use of class-balanced losses [3] or the supervision to intermediate layers [32].

The rest of the manuscript is organized as follows: A general overview of the proposed approach, along with a description of the pre-training and target tasks is depicted in Section 2. The network architecture and the training strategy are also detailed in this section. The description of the conducted experiments and the obtained results are presented in . Section 4 is focused on the discussion of results and the final conclusions are drawn in Section 5.

2. Methodology

A general scheme that summarizes the proposed methodology is depicted in Fig. 2. Particularly, the self-supervised reconstruction of fluorescein angiography from retinography is used as pre-training. Then, the pre-trained neural network is fine-tuned on the different target tasks. Given that the multimodal reconstruction covers the whole anatomy of the retina, it is expected that the internal neural network representations that are used for the reconstruction are also useful for the detection and segmentation of the different anatomical structures.

The exact same network architecture and training strategy are employed for all the considered tasks, with the only difference of the loss function. In particular, for the pre-training task a reconstruction loss is used, whereas for the target tasks two different losses are used depending on the objective: a localization loss and a segmentation loss.

2.1. Self-supervised multimodal reconstruction

The multimodal reconstruction of fluorescein angiography from retinography is conceived as a self-supervised task due to the use of aligned retinography–angiography pairs from the same eye [19]. In this scenario, there is a pixel-wise correspondence between the input retinography and the target angiography. This enables the use of full-reference metrics for the reconstruction loss, which provides a supervisory training signal that involves fine image details and does not need any human labelling effort.

The aligned multimodal data for training the network is obtained after the registration of retinographies and angiographies of the same eye. This registration is performed following a domain-specific methodology that relies on the presence of retinal vessels in both image modalities [33]. This registration methodology is divided into two main steps: an initial landmark-based registration that globally aligns the images followed by a refined pixel-wise registration that corrects the remaining small misalignments between the images.

Both retinography and angiography display the eye fundus in a circular Field of View (FOV). After the image alignment, the area containing information from both modalities, denoted as the multimodal FOV, Ω_M, will be typically smaller than the individual FOVs of the original images. This area is defined as:

$$\Omega_M = \Omega_R \cap \Omega_A \tag{1}$$

where Ω_R and Ω_A denote the circular FOVs of the retinography and the angiography respectively. Consequently, Ω_M represents

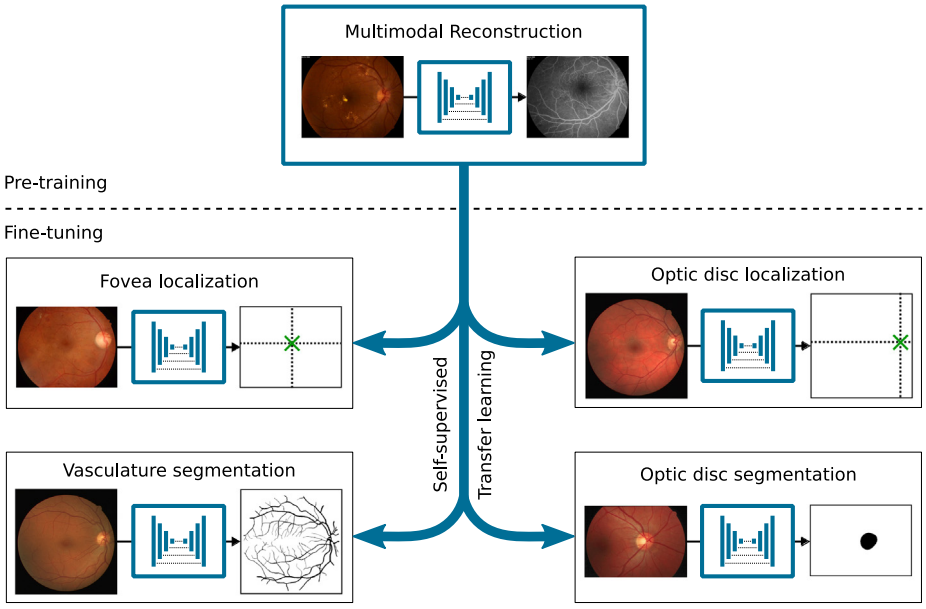


Fig. 2. Scheme of the proposed methodology. The self-supervised multimodal reconstruction of angiography from retinography is used as pre-training task. The pre-trained network is fine-tuned on different target tasks aiming at the analysis of the main anatomical structures in retinography.

the region where the reconstruction loss is computed during the training. The reconstruction loss $\mathcal{L}_R(\mathbf{g}(\mathbf{r}), \mathbf{a})$ is given by:

$$\mathcal{L}_R(\mathbf{g}(\mathbf{r}), \mathbf{a}) = -\sum_{\Omega_M} \mathbf{SSIM}(\mathbf{g}(\mathbf{r}), \mathbf{a}) \tag{2}$$

where $\mathbf{r}$ is the input retinography, $\mathbf{a}$ the target angiography, $\mathbf{g}(\mathbf{r})$ the output of the network, and **SSIM** the Structural Similarity (SSIM) index map between the target angiography and the network output [19]. SSIM is frequently used as test metric for the evaluation of deep learning models that were trained with other losses. However, in our context, the direct optimization of the SSIM has demonstrated an improved performance with respect to other common metrics in the presented task [19]. The **SSIM** map is obtained as:

$$\mathbf{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{y}} + C_1) + (2\boldsymbol{\sigma}_{\mathbf{xy}} + C_2)}{(\boldsymbol{\mu}_{\mathbf{x}}{}^2 + \boldsymbol{\mu}_{\mathbf{y}}{}^2 + C_1)(\boldsymbol{\sigma}_{\mathbf{x}}{}^2 + \boldsymbol{\sigma}_{\mathbf{y}}{}^2 + C_2)} \tag{3}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote two single channel images, $\boldsymbol{\mu}_{\mathbf{x}}$ and $\boldsymbol{\mu}_{\mathbf{y}}$ the local averages of $\mathbf{x}$ and $\mathbf{y}$ respectively, $\boldsymbol{\sigma}_{\mathbf{x}}$ and $\boldsymbol{\sigma}_{\mathbf{y}}$ the local standard deviations of $\mathbf{x}$ and $\mathbf{y}$, respectively, $\boldsymbol{\sigma}_{\mathbf{xy}}$ the local covariance between $\mathbf{x}$ and $\mathbf{y}$, and C_1 and C_2 are constant values used to avoid instability when the denominator terms are close to zero [34]. The local statistics for each pixel are computed using a Gaussian window with $\sigma = 1.5$ [34].

2.2. Localization of anatomical structures of the retina

The localization of the fovea and optic disc centres is obtained following the same task formulation. In this regard, the localization tasks consist in the regression of pixel coordinates, which can be directly approached using a DNN with fully connected layers that produce the coordinate values. However, this kind of regression settings can be difficult to train, and does not take full advantage of the shared weights and local connectivity of convolutional networks. An straightforward alternative is to predict a target map with two classes: the pixel of the target location and the rest of the image. In this case, the difficulty is that the target maps are heavily unbalanced. An alternative to improve this is to augment the ground truth annotations by the means of a distance map to the target pixel [5]. Using the Euclidean norm, this distance map is given by:

$$d_T(x_i, y_i) = \sqrt{(x_i - x_T)^2 + (y_i - y_T)^2} \tag{4}$$

where (x_T, y_T) are the coordinates of the target pixel and (x_i, y_i) the coordinates of each pixel in the image. The distance map $\mathbf{d}_T$ provides additional information for training the localization task. Nevertheless, the accurate prediction of the norm values for the most distant pixels is difficult given that less visual cues are present. This has a negative effect on the global accuracy of the prediction due to the excessive importance given to the less relevant distant pixels. Thus, we use a location map with higher variability near the target location, which is obtained by applying an exponential decay that saturates at the distant pixels. The proposed location map $\mathbf{y}_L$ is defined as:

$$\mathbf{y}_L = 1 + tanh\left(-\mathbf{d}_T \frac{\pi}{\beta}\right) \tag{5}$$

where *tanh* is a hyperbolic tangent function, β the saturation distance, and $\mathbf{d}_T$ the original Euclidean distance map. For the experiments in this work, we set the saturation distance β to the value of the approximate optic disc radius. An illustration of the proposed location map for a given target location is shown in Fig. 3. The localization tasks are then trained using a mean squared error (MSE) loss between the target location map $\mathbf{y}_L$ and the network output.

A straightforward approach can be used to recover the resulting location coordinates from the predicted location map by detecting the pixel of maximum response.

2.3. Segmentation of anatomical structures of the retina

The segmentation of the retinal vasculature and the optic disc is approached following the same formulation. Both tasks consist in the prediction of pixel-level labels within two categories: the anatomical structure of interest and the background. The training of these tasks is performed with a set $\{(\mathbf{r}, \mathbf{y_s})_1, \ldots, (\mathbf{r}, \mathbf{y_s})_N\}$ where $\mathbf{r}$ denotes the fundus image and $\mathbf{y_s}$ denotes its corresponding ground truth segmentation map. The objective is to obtain the transformation $\mathbf{f}_s$ that assigns the likelihood of belonging to the anatomical structure of interest to each pixel of the fundus image.

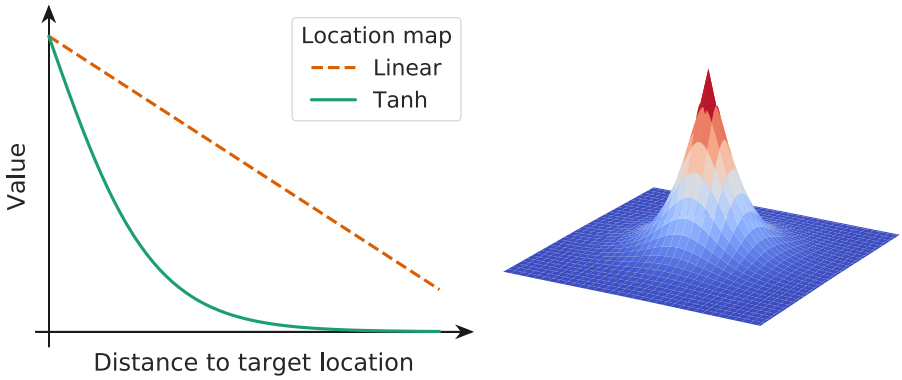


Fig. 3. (a) Value of the location map as a function of the distance to the target location. The hyperbolic tangent (tanh) version is the one used in this work, whereas the linear version is provided for comparison. (b) The location map represented as a three-dimensional surface.

These binary classifications are trained optimizing the cross-entropy loss between ground truth and network output, defined as:

$$\mathcal{L}_S(\mathbf{f}_s(\mathbf{r}), \mathbf{y}_s) = -\sum_{\Omega_R} \mathbf{y}_s log(\mathbf{f}_s(\mathbf{r})) + (1 - \mathbf{y}_s)(log(1 - \mathbf{f}_s(\mathbf{r}))) \quad (6)$$

where $\mathbf{r}$ is the input retinography, $\mathbf{y}_s$ the corresponding ground truth binary map, $\mathbf{f}_s(\mathbf{r})$ the output of the network, and Ω_R the retinography FOV where the loss is computed.

2.4. Network architecture

In this work, we use the U-Net architecture [35] for all the reconstruction, localization, and segmentation tasks. U-Net is a commonly used network in many medical imaging applications, and a well-known and proven baseline. In that sense, in order to ensure an strongly validated baseline, we use the same exact network that was proposed by Ronneberger et al. [35], including the same number of layers and channels, without any additional adjustments. The only exception is the number of output channels, which inevitably depends on the output that is required for each specific problem. A general scheme of the network, including details of the different layers, is depicted in Fig. 4. Specifically, U-Net is a fully convolutional neural network with output and input of the same size. This allows the estimation of a full size target image map, which represents an useful property for segmentation or reconstruction tasks, as well as for the prediction of location maps.

This architecture presents a multiscale encoder–decoder structure, featuring skip connections between their respective inner blocks. In the encoder part, the width and height image dimensions are progressively reduced by half at subsequent blocks, using max pooling operations. Following the idea of the VGG networks [36], these blocks are composed of two convolutional layers with kernel size 3×3 followed by the spatial max pooling operation. The objective of the progressive reduction in space is to enforce the learning of broad and abstract patterns from the data. This helps to produce a hierarchical representation from low to high level features in which the input data is transformed. The decoder part progressively recovers the width and height of the input images, by building the output from the high level abstractions to the low level details. The progressive upsampling is produced with strided transposed convolutions that increase the spatial dimensions by a factor of 2 at each block. These transposed convolutions are interleaved between convolution layers like those in the encoder.

The width and height variations across the network create a bottleneck effect that enforces the learning of high level patterns. However, the spatial contraction penalizes the tracking of the precise localization of the extracted features. U-Net successfully improves the localization and generation of small details with the inclusion of skip connections between encoder and decoder. These connections transfer features from the encoder to the decoder at different resolutions, providing alternative paths to propagate precise spatial localizations.

All the convolutional layers of the network are followed by ReLU activation functions except for the last layer. In the case of the segmentation tasks, a sigmoid activation function is used at the output layer of the network, whereas for the localization and the multimodal reconstruction tasks a linear activation function is used instead.

2.5. Network training

When the network is trained from scratch, the parameters are randomly initialized following the method proposed by He et al. [37]. The Adam [38] algorithm is used for the optimization of the loss functions. The decay rates for the first and second order moments of Adam are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively, as originally proposed by Kingma and Ba [38]. The initial learning rate is set to $\alpha = 1e-4$ for the multimodal reconstruction and $\alpha = 1e-5$ for all the localization and segmentation tasks. The learning rate schedule is the same for all the experiments. It consists in the reduction of the learning rate by a factor of 10 when the validation loss does not improve for 2500 iterations. Each iteration consists in a network parameters update due to the presentation of a training minibatch, which is fixed to consist of one image in all the experiments. The training stops when the validation loss stalls after reaching a learning rate of $\alpha = 1e-7$. These parameters were empirically established as those that were observed to provide enough training for all the tasks.

For the target tasks the datasets are initially divided into training and hold-out test sets, whereas for the pre-training task the whole dataset is used during training. In order to control the learning rate schedule and the stopping criteria, the training sets are additionally divided into training and validation subsets. In this work, several experiments are performed varying the number of training samples used. Therefore, for each experiment, the samples that are not selected for training are included into the validation subset. In the experiments where the whole training data is used, there is no validation subset, and the schedule

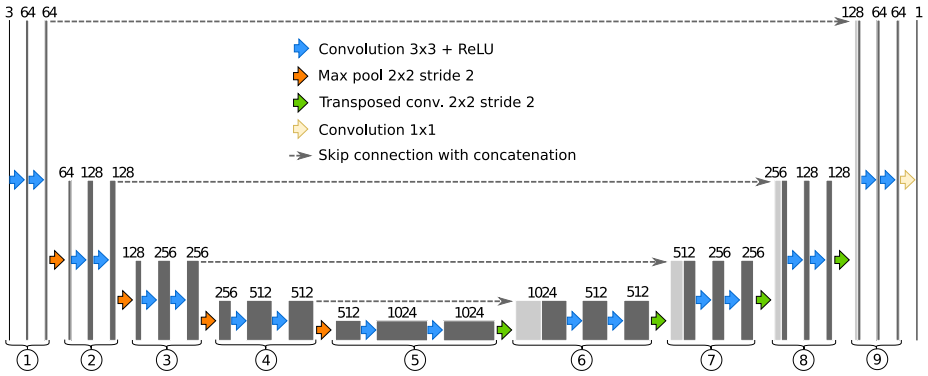


Fig. 4. Description of the U-Net architecture.

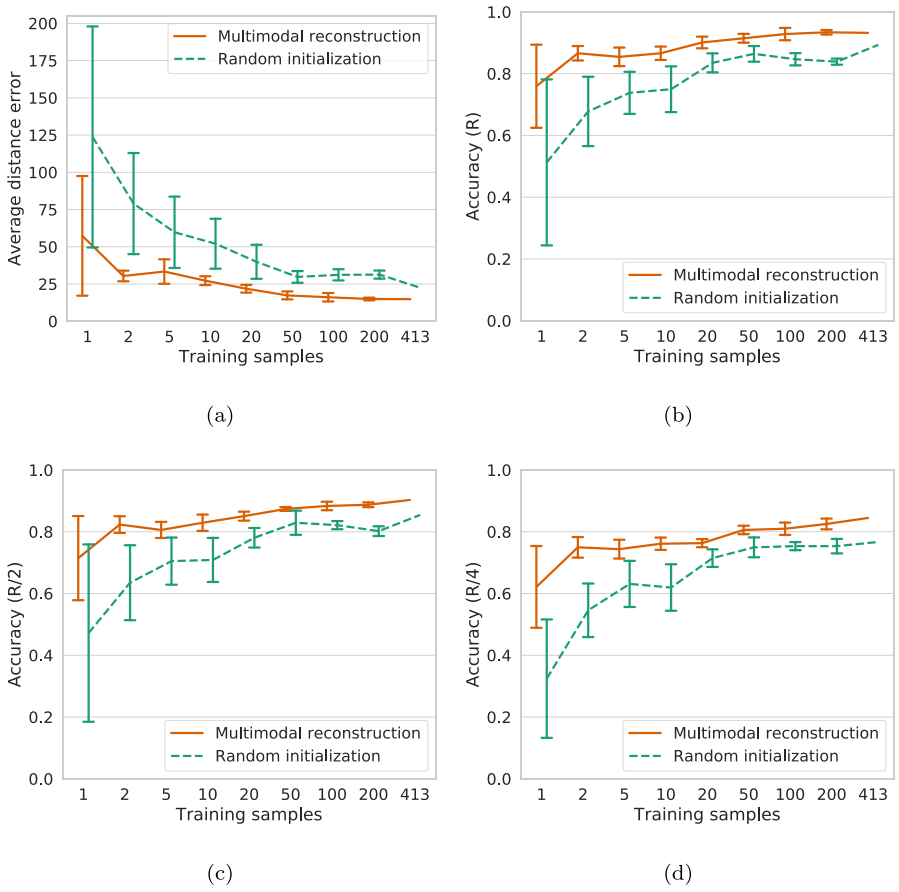


Fig. 5. Results of the fovea localization for a varying number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against the training from scratch (Random initialization). (a) Average distance error in pixels and ((b), (c), (d)) accuracy for (b) R, (c) R/2, and (d) R/4 criteria. The means and standard deviations are computed for each experiment from 5 repetitions with 5 different training subsets.

resulting from the previous experiment with more data samples is applied.

To avoid excessive overfitting, data augmentation techniques and dropout are also used in both the target and pre-training tasks. In that sense, we apply the same data augmentation techniques as Hervella et al. [20], including colour and spatial augmentations. The colour augmentations consists in random linear transformations of the image channels using the HSV colour representation. The spatial augmentation consists in random affine transformations with scaling, rotation, and shearing components. Dropout layers with probability $p = 0.2$ are added to the network after the convolutional blocks 2, 3, 4, 5, and 6, which are depicted in Fig. 4.

3. Experiments and results

In order to quantify and demonstrate the advantages of the proposed approach, the self-supervised multimodal pretraining is compared against training the networks from scratch, which is the standard alternative without requiring additional annotated data. In this way, the same experiments were conducted for two different frameworks:

- **Multimodal reconstruction**: The neural network is pre-trained on the unlabelled multimodal data using the self-supervised multimodal reconstruction. Then, the network is fine-tuned using the annotated data of the target task.

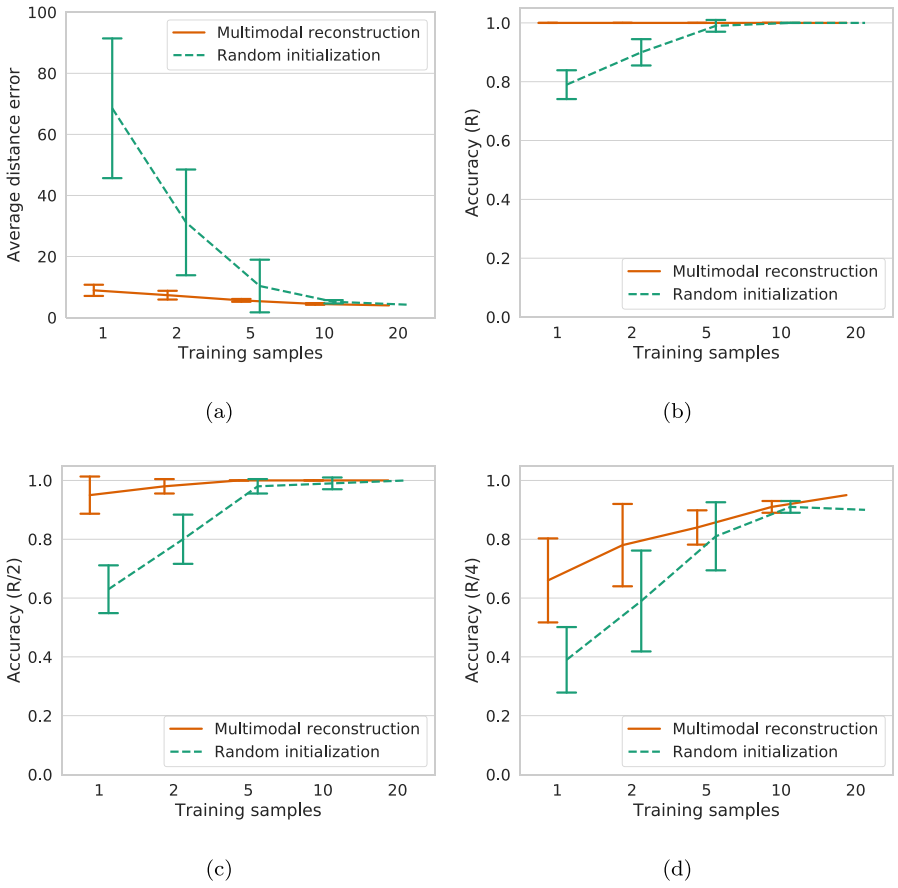


Fig. 6. Results of the optic disc localization for a varying number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against the training from scratch (Random initialization). (a) Average distance error in pixels and ((b), (c), (d)) accuracy for (b) R, (c) R/2, and (d) R/4 criteria. The means and standard deviations are computed for each experiment from 5 repetitions with 5 different training subsets.

- **Random initialization**: The neural network is randomly initialized and trained from scratch using the annotated data of the target task.

In order to guarantee an adequate and fair comparison, the same network architecture and training strategy was used for both frameworks, as described in Section 2. Additionally, the same settings were also used in all the studied tasks, except for the training loss and the output layer of the network, which require to be specific for each task objective (segmentation or localization).

In general terms, several experiments were mainly performed to study whether the use of the multimodal reconstruction as self-supervised pre-training may alleviate the impact of having a very small number of annotated samples. To that end, we performed experiments with a varying progressive number of training samples, ranging from a single image to the whole training set, while keeping the same hold-out test set for the evaluation. In most of these experiments, only a subset of the available training data is actually used for training. Thus, for each experiment, different combinations of the available training samples are possible. The variability regarding the selection of these training samples may have an effect in the performance of the networks. In order to take this variability into account, we performed 5 repetitions for each experiments using 5 different training subsets. These subsets are randomly selected from all the possible combinations of the available training data. The only exception to this procedure was the experiment with the whole training set, where all the training data was used for a single repetition. Additionally, in order to ensure a fair comparison, the same randomly selected training subsets were used for both frameworks.

Finally, the performance of both frameworks is compared against that of state-of-the-art approaches for fovea localization, optic disc localization, vessel segmentation, and optic disc segmentation. The objective of this comparison is to ensure that the proposed methods, despite being general and of straightforward use, can reach state-of-the-art performance in the tested tasks.

3.1. Datasets

The experiments presented in this paper were all conducted using five of the most representative publicly available datasets, which are described below:

- Isfahan MISP [39]: This dataset was used for the self-supervised pre-training consisting in the multimodal reconstruction between retinography and angiography. The dataset comprises 59 retinography–angiography pairs with image sizes of 720 × 576 pixels. Half of the samples correspond to pathological cases that were obtained from patients diagnosed with diabetic retinopathy. The other half correspond to healthy cases. All the images in this dataset are used for training.
- DRIVE [40]: This dataset was used for the training and evaluation of the blood vessel segmentation and optic disc localization. DRIVE is a collection of 40 retinographies with their corresponding ground truth vessel segmentations. The ground truth optic disc locations, instead, are not publicly available and were manually annotated by a clinical expert

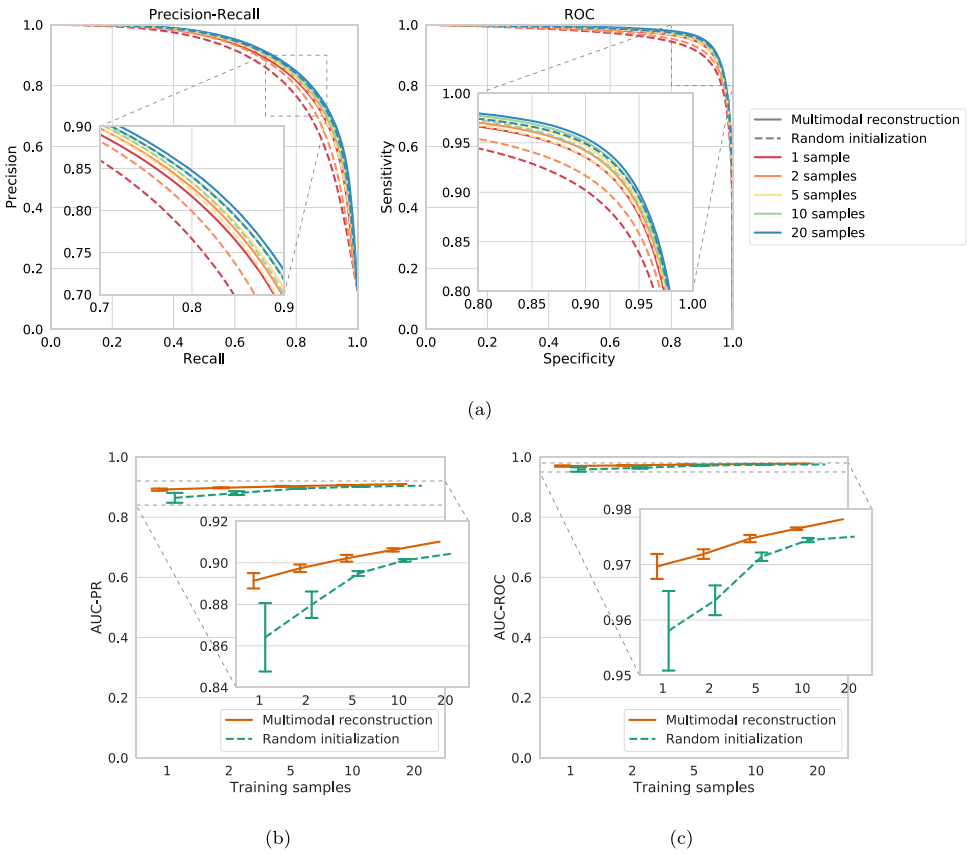


Fig. 7. Results of the blood vessels segmentation for a varying number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against the training from scratch (Random initialization). (a) Mean PR and ROC curves, (b) AUC-PR, and (c) AUC-ROC for a varying number of training samples. The means and standard deviations are computed for each experiment from 5 repetitions with 5 different training subsets.

in our case. These annotations consist of the pixel coordinates for the optic disc centre. The images present a size of 565 × 584 pixels and the approximate optic disc radius is 40 pixels. This value is used as the saturation distance β in Eq. (5) to compute the optic disc location maps. We use the standard split for this dataset, which results in 20 images used as training set and the remaining 20 images hold out for the evaluation.

- DRIONS [41]: This dataset was used for the training and evaluation of the optic disc segmentation. DRIONS includes a collection of 110 retinographies with their corresponding ground truth optic disc segmentations. The images have a size of 700 × 605 pixels. We use the same data split that Maninis et al. [3], consisting of 60 images for training and the remaining 50 images hold out for the evaluation.
- IDRiD [42]: This dataset was used for the training and evaluation of the fovea localization. IDRiD contains 516 retinographies including different grades of diabetic retinopathy. The provided ground truth annotations for the fovea localization consist in the pixel coordinates of the fovea centre. The images have a size of 4288 × 2848 pixels, being, therefore, significantly larger than the images from the Isfahan MISP dataset used for pre-training. The size of the retinal structures in the images also differs. For this reason, the images are rescaled to a fixed size of 858 × 570, for which the approximate optic disc radius is 50 pixels. This value is used as the saturation distance β in Eq. (5) to compute the fovea location maps. We use the standard split for this dataset, consisting of 413 images for training and the remaining 103 images hold out for the evaluation.
- MESSIDOR [8]: This dataset was used for the evaluation of the fovea localization. MESSIDOR is a collection of 1200 retinographies including different grades of diabetic retinopathy. From them, we use 1136 images, for which the ground truth fovea localizations were provided by Gegundez-Arias et al. [22]. The dataset includes images of three different sizes. As happens with IDRiD, the scale of the retinal structures is significantly different to that of the pre-training dataset. Therefore, the original image sizes of 2240 × 1488, 1440 × 960, and 2304 × 1536 are rescaled to 1120 × 744, 1080 × 720, and 1152 × 768, respectively, to match the scale of the other datasets. The approximate optic disc radii are also provided by Gegundez-Arias et al. [22] and are rescaled in the same proportion than the images. In this case, all the images are used as test set for comparison with the state-of-the-art.

3.2. Evaluation metrics

For the localization of the optic disc and the fovea, the performance was evaluated following the strategy that is typically used in the literature [4,23]. First, the euclidean distance between the predicted location and the ground truth location is computed. If this distance is lower than a certain threshold, the prediction is considered successful. The accuracy, defined as the ratio between the successful predictions and the total number of images, is used for the assessment of the performance. In order to obtain a more complete analysis, this accuracy is computed using different progressive thresholds. Particularly, we use R, R/2, and R/4, where R denotes the approximate optic disc radius, which is indicated for each dataset in Section 3.1. Additionally, the average distance in pixels is also used as evaluation metric.

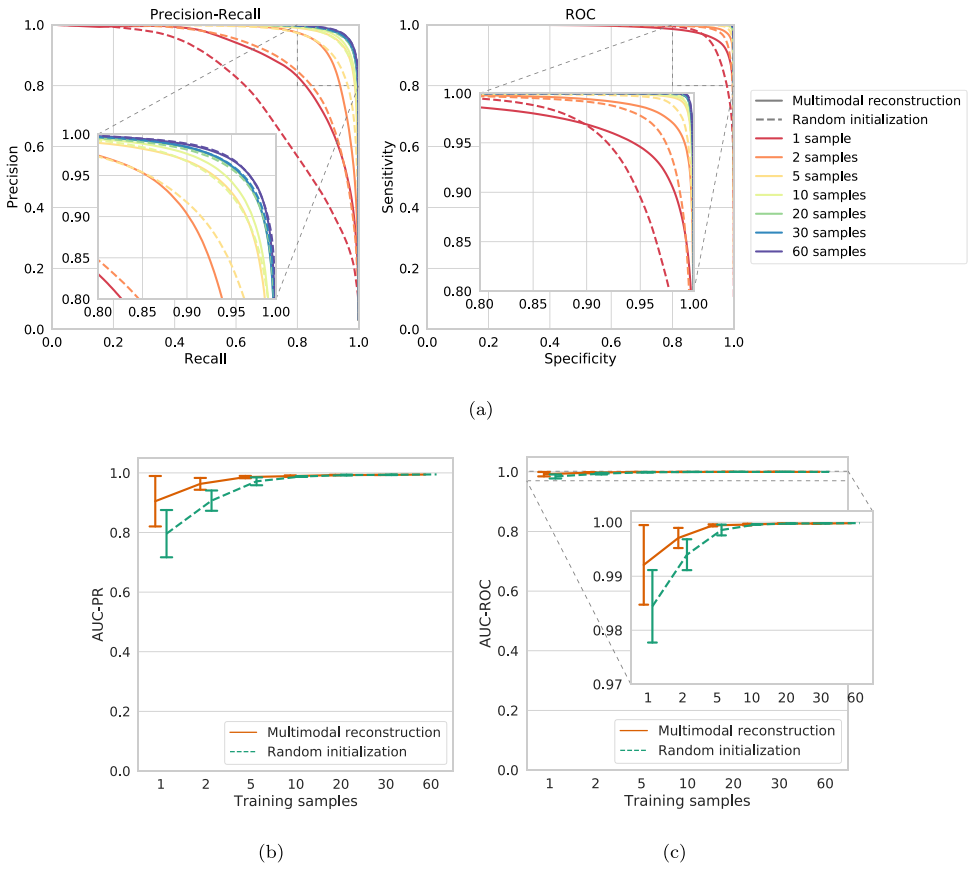


Fig. 8. Results of the optic disc segmentation for a varying number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against the training from scratch (Random initialization). (a) Mean PR and ROC curves, (b) AUC-PR, and (c) AUC-ROC for a varying number of training samples. The means and standard deviations are computed for each experiment from 5 repetitions with 5 different training subsets.

Regarding the segmentation tasks, Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves were used to assess the performance. Both curves are commonly used in binary decision problems, allowing the evaluation of the generated probability maps without selecting the decision threshold. Note that the difference between ROC and PR curves is significative when the target classes are unbalanced. In our case, for the vessels and the optic disc segmentation, the number of samples from the positive class, i.e., vessels or optic disc, is significantly lower than the number of samples from the negative class, i.e., background. In this scenario, PR curves are more sensitive to variations in the false positive number, which leads to a greater performance discrimination ability. Despite this, ROC curves are widely used in the literature as a default metric in retinal imaging, specially for vessel segmentation [31,32]. For such reason, we include both complementary curves in our evaluation. Additionally, the area under the ROC curve (AUC-ROC) and the area under the Precision–Recall curve (AUC-PR) were used.

Finally, for all the target tasks, mean values and standard deviations of the evaluation metrics are computed from the 5 repetitions with 5 different trainings subsets that are performed for each experiment. Additionally, in the case of the segmentation tasks, mean ROC and PR curves are also computed. The only exception to this procedure happens for the experiments with the whole training set, given that all the training samples are used for a single repetition in that case.

3.3. Results

The results for the fovea localization and the optic disc localization are depicted in Figs. 5 and 6, respectively. It is observed that the use of the self-supervised multimodal pre-training improves the performance of the localization process of both anatomical structures. In particular, this improvement happens in terms of both average value and standard deviation. In the case of the fovea (Fig. 5), the improvement is significant for any number of training samples, whereas in the case of the optic disc (Fig. 6), the random initialization approach reaches the performance of the proposed method only when all the training data is used. The latter is due to the fact that the multimodal reconstruction framework has already almost converged to the maximum performance with a smaller number of annotated samples.

The results for blood vessel segmentation and optic disc segmentation are depicted in Figs. 7 and 8, respectively. It is observed that the use of the self-supervised multimodal pre-training also improves the performance for the segmentation of both anatomical structures. In the case of the optic disc segmentation (Fig. 8), the random initialization approach reaches the performance of the proposed method when half the training data is used. As with the optic disc localization, this is due to the fact that the multimodal reconstruction framework has already converged. Regarding the blood vessel segmentation (Fig. 7), the improvement is obtained using any number of training samples. In fact, as illustrated in the plots of Fig. 7(b) and (c), the trend may have continued if more training samples were also used.

A notorious difference between the localization and the segmentation results is that the latter show a smaller difference between the two frameworks in the comparison. This is a consequence of the high performance that is already achieved by training the networks from scratch, which leaves only a little gap for improvement. However, even in this highly competitive scenario, the self-supervised multimodal pre-training gets to improve the performance.

Table 1
Comparison with state-of-the-art methods for the fovea localization. The means and standard deviations in our experiments are computed from 5 repetitions with 5 different training subsets.

		Accuracy (%)		
		R	R/2	R/4
Evaluation on MESSIDOR				
Gegundez-Arias et al. [22]		96.50	95.88	94.25
Yu et al. [28]		98.00	94.00	64.88
Niemeijer et al. [21]		97.38	96.00	93.25
Dashtbozorg et al. [26]		98.87	93.75	66.50
Al-Bander et al. [4]		96.60	91.40	66.80
Meyer et al. [5]		99.74	97.71	94.01
Ours (1 image)	Random init.	54.52 ± 36.23	53.86 ± 36.01	48.98 ± 32.08
	Multimodal	86.09 ± 19.02	85.49 ± 19.18	80.07 ± 21.20
Ours (2 images)	Random init.	83.64 ± 9.41	83.17 ± 9.36	78.93 ± 9.51
	Multimodal	98.33 ± 0.59	97.94 ± 0.52	94.35 ± 1.21
Ours (200 images)	Random init.	99.47 ± 0.06	99.26 ± 0.09	97.02 ± 0.38
	Multimodal	99.84 ± 0.07	99.54 ± 0.13	97.80 ± 0.15
Ours (413 images)	Random init.	99.91	99.56	97.54
	Multimodal	100.00	99.65	97.98
Evaluation on IDRiD				
Ours (1 image)	Random init.	51.26 ± 26.87	47.18 ± 28.72	32.43 ± 19.17
	Multimodal	75.92 ± 13.46	71.46 ± 13.63	62.14 ± 13.24
Ours (2 images)	Random init.	67.77 ± 11.23	63.50 ± 12.14	54.56 ± 8.67
	Multimodal	86.60 ± 2.33	82.33 ± 2.70	74.95 ± 3.33
Ours (200 images)	Random init.	83.88 ± 0.99	80.19 ± 1.58	75.34 ± 2.35
	Multimodal	93.40 ± 0.73	88.74 ± 0.78	82.52 ± 1.74
Ours (413 images)	Random init.	89.32	85.44	76.70
	Multimodal	93.20	90.29	84.47

Table 2
Comparison with state-of-the-art methods for the optic disc localization. The means and standard deviations in our experiments are computed from 5 repetitions with 5 different training subsets.

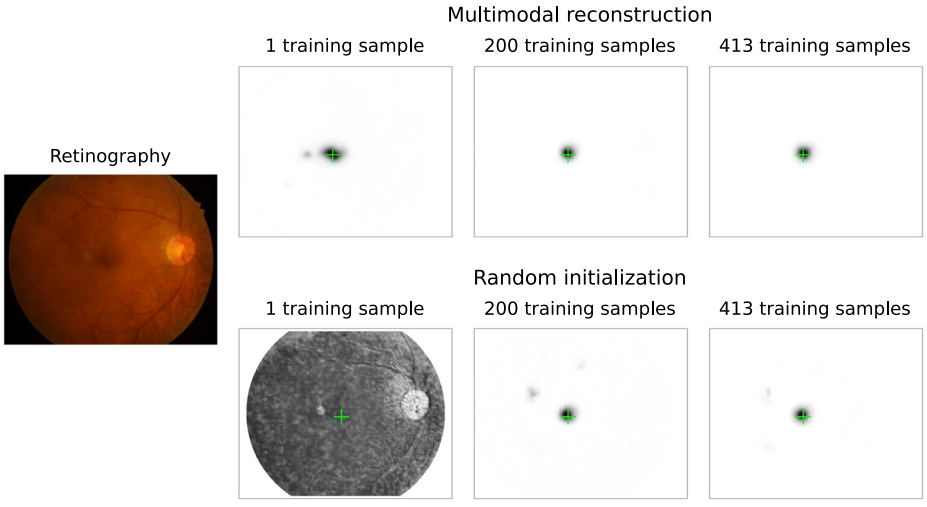
		Accuracy (%)		
		R	R/2	R/4
Al-Bander et al. [4]	(MESSIDOR)	97.00	95.00	83.60
Marin et al. [23]	(MESSIDOR)	99.75	99.50	97.75
Zhu et al. [24]		90.00	–	–
Qureshi et al. [25]		100.00	–	–
Dehghani et al. [27]		100.00	–	–
Ours (1 image)	Random init.	79.00 ± 4.90	63.00 ± 8.12	39.00 ± 11.14
	Multimodal	100.00 ± 0.00	95.00 ± 6.32	66.00 ± 14.28
Ours (2 images)	Random init.	90.00 ± 4.47	80.00 ± 8.37	59.00 ± 17.15
	Multimodal	100.00 ± 0.00	98.00 ± 2.45	78.00 ± 14.00
Ours (10 images)	Random init.	100.00 ± 0.00	99.00 ± 2.00	91.00 ± 2.00
	Multimodal	100.00 ± 0.00	100.00 ± 0.00	91.00 ± 2.00
Ours (20 images)	Random init.	100.00	100.00	90.00
	Multimodal	100.00	100.00	95.00

Table 3
Comparison with state-of-the-art methods for the blood vessels segmentation. The means and standard deviations in our experiments are computed from 5 repetitions with 5 different training subsets.

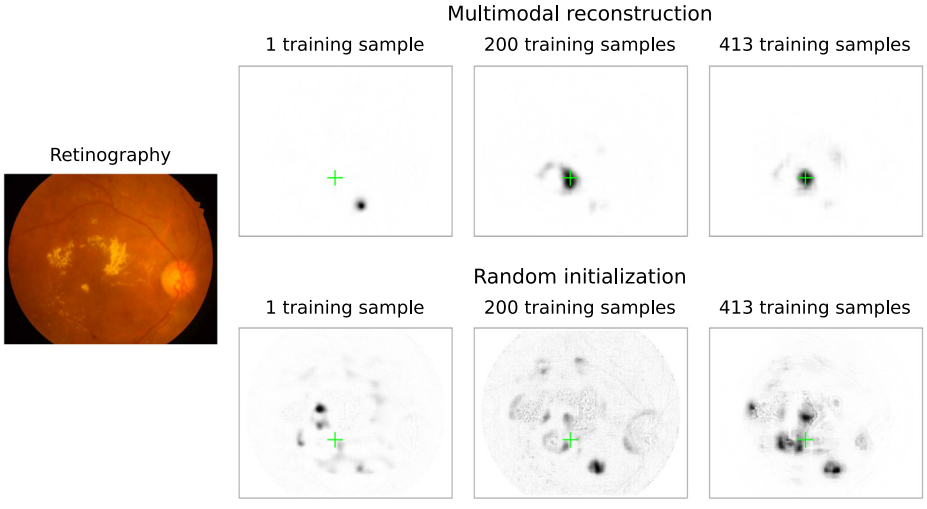
		AUC-PR (%)	AUC-ROC (%)
Fraz et al. [29]		–	97.47
Liskowski and Krawiec [31]		–	97.90
Li et al. [30]		–	97.38
Maninis et al. [3]		90.64	97.93
Mo and Zhang [32]		–	97.82
Ours (1 image)	Random init.	86.41 ± 1.65	95.81 ± 0.72
	Multimodal	89.14 ± 0.37	96.97 ± 0.22
Ours (2 images)	Random init.	87.98 ± 0.64	96.36 ± 0.27
	Multimodal	89.74 ± 0.18	97.19 ± 0.08
Ours (10 images)	Random init.	90.12 ± 0.06	97.44 ± 0.04
	Multimodal	90.62 ± 0.08	97.65 ± 0.02
Ours (20 images)	Random init.	90.44	97.51
	Multimodal	91.02	97.82

Table 4
Comparison with state-of-the-art methods for the optic disc segmentation. The means and standard deviations in our experiments are computed from 5 repetitions with 5 different training subsets.

		AUC-PR (%)	AUC-ROC (%)
Maninis et al. [3]		99.57	99.98
Ours (1 image)	Random init.	79.62 ± 7.93	98.44 ± 0.67
	Multimodal	90.50 ± 8.46	99.21 ± 0.74
Ours (2 images)	Random init.	90.69 ± 3.41	99.40 ± 0.29
	Multimodal	96.33 ± 1.98	99.71 ± 0.19
Ours (30 images)	Random init.	99.37 ± 0.10	99.98 ± 0.00
	Multimodal	99.31 ± 0.03	99.98 ± 0.00
Ours (60 images)	Random init.	99.49	99.98
	Multimodal	99.45	99.98



(a)

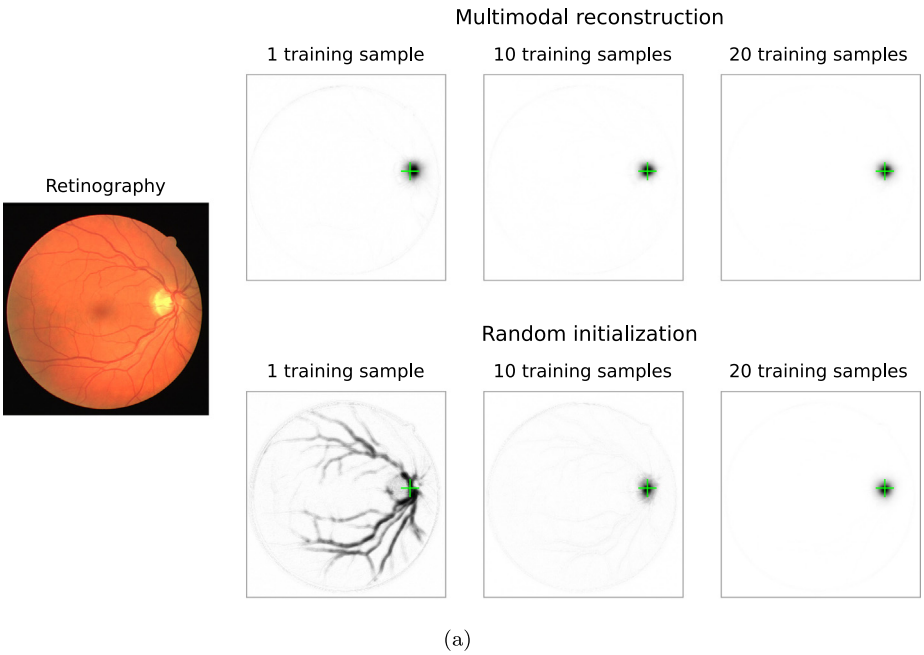


(b)

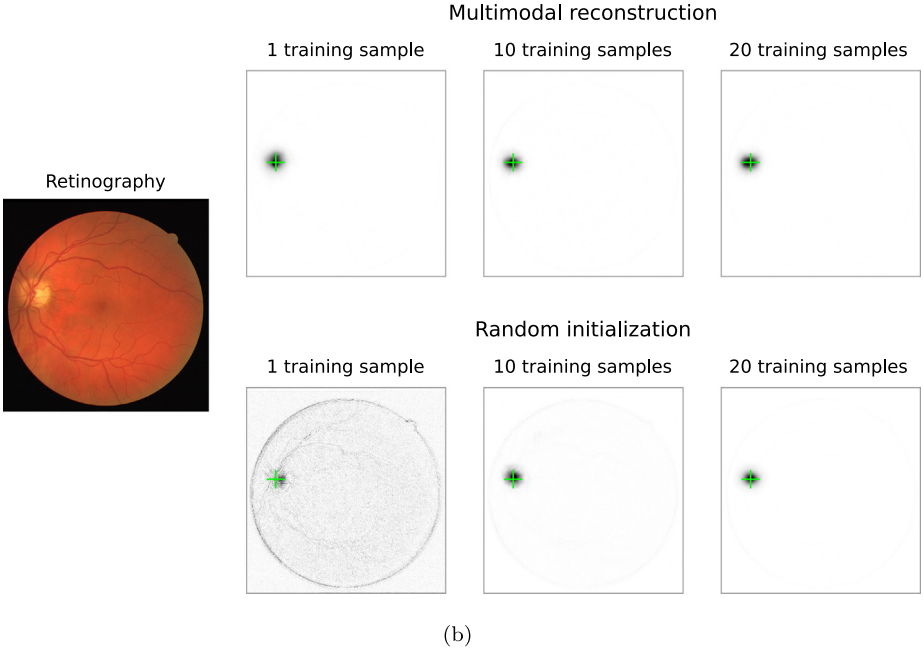
Fig. 9. Examples of predicted location maps for the fovea using different number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against training from scratch (Random initialization). The green cross depicts the ground truth location.

In addition, the comparison with state-of-the-art methods is respectively shown in Table 1 for the fovea localization, Table 2 for the optic disc localization, Table 3 for the blood vessel segmentation, and Table 4 for the optic disc segmentation. It is observed that both the multimodal reconstruction and the random initialization frameworks reached competitive performance in all the studied tasks. However, we would like to remark that the proposed self-supervised multimodal pre-training approach leads to state-of-the-art performance with much less annotated data.

Regarding the fovea localization, our experiments were performed using the recently published IDRiD dataset. In order to perform a comparison with state-of-the-art approaches we include additional results of our proposal evaluated on the MESSIDOR dataset. This additional evaluation is performed using the networks that were previously trained using the IDRiD dataset.



(a)



(b)

Fig. 10. Examples of predicted location maps for the optic disc using different number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against training from scratch (Random initialization). The green cross depicts the ground truth location.

Table 1 shows that the results obtained for MESSIDOR are better than those obtained for IDRiD. We have to consider, in this case, the higher percentage of pathological cases and advanced severity stages that is present in the IDRID dataset.

In the case of the optic disc localization, existent approaches evaluated on the DRIVE dataset only report accuracy for a distance threshold of value R, i.e., the approximate optic disc radius. Thus, as additional reference, we include two representative works that were evaluated using the MESSIDOR dataset and manually labelled ground truths. In this case, the labels were not publicly available.

Finally, as illustration for qualitative comparison, examples of results obtained with the multimodal reconstruction and the random initialization frameworks are provided in Figs. 9, 10, 11, and 12. In particular, Figs. 9 and 10 depict representative examples of predicted location maps for the fovea and the optic disc, respectively. In addition, Figs. 10 and 11 depict representative examples of predicted segmentation maps for the vasculature and the optic disc, respectively. All the examples correspond to images from the evaluation sets, and the ground truth annotations are provided as reference.

In general, it is observed that the multimodal reconstruction approach produces similar or even better results than the random initialization approach when all the training data is used. Nevertheless, due to the competitive performance of both frameworks, the visual comparison of the results can be difficult, requiring a more detailed analysis that is out of the scope of this paper. In contrast, when the training data is reduced, the contribution of the self-supervised multimodal pre-training is easier to appreciate with a rough visual analysis. In that sense, the improvement is

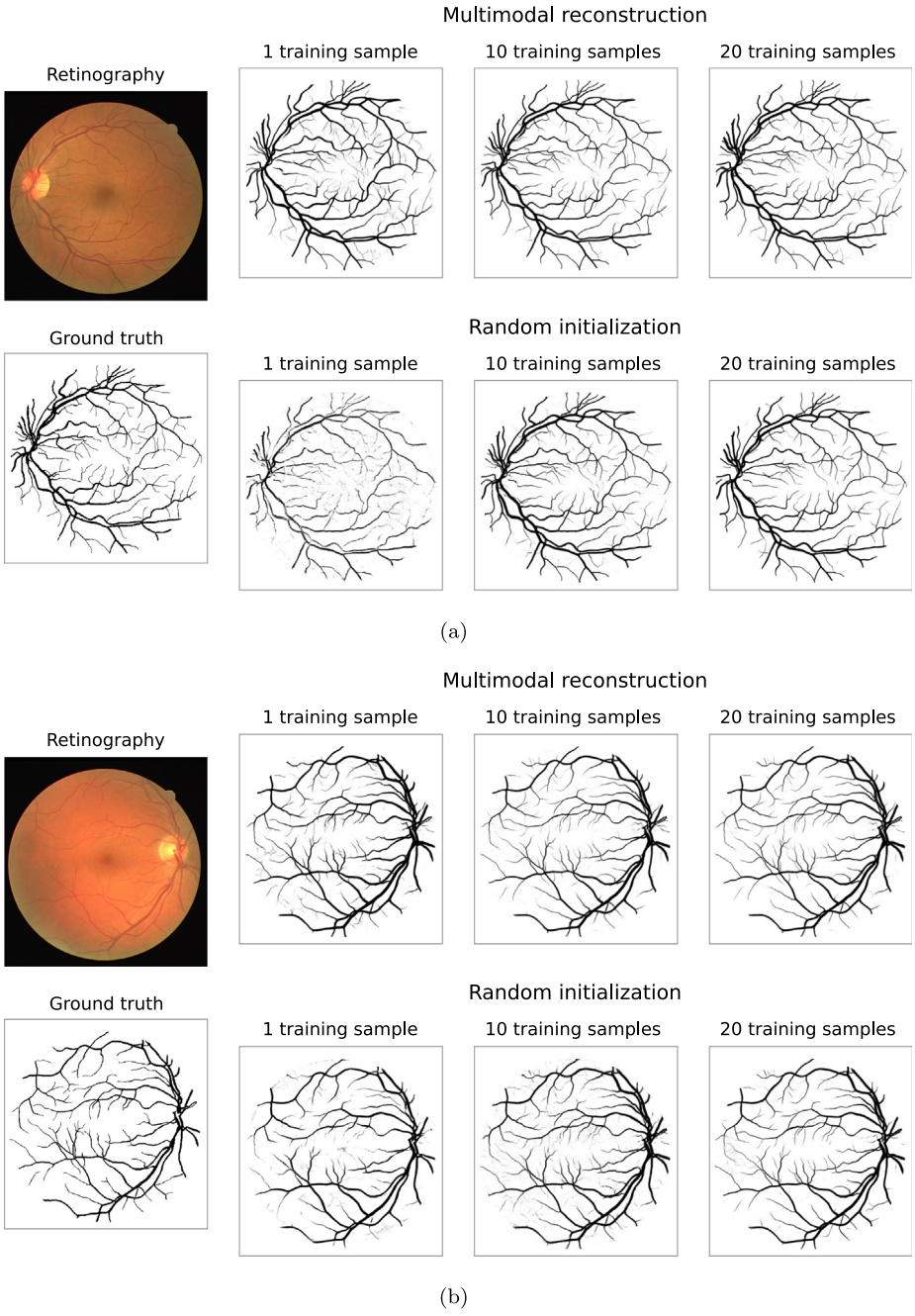


Fig. 11. Examples of predicted segmentation maps for the retinal vasculature using different number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against training from scratch (Random initialization).

especially significant when a single training sample is used, which represents the most challenging scenario in this scope.

Additionally, even greater improvement is that of the example in Fig. 9(b). In this case, the multimodal reconstruction leads to an important improvement when all the training data is used with respect to the random initialization counterpart. This is caused by the presence of lesions in the retina, which evidences that the proposed self-supervised multimodal pre-training presents the potential of being especially helpful in the more complex pathological cases.

4. Discussion

In this work, we address the problem of training DNNs for the localization and segmentation of the main anatomical structures of the eye fundus in retinography using scarce annotated data. To that end, we propose the use of the multimodal reconstruction between retinography and fluorescein angiography as a common self-supervised pre-training task; and the later fine-tuning of the pre-trained DNN for fovea localization, optic disc localization, blood vessel segmentation, and optic disc segmentation using a limited amount of task-specific annotated data. Given that obtaining the best possible results is not our main objective, we use the same network and training methodology for all the considered tasks. The only difference is the training loss, which requires to be specific for each kind of task: reconstruction, localization, or segmentation. Additionally, as neural network architecture, we employ the original U-Net [35], which is a reliable baseline that was previously applied in this retinal context with a satisfactory

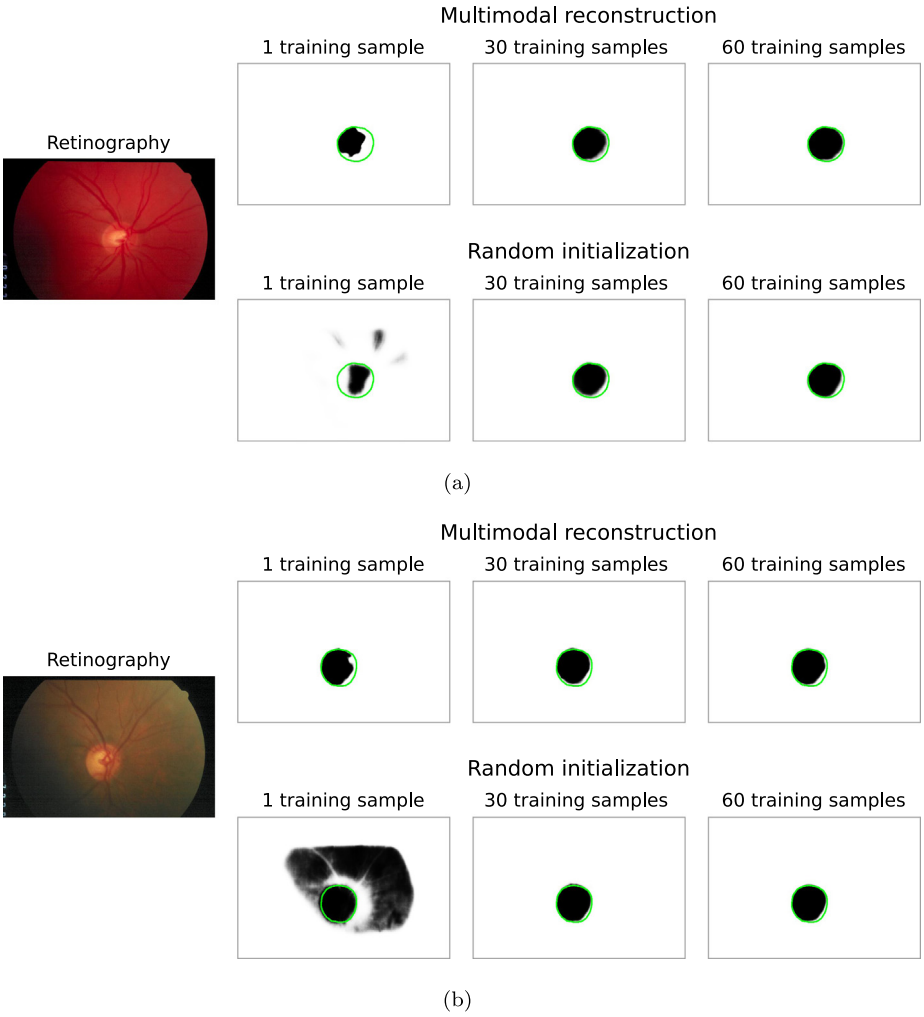


Fig. 12. Examples of predicted segmentation maps for the optic disc using different number of training samples and comparison of the proposed self-supervised pre-training (Multimodal reconstruction) against training from scratch (Random initialization). The boundary of the ground truth segmentation is depicted in green.

performance [19]. Incidentally, the experimental results demonstrate that state-of-the-art performance can be achieved in all the studied tasks with the same network architecture and training strategy without further specific tuning.

From the comparison between the multimodal reconstruction and the random initialization frameworks, it is observed that the proposed self-supervised multimodal pre-training improves the obtained performance in all the studied tasks. Nevertheless, the extent of this improvement is not the same for all the tasks or training data sizes. The most remarkable improvement is observed in all the tasks when only few annotated images are used for training. In fact, the results that are obtained training from scratch with all the annotated data can be achieved using a fraction of the annotations if the networks are, instead, pre-trained with the proposed multimodal reconstruction. A negligible improvement of the proposed approach happens only for the cases where highly competitive performance is already obtained by the random initialization counterpart. Naturally, the beneficial effect of using the multimodal reconstruction as pre-training is limited by the room left for improvement by the baseline approach. For example, this is the case of some experiments involving the optic disc. However, in any case, the multimodal reconstruction approach converges to the maximum performance with less annotated data. In this regard, the results indicate that the optic disc localization and segmentation tasks are easier in comparison to the others in our experiments.

The provided comparison with state-of-the-art works shows that both the multimodal reconstruction and the random initialization frameworks produce competitive results when using all the training data. In that sense, the strong baseline ensures the practical relevance of the conclusions drawn from our analysis. Additionally, for some experiments, the random initialization framework behaves reasonably well with moderate reductions in the training data. This shows that modern data augmentation practices, adequate training schedules, and well designed loss functions are key to the successful application of DNNs to standard medical image analysis applications, without even needing any bells and whistles to fine tune the network architecture.

Regarding the self-supervised multimodal pre-training, the provided comparisons demonstrate that competitive results can also be achieved using a fraction of the total annotated training data. This is a strong result, indicating that clinical applications based on deep learning methods can be produced without requiring large amounts of manually annotated images. Additionally, we have demonstrated the advantages of the multimodal reconstruction as stand-alone transfer learning strategy. However, the proposed pre-training could also be applied together with other complementary self-supervised tasks in settings similar to those already explored in other domains [16]. In that sense, future works could explore the complementary application of the multimodal reconstruction and other self-supervised approaches in the medical domain.

Finally, other benefit of the proposed self-supervised pre-training is that, in general, the variability due to the use of different training samples is significantly reduced. However, this variability is still high when fewer annotations are used. Incidentally, this indicates that some images are considerably more adequate for training than others in order to achieve a better generalization. Thus, despite that a competitive performance can be achieved with very scarce annotations, for some applications this labelled data efficiency could be limited by the appropriate selection of particular training samples. In those situations, it would be interesting to explore the use of techniques aiming at the selection of the most informative images for being annotated.

5. Conclusions

Despite the great success of deep neural networks, the scarcity of annotated data is still a significant limiting factor to apply deep learning solutions to new clinical applications. In this regard, we propose to use the multimodal reconstruction as a self-supervised pre-training for different target tasks in the same application domain. We demonstrate the advantages of this proposal in the context of retinal image analysis. In particular, this work focuses on the localization and the segmentation of the main anatomical structures of the eye fundus, namely the fovea, the retinal vasculature, and the optic disc. For that purpose, we use the self-supervised multimodal reconstruction between retinography and fluorescein angiography to pre-train the networks.

The performed experiments demonstrate that using the multimodal reconstruction as self-supervised pre-training improves the performance of the considered target tasks. In particular, the proposed self-supervised transfer learning strategy allows to produce state-of-the-art results with a significant reduction of the annotated training data. This outcome has remarkable implications for future applications of neural networks in many fields of medical imaging where multimodal data can be easily gathered.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Álvaro S. Hervella: Methodology, Software, Validation, Writing - original draft, Visualization. **José Rouco:** Conceptualization, Validation, Writing - review & editing, Supervision. **Jorge Novo:** Conceptualization, Validation, Writing - review & editing, Supervision. **Marcos Ortega:** Conceptualization, Supervision, Project administration, Funding acquisition.

Acknowledgements

This work is supported by Instituto de Salud Carlos III, Government of Spain, and the European Regional Development Fund (ERDF) of the European Union (EU) through the DTS18/00136 research project, and by Ministerio de Economía, Industria y Competitividad, Government of Spain, through the DPI2015-69948-R research project. The authors of this work also receive financial support from the ERDF and Xunta de Galicia (Spain) through Grupo de Referencia Competitiva, ref. ED431C 2016-047, and from the European Social Fund (ESF) of the EU and Xunta de Galicia (Spain) through the predoctoral grant contract ref. ED481A-2017/328. CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia (Spain), through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

References

[1] R. Besenczi, J. Tóth, A. Hajdu, A review on automatic analysis techniques for color fundus photographs, Comput. Struct. Biotechnol. J. 14 (2016) 371–384, http://dx.doi.org/10.1016/j.csbj.2016.10.001.

[2] E.D. Cole, E.A. Novais, R.N. Louzada, N.K. Waheed, Contemporary retinal imaging techniques in diabetic retinopathy: a review, Clin. Exp. Ophthalmol. 44 (4) (2016) 289–299, http://dx.doi.org/10.1111/ceo.12711.

[3] K.K. Maninis, J. Pont-Tuset, P. Arbeláez, L.V. Gool, Deep retinal image understanding, in: Medical Image Computing and Computer-Assisted Intervention, MICCAI, 2016, http://dx.doi.org/10.1007/978-3-319-46723-8_17.

[4] B. Al-Bander, W. Al-Nuaimy, B.M. Williams, Y. Zheng, Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc, Biomed. Signal Process. Control 40 (2018) 91–101, http://dx.doi.org/10.1016/j.bspc.2017.09.008.

[5] M.I. Meyer, A. Galdran, A.M. Mendon, A pixel-wise distance regression approach for joint retinal optical disc and fovea detection, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, 2018, http://dx.doi.org/10.1007/978-3-030-00934-2.

[6] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, J. Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, Appl. Soft Comput. (ISSN: 15684946) 70 (2018) 41–65, http://dx.doi.org/10.1016/j.asoc.2018.05.018.

[7] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, http://dx.doi.org/10.1016/j.media.2017.07.005.

[8] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, J.-C. Klein, Feedback on a punlicly distributed image database: the MESSIDOR database, Image Anal. Stereol. 33 (3) (2014) 231, http://dx.doi.org/10.5566/ias.1155.

[9] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans. Med. Imaging 35 (5) (2016) 1299–1312, http://dx.doi.org/10.1109/TMI.2016.2535302.

[10] A.D. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, IEEE Trans. Med. Imaging 19 (3) (2000) 203–210, http://dx.doi.org/10.1109/42.845178.

[11] B. van Ginneken, A.A.A. Setio, C. Jacobs, F. Ciompi, Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans, in: International Symposium on Biomedical Imaging, ISBI, 2015, http://dx.doi.org/10.1109/ISBI.2015.7163869.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2009, http://dx.doi.org/10.1109/CVPR.2009.5206848.

[13] A.I. Namburete, W. Xie, M. Yaqub, A. Zisserman, J.A. Noble, Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning, Med. Image Anal. 46 (2018) 1–14, http://dx.doi.org/10.1016/j.media.2018.02.006.

[14] K.C. Wong, T. Syeda-Mahmood, M. Moradi, Building medical image classifiers with very limited data using segmentation networks, Med. Image Anal. 49 (2018) 105–116, http://dx.doi.org/10.1016/J.MEDIA.2018.07.010.

[15] G. Maicas, A.P. Bradley, J.C. Nascimento, I.D. Reid, G. Carneiro, Training medical image analysis systems like radiologists, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, 2018, http://dx.doi.org/10.1007/978-3-030-00928-1_62.

[16] C. Doersch, A. Zisserman, Multi-task self-supervised visual learning, in: International Conference on Computer Vision, ICCV, 2017, http://dx.doi.org/10.1109/ICCV.2017.226.

[17] T. Ross, D. Zimmerer, A. Vemuri, F. Isensee, M. Wiesenfarth, S. Bodenstedt, F. Both, P. Kessler, M. Wagner, B. Müller, H. Kenngott, S. Speidel, A. Kopp-Schneider, K. Maier-Hein, L. Maier-Hein, Exploiting the potential of unlabeled endoscopic video data with self-supervised learning, Int. J. Comput. Assist. Radiol. Surg. 13 (6) (2018) 925–933, http://dx.doi.org/10.1007/s11548-018-1772-0.

[18] A. Jamaludin, T. Kadir, A. Zisserman, Self-supervised learning for spinal MRIs, 2017, pp. 294–302, http://dx.doi.org/10.1007/978-3-319-67558-9_34.

[19] A.S. Hervella, J. Rouco, J. Novo, M. Ortega, Retinal image understanding emerges from self-supervised multimodal reconstruction, in: Medical Image Computing and Computer-Assisted Intervention, MICCAI, 2018, http://dx.doi.org/10.1007/978-3-030-00928-1_37.

[20] A.S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised deep learning for retinal vessel segmentation using automatically generated labels from multimodal data, in: International Joint Conference on Neural Networks, IJCNN, 2019, http://dx.doi.org/10.1109/IJCNN.2019.8851844.

[21] M. Niemeijer, M.D. Abràmoff, B. van Ginneken, Fast detection of the optic disc and fovea in color fundus photographs, Med. Image Anal. 13 (6) (2009) 859–870, http://dx.doi.org/10.1016/J.MEDIA.2009.08.003.
[22] M.E. Gegundez-Arias, D. Marin, J.M. Bravo, A. Suero, Locating the fovea center position in digital fundus images using thresholding and feature extraction techniques, Comput. Med. Imaging Graph. 37 (5–6) (2013) 386–393, http://dx.doi.org/10.1016/j.compmedimag.2013.06.002.
[23] D. Marin, M.E. Gegundez-arias, A. Suero, J.M. Bravo, Obtaining optic disc center and pixel region by automatic thresholding methods on morphologically processed fundus images, Comput. Methods Programs Biomed. 118 (2) (2014) 173–185, http://dx.doi.org/10.1016/j.cmpb.2014.11.003.
[24] X. Zhu, R.M. Rangayyan, A.L. Ells, Detection of the optic nerve head in fundus images of the retina using the hough transform for circles, J. Digit. Imaging 23 (3) (2010) 332–341, http://dx.doi.org/10.1007/s10278-009-9189-5.
[25] R.J. Qureshi, L. Kovacs, B. Harangi, B. Nagy, T. Peto, A. Hajdu, Combining algorithms for automatic detection of optic disc and macula in fundus images, Comput. Vis. Image Underst. 116 (1) (2012) 138–145, http://dx.doi.org/10.1016/j.cviu.2011.09.001.
[26] B. Dashtbozorg, J. Zhang, F. Huang, B.M. ter Haar Romeny, Automatic optic disc and fovea detection in retinal images using super-elliptical convergence index filters, 2016, http://dx.doi.org/10.1007/978-3-319-41501-7_78.
[27] A. Dehghani, H.A. Moghaddam, M.S. Moin, Optic disc localization in retinal images using histogram matching, Eurasip J. Image Video Process. 2012 (2012) 1–11, http://dx.doi.org/10.1186/1687-5281-2012-19.
[28] H. Yu, S. Barriga, C. Agurto, S. Echegaray, M. Pattichis, G. Zamora, W. Bauman, P. Soliz, Fast localization of optic disc and fovea in retinal images for eye disease screening, in: Proceedings of SPIE, 2011, http://dx.doi.org/10.1117/12.878145.
[29] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen, S.A. Barman, An ensemble classification-based approach applied to retinal blood vessel segmentation, IEEE Trans. Biomed. Eng. 59 (9) (2012) 2538–2548, http://dx.doi.org/10.1109/TBME.2012.2205687.
[30] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, T. Wang, A cross-modality learning approach for vessel segmentation in retinal images, IEEE Trans. Med. Imaging 35 (1) (2016) 109–118, http://dx.doi.org/10.1109/TMI.2015.2457891.
[31] P. Liskowski, K. Krawiec, Segmenting retinal blood vessels with deep neural networks, IEEE Trans. Med. Imaging 35 (11) (2016) 2369–2380, http://dx.doi.org/10.1109/TMI.2016.2546227.
[32] J. Mo, L. Zhang, Multi-level deep supervised networks for retinal vessel segmentation, Int. J. Comput. Assist. Radiol. Surg. 12 (12) (2017) 2181–2193, http://dx.doi.org/10.1007/s11548-017-1619-0.
[33] A.S. Hervella, J. Rouco, J. Novo, M. Ortega, Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, KES, 2018, http://dx.doi.org/10.1016/j.procs.2018.07.213.
[34] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.
[35] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, MICCAI, 2015, http://dx.doi.org/10.1007/978-3-319-24574-4_28.
[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.
[37] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imageNet classification, in: International Conference on Computer Vision, ICCV, 2015, http://dx.doi.org/10.1109/ICCV.2015.123.
[38] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, ICLR, 2015.
[39] S.H.M. Alipour, H. Rabbani, M.R. Akhlaghi, Diabetic retinopathy grading by digital curvelet transform, Comput. Math. Methods Med. 2012 (2012) http://dx.doi.org/10.1155/2012/761901.
[40] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, B. van Ginneken, Ridge based vessel segmentation in color images of the retina, IEEE Trans. Med. Imaging 23 (4) (2004) 501–509, http://dx.doi.org/10.1109/TMI.2004.825627.
[41] E.J. Carmona, M. Rincón, J. García-Feijoó, J.M. Martínez-de-la Casa, Identification of the optic nerve head with genetic algorithms, Artif. Intell. Med. 43 (3) (2008) 243–259, http://dx.doi.org/10.1016/j.artmed.2008.04.005.
[42] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, F. Meriaudeau, Indian diabetic retinopathy image dataset (IDRiD), 2018, http://dx.doi.org/10.21227/H25W98.