Expert Systems With Applications 185 (2021) 115598

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis

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

Centro de Investigación CITIC, Universidade da Coruña, A Coruña, Spain
VARPA Research Group, Instituto de Investigación Biomédica de A Coruña (INIBIC), Universidade da Coruña, A Coruña, Spain

ARTICLE INFO

Keywords:
Deep learning
Medical imaging
Self-supervised learning
Eye fundus
Transfer learning
Computer-aided diagnosis

ABSTRACT

Computer-aided diagnosis using retinal fundus images is crucial for the early detection of many ocular and systemic diseases. Nowadays, deep learning-based approaches are commonly used for this purpose. However, training deep neural networks usually requires a large amount of annotated data, which is not always available. In practice, this issue is commonly mitigated with different techniques, such as data augmentation or transfer learning. Nevertheless, the latter is typically faced using networks that were pre-trained on additional annotated data.

An emerging alternative to the traditional transfer learning source tasks is the use of self-supervised tasks that do not require manually annotated data for training. In that regard, we propose a novel self-supervised visual learning strategy for improving the retinal computer-aided diagnosis systems using unlabeled multimodal data. In particular, we explore the use of a multimodal reconstruction task between complementary retinal imaging modalities. This allows to take advantage of existent unlabeled multimodal data in the medical domain, improving the diagnosis of different ocular diseases with additional domain-specific knowledge that does not rely on manual annotation.

To validate and analyze the proposed approach, we performed several experiments aiming at the diagnosis of different diseases, including two of the most prevalent impairing ocular disorders: glaucoma and age-related macular degeneration. Additionally, the advantages of the proposed approach are clearly demonstrated in the comparisons that we perform against both the common fully-supervised approaches in the literature as well as current self-supervised alternatives for retinal computer-aided diagnosis. In general, the results show a satisfactory performance of our proposal, which improves existing alternatives by leveraging the unlabeled multimodal visual data that is commonly available in the medical field.

1. Introduction

Deep learning has become a fundamental part of modern computer-aided diagnosis (CAD) systems. The use of deep neural networks (DNNs) has improved the performance over traditional methods without requiring the ad-hoc design of complex processing algorithms. However, in return, DNNs need to be fed with expert knowledge in the form of large annotated datasets (Jing & Tian, 2020; Litjens et al., 2017).

Gathering enough annotated data for training a DNN can be challenging. In fact, in medical imaging, the manual labeling of the images should be performed by expert clinicians. This requirement commonly leads to a limited number of available annotated samples, given the time that takes to produce high-quality annotations (Tajbakhsh et al., 2016). This issue has motivated the adoption of numerous techniques aiming at the improvement of the deep learning methods without the necessity of additional annotated data (Cheplygina, de Bruijne, & Pluim, 2019). For instance, data augmentation techniques, which consist in creating new data samples through a set of plausible transformations for the application domain, are applied by default in order to successfully train DNNs (Bloice, Roth, & Holzinger, 2019; Litjens et al., 2017). Additionally, transfer learning techniques, which consist in taking advantage of already trained models for other applications, are also commonly employed to further improve the performance of the networks (Cheplygina et al., 2019; Litjens et al., 2017).

A common approach to transfer learning in image analysis is the use of DNNs that were pre-trained on extensive annotated datasets (Cheplygina et al., 2019; Houssein, Emam, Ali, & Suganthan, 2020). However, the available datasets of this kind are typically focused on broad

* Corresponding author at: Centro de Investigación CITIC, Universidade da Coruña, A Coruña, Spain.
E-mail addresses: a.suarezh@udc.es (Á.S. Hervella), jrouco@udc.es (J. Rouco), jnovo@udc.es (J. Novo), mortega@udc.es (M. Ortega).

https://doi.org/10.1016/j.eswa.2021.115598
Received 4 December 2020; Received in revised form 15 May 2021; Accepted 10 July 2021
Available online 24 July 2021

domain applications, such as the natural image classification challenge in the ImageNet dataset (Deng et al., 2009). It can be argued that the different nature of these images with respect to, for example, medical images, can represent a limiting factor for transfer learning purposes. In fact, when large scale annotated datasets of medical images are available, the performance benefit due to ImageNet pre-training is very limited (Raghu, Zhang, Kleinberg, & Bengio, 2019). However, in practice, the medical image datasets typically present a reduced number of annotations. In these scenarios, ImageNet classification pre-training has demonstrated to provide a general knowledge that improve the training of very deep convolutional networks in the medical imaging field (Cheplygina et al., 2019; Houssein et al., 2020).

Another alternative is to exploit the availability of heterogeneous or complementary labels within the same application domain, e.g. segmentation and classification labels. This allows to produce supervised auxiliary tasks that are restricted to the target application domain (Cheplygina et al., 2019). In this case, applying transfer or multi-task learning, the target task benefits from the increased amount of domain-specific knowledge. This auxiliary task alternative within the same domain has demonstrated to be superior than the use of additional data from broad domain natural images, even when the total number of involved annotations is lower (Wong, Syeda-Mahmood, & Moradi, 2018). The inconvenience in this case is that the additional labels in the target application domain carry an additional annotation effort.

Recently, self-supervised learning has arisen as a promising alternative to the traditional supervised approaches for transfer learning (Jing & Tian, 2020). Self-supervised learning is based on the use of pretext tasks that are trained with conventional supervised methods but do not require manual annotations. Instead, the training labels for these tasks are automatically generated from the unlabeled data. This allows the learning of useful representations for a target task using unlabeled data from the same application domain. Nowadays, the most common approaches to self-supervised learning are focused on either the prediction of hidden portions of the data or the prediction of hidden relations in the data (Jing & Tian, 2020). For instance, a representative self-supervised pretext task is colorization (Zhang, Isola, & Efros, 2016), which consists in the prediction of the different color components from the grayscale input image. Similarly, it is also possible to create an image inpainting task that requires the prediction of the original content of masked regions in the input image (Pathak, Krähenbühl, Donahue, Darrell, & Efros, 2016). Alternatively, solving jigsaw puzzles of the input image (Noroozi & Favaro, 2016) or predicting the geometric relationship between automatically extracted object proposals (Oh, Ham, Kim, Hilton, & Sohn, 2019) are representative examples of predicting the relations in the data. In this line, an emerging trend is the use of instance discrimination tasks (Chen, Kornblith, Norouzi, & Hinton, 2020; Ye, Zhang, Yuen, & Chang, 2019) that are performed via contrastive learning (Hadsell, Chopra, & LeCun, 2006). In these cases, the network learns to discriminate the individual images after being substantially altered via common data augmentations pipelines.

In medical imaging, given the difficulty for gathering large annotated datasets, there is an increasing interest for exploring these approaches. In particular, several works have adapted or extended existing paradigms previously proposed for natural images. For instance, Ross et al. (2018) propose colorization as auxiliary task for improving the segmentation of endoscopic video data. More recently, Taleb et al. (2020) extended several state-of-the-art self-supervised approaches to 3D medical data, including, e.g., jigsaw puzzles or instance discrimination with contrastive learning. Similarly, the instance discrimination paradigm was extended by including the synthesis of a complementary image modality as an additional transformation of the input image (Li, Jia, Islam, Yu, & Xing, 2020). Additionally, other novel self-supervised paradigms have been directly proposed in the medical imaging field. For instance, Chen et al. (2019) propose a context restoration task that requires to predict the original content of an image where different random patches are swapped. In contrast, Chaitanya, Erdil, Karani, and Konukoglu (2020) extended the contrastive learning paradigm to local features, producing a more adequate auxiliary task for segmentation.

Alternatively, instead of building the pretext task by manipulating the input image, in medical imaging it is also possible to directly use multimodal visual data for self-supervised learning purposes. In particular, Hervella, Rouco, Novo, and Ortega (2020a) propose the multimodal reconstruction between complementary image modalities as auxiliary task for segmentation or localization. The use of different imaging techniques is common in the modern clinical practice, including the use of complementary image modalities that represent the same organs or tissues. These complementary image modalities can be used to create a self-supervised multimodal reconstruction task consisting in the prediction of one image modality from other (Hervella, Rouco, Novo, & Ortega, 2020b). In order to solve this complex task, a neural network will have to learn relevant domain-specific patterns from the unlabeled data. Hence, the internal representations learned during this self-supervised task should be useful to improve the training of other target tasks in the same application domain.

However, self-supervision based on multimodal reconstruction of medical images has not yet been explored for improving deep learning CAD systems. In this regard, although Li et al. (2020) aim at using complementary modalities to aid self-supervision, their setting is not based on direct prediction. Instead, they explore the use of synthetic complementary image modalities as an additional augmentation strategy in a contrastive learning instance discrimination setting. In that case, it is expected that the learned representations are invariant to the synthetic multimodal transformation. However, that approach does not provide any incentive for the network to detect all the important patterns involved in the complex causal relations between modalities. In this sense, the network could just represent the patterns that are evident and similar in both modalities, disregarding the particular image contents that evidence the different complementary visualizations of the same reality. Additionally, due to the use of a synthetic image modality, the network only has access to a rough estimate of the true multimodal data. In contrast, the multimodal reconstruction task directly provides the network with the true multimodal data and the network must precisely learn the complex relationship between modalities to solve the task. Thus, the self-supervised multimodal reconstruction provides the network with a deep understanding of the image contents, which is expected to further facilitate the training of the desired deep learning CAD systems.

In this work, we propose to use the multimodal reconstruction between complementary retinal image modalities as self-supervised pre-training for deep learning-based retinal CAD systems. Specifically, we use the multimodal reconstruction between retinography and fluorescein angiography (Hervella et al., 2020b). Fig. 1 depicts a representative example of retinography and fluorescein angiography for the same eye. These two imaging modalities are obtained with different capture processes and represent complementary information about the different anatomical structures and pathological lesions in the retina. In particular, the retinography is directly obtained as a color photograph of the retina, whereas the angiography is captured after injecting a contrast dye into the patient's bloodstream. The proposed approach exploits this kind of existent unlabeled multimodal image pairs for learning useful representations of the data. This idea has been previously explored for improving pixel-wise prediction tasks, such as segmentation and localization, where the same neural network can be used for pre-training and target tasks (Hervella et al., 2020a). However, CAD systems require a completely different network architecture, which prevents the direct adoption of existing methodologies. In this regard, we provide a complete methodology for taking advantage of the multimodal reconstruction and improve the training of deep learning-based retinal CAD systems. In particular, this work focuses on the diagnosis of different retinal diseases, including two of the most prevalent impairing ocular disorders: glaucoma and age-related

(a) (b)

Fig. 1. Representative example of (a) retinography and (b) fluorescein angiography for the same eye.

macular degeneration (AMD). In this context, we perform several experiments that allow to better understand the proposed approach and we perform a comparison against two common fully-supervised approaches: training the network from scratch in the target task and pre-training in the annotated ImageNet dataset. Additionally, we also provide a comparison against previous self-supervised approaches in retinal image analysis.

2. Methodology

The proposed multimodal self-supervised transfer learning paradigm for the training of retinal CAD systems is summarized in the scheme of Fig. 2. The objective of the retinal CAD system is to predict the clinical diagnosis for a certain disease using the retinography of the patient as single input data. This target classification task is trained using an application-specific dataset containing annotated retinographies. In order to improve the performance of the target task and reduce the necessity of a large annotated dataset, we propose a domain-specific pre-training using unlabeled images. In particular, the pre-training task consists in the generation of fluorescein angiography from retinography. This multimodal reconstruction of the eye fundus is a self-supervised task that does not require manually annotated data for the training. Instead, it takes advantage of existent unlabeled multimodal image pairs. In order to take advantage of these image pairs, the retinography and the angiography of the same eye are aligned together. This establishes a pixel-wise correspondence between both images, resulting in a richer source of information in comparison with the unaligned counterparts.

Regarding the retinal CAD system, the study of different diseases typically requires the analysis of different regions in the retinal images. Thus, the Region Of Interest (ROI) for each disease is automatically extracted from the input retinography before feeding the image to the neural network. Simultaneously, the ROI for each disease is also extracted in the images of the unlabeled multimodal dataset. Thus, during the pre-training phase, the neural network will have to learn retinal patterns similar to those required for the target application. The proposed transfer learning paradigm is applied by fine-tuning, in the target classification task, the previously trained multimodal reconstruction network. In particular, a fully convolutional encoder–decoder network is used for the multimodal reconstruction. Then, the encoder part of the network is reused for the target classification task. In this regard, given the different network architecture requirements of both tasks, we explore different alternatives for performing an effective transfer learning between multimodal reconstruction and classification.

2.1. Deep learning-based retinal CAD

The automated diagnosis of AMD and glaucoma from the retinography is approached as a binary classification task. Therefore, for each disease, a neural network is trained to predict whether an input retinography is healthy or pathological. In this regard, in clinical practice, AMD is typically diagnosed by the presence of certain pathological structures or lesions around the macula, such as drusen, exudates, or epithelial abnormalities, among others (AREDS Research Group, 2001). In contrast, glaucoma is typically diagnosed after a detailed analysis of the optic disc morphology, including the optic cup and rim (Weinreb, Aung, & Medeiros, 2014). These clinical criteria are adopted by cropping squared ROIs centered at the macula and the optic disc for the cases of AMD and glaucoma, respectively. Following the clinical standards, the cropped regions present a size of four times the average optic disc diameter for AMD (AREDS Research Group, 2001) and two times the average optic disc diameter for glaucoma (Weinreb et al., 2014). The automated detection of the macula center and the optic disc is performed following the method proposed in Hervella et al. (2020a).

Representative examples of retinographies and the corresponding ROIs for AMD and glaucoma are depicted in Fig. 3. In the case of AMD, these examples show the great variety of pathological structures that may be present in this disease, ranging from very tiny lesions to larger structures that cover a substantial area in the macula. With regards to glaucoma, the examples show the common subtle differences between glaucomatous and non-glaucomatous eyes. In this case, the differences are typically focused on the internal optic disc morphology.

For each disease, the network training is performed using the binary cross-entropy (BCE) as loss function. Thus, the training loss for diagnosis is computed as:

$$\mathcal{L}_D = BCE(\mathbf{f}(\mathbf{r}), \mathbf{y}) \tag{1}$$

where $\mathbf{r}$ denotes the cropped retinography ROI, $\mathbf{y}$ its corresponding ground truth label, and $\mathbf{f}$ the transformation that assigns to each retinography $\mathbf{r}$ the likelihood of being a pathological sample.

2.2. Self-supervised multimodal pre-training

The multimodal reconstruction of fluorescein angiography from retinography is approached by using aligned retinography–angiography pairs as training data. The use of aligned image pairs results in a strong pixel-level supervision for learning the multimodal reconstruction task, as it allows the use of full-reference metrics between the network output and the aligned target image as loss function (Hervella, Rouco, Novo, & Ortega, 2018b). The alignment of the multimodal image pairs is automatically performed following the domain-specific methodology proposed in Hervella, Rouco, Novo, and Ortega (2018a).

The multimodal reconstruction pre-training is applied to the specific ROI of each target task. The ROI required for the analysis of each disease is extracted from both the retinography and the angiography following the same criteria and methods indicated for the target classification task (Section 2.1). In this way, during the pre-training, the

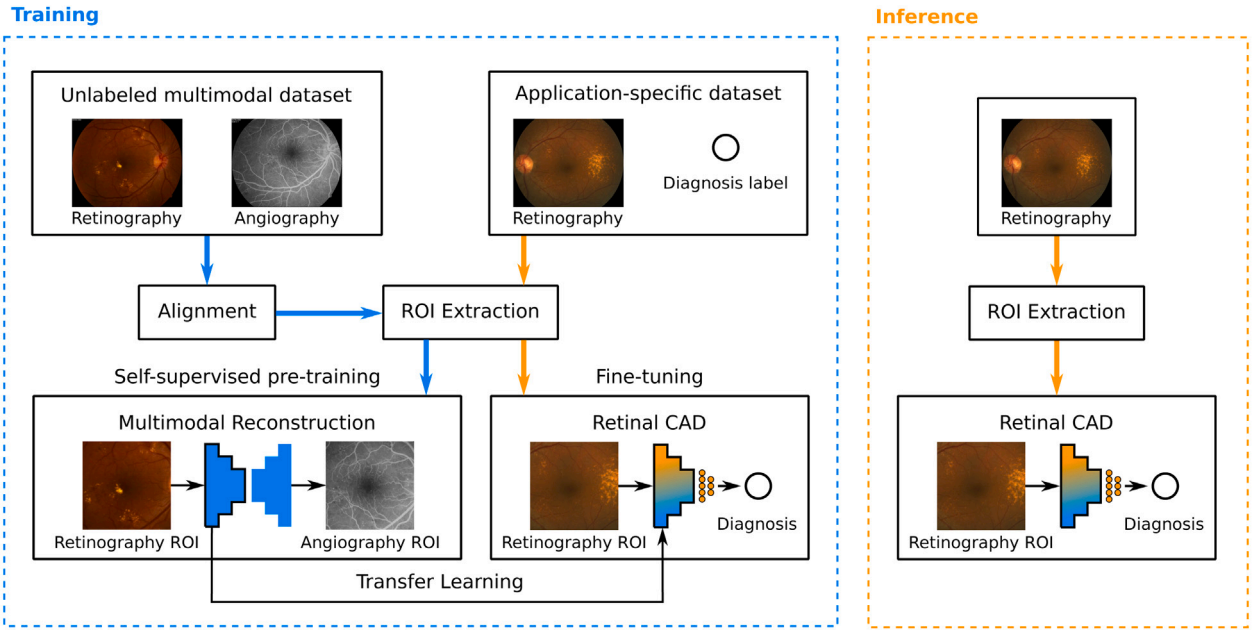

Fig. 2. Scheme of deep learning-based retinal CAD system using the proposed multimodal self-supervised pre-training.

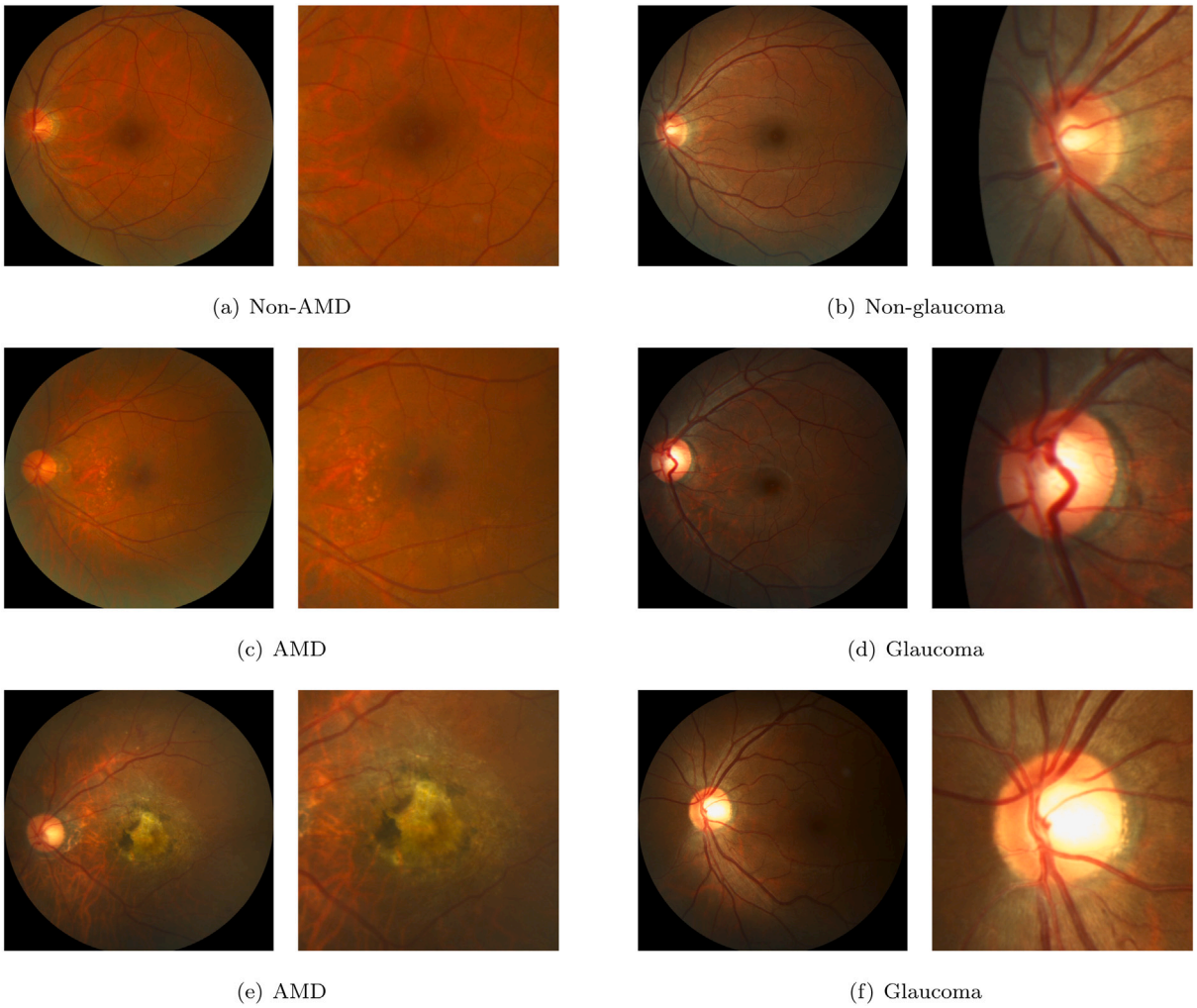

Fig. 3. Examples of retinographies and ROIs used for the diagnosis of ((a),(c),(e)) AMD and ((b),(d),(f)) glaucoma. For each image pair the retinography is in the left and the cropped ROI in the right.

neural network will learn to recognize those retinal structures that are relevant for the specific target application.

For each disease, the multimodal reconstruction pre-training is performed using the negative Structural Similarity (SSIM) as loss function. The use of SSIM has demonstrated to provide a superior performance for the multimodal reconstruction in comparison to other common metrics (Hervella et al., 2018b). SSIM is a similarity metric that takes into account intensity, contrast, and structural differences between the images. For that purpose, SSIM requires the computation of a series of local statistics at each pixel position, such as the mean and the variance in each individual image, and the covariance between the images. These statistics are computed locally considering a small neighborhood for each pixel. Then, given a pair of pixels (x, y), the SSIM value between x and y is computed as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1) + (2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2}$$

where μ_x and μ_y denote the local means of x and y respectively, σ_x^2 and σ_y^2 the local variances of x and y respectively, σ_{xy} the local covariance between x and y, and c_1 and c_2 are constant values used to avoid instability when the denominator terms are close to zero (Wang, Bovik, Sheikh, & Simoncelli, 2004). To avoid artifacts in the output, the local statistics are computed weighting the neighborhood of each pixel with a Gaussian window of $\sigma = 1.5$ (Wang et al., 2004).

Finally, the training loss for the multimodal reconstruction is computed as the negative mean SSIM between the network prediction and the target:

$$\mathcal{L}_{MR} = -\frac{1}{N}\sum_{n=1}^{N} SSIM(\mathbf{g}(\mathbf{r})_n, \mathbf{a}_n) \tag{3}$$

where $\mathbf{r}$ denotes the cropped retinography ROI, $\mathbf{a}$ the corresponding angiography ROI, $\mathbf{g}$ the transformation that maps each retinography to its angiography counterpart, and N the number of pixels in the ROI.

2.3. Network architecture

In order to demonstrate the advantages of the proposed approach and provide a reference well-proven baseline, we adopt standard network architectures for both target and pre-training tasks. In particular, we use VGG-Net (Simonyan & Zisserman, 2015) for the target classification tasks and U-Net (Ronneberger, Fischer, & Brox, 2015) for the multimodal reconstruction pre-training. Both VGG-Net and U-Net represent well-proven network architectures for image-level and pixel-level prediction tasks, respectively (Houssein et al., 2020; Tariq et al., 2020). Additionally, both networks share numerous characteristics due to the fact that the U-Net layers are precisely based on the design of VGG-Net. As consequence, these networks allow for a straightforward transfer learning strategy by directly reusing the pre-trained U-Net encoder as the encoder of the VGG-Net in the target classification task. Fig. 4 depicts a joint diagram of these networks that shows the close relationship between them.

Particularly, in this work, we use a VGG-Net with 13 layers (VGG-B) (Simonyan & Zisserman, 2015). This network consists of 10 convolutional layers followed by 3 fully connected layers. All the convolutions present kernels of size 3×3 and after every two convolutions there is a max pooling operation. In comparison with the 1000 classes of the ImageNet challenge (Deng et al., 2009), for which the network was originally designed, the classification tasks in this work only require the prediction of two classes: healthy or pathological. Thus, we adapt the network architecture by reducing the number of units in the 3 fully connected layers to 512, 128 and 1. A sigmoid activation function is used in the last layer to generate the binary prediction whereas the other layers have ReLU activation functions.

Regarding U-Net, this network architecture has already extensively demonstrated to be adequate for both the multimodal reconstruction (Hervella et al., 2020b) and transfer learning in this same application domain (Hervella et al., 2020a). In particular, U-Net is a fully convolutional network with a symmetric encoder–decoder structure and skip connections between encoder and decoder. These skip connections concatenate feature maps from the encoder with those of the same spatial resolution in the decoder. This particular design provides two main benefits. Firstly, precise spatial locations of the different extracted patterns are available in the decoder through the skip connections. This allows the precise generation of subtle details in the network output. Secondly, the skip connections ease the gradients back-propagation towards the early layers, which improves the network training.

The different convolutional blocks in U-Net are like those in VGG-Net. In particular, all the convolutions present kernels of size 3×3 and, in the encoder, there is a max pooling operation after every two convolutions. Similarly, in the decoder, there is a transpose convolution for upsampling every two convolutions. Then, the last layer consists of a convolution with kernel of size 1×1 and a linear activation function, whereas the other layers have ReLU activation functions.

Regarding the previously described transfer learning strategy, which consists in reusing the pre-trained U-Net encoder, we argue that it may be negatively affected by the skip connections in U-Net. In this sense, we should consider the effect of the skip connections in the high level representations learned by the encoder. Despite the positive effects in the network training, some relevant information related to the patterns that are learned in the early layers may never reach the network bottleneck (i.e., the encoder output), as they are directly forwarded to the decoder through the skip connections. In this case, the high level representations in the encoder will lack some information that may be relevant for the target classification task.

In this work, besides the standard U-Net architecture, we consider some variations of this network with a reduced number of skip connections for pre-training. The aim of this is to enforce that most of the relevant information reaches the network bottleneck during the multimodal reconstruction training. Additionally, this alternative keeps unchanged the classification network, which facilitates the comparison with other standard approaches for the target task. Initially, focusing on the requirements of the target classification task, it could be argued that avoiding all the skip connections would be the best alternative. However, this may excessively complicate the multimodal reconstruction training due to the difficulty of propagating precise spatial locations through the low resolution network bottleneck. In this regard, an inadequate pre-training could compromise the learning of representations that are useful for the target task. Thus, in order to study the most adequate configuration for transfer learning, we perform experiments with a varying numbers of skip connections, ranging from 0 to 4. In particular, in the performed experiments, the skip connections are added one at a time from the innermost to the outermost, as indicated in the numbering of Fig. 4.

2.4. Training details

Regarding the neural networks, the initial parameters are drawn from a zero-centered normal distribution following the approach proposed by He, Zhang, Ren, and Sun (2015). The optimization is performed using the Adam optimization algorithm (Kingma & Ba, 2015), with the default decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and batch size of one image. The initial learning rate is set to $\alpha = 1e{-}4$ for the multimodal reconstruction and $\alpha = 1e{-}5$ for the classification tasks. Additionally, we apply a learning rate schedule that reduces the learning rate by a factor of 10 when the validation loss does not improve for 25 epochs. Finally, the training is stopped after 100 epochs without improvement in the validation loss. These particular values are empirically set by taking as reference previous works in the literature and analyzing the learning curves during the training. In order to apply the described settings, 25% of the training data is used as validation subset. Additionally, in order to adapt the images to the input requirements of the classification network, the cropped ROI for each disease is rescaled to a size of 224×224 pixels.

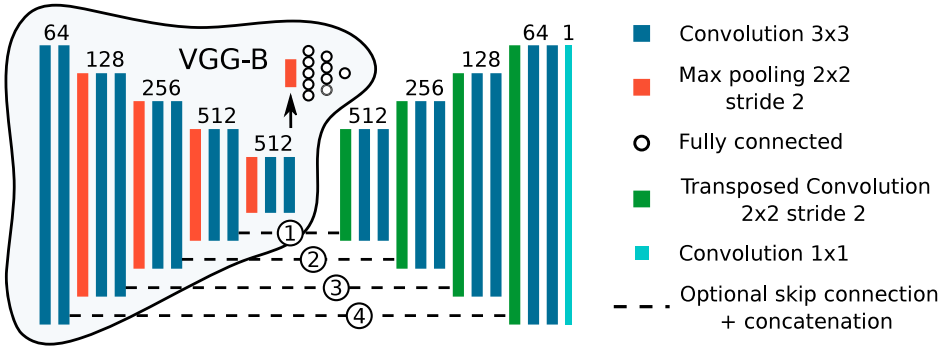

Fig. 4. Network architectures that are used in this work. The multimodal self-supervised pre-training and the target classification task are performed using U-Net and VGG-Net, respectively. Both network share the layers of the convolutional encoder.

To avoid overfitting in both pre-training and target tasks, we apply online data augmentation consisting of random spatial and color transformations. The spatial transformations are comprised of rotation, scaling, and shearing for the multimodal reconstruction and rotation and shearing for the classification tasks. The color transformations were applied in HSV color space as proposed in Hervella et al. (2020b). Additionally, the range of the transformation parameters was selected so that transformed images are still considered as valid in appearance. Finally, in order to take into account the stochasticity of the networks training, we perform 5 repetitions with different random seeds for each experiment in the target classification tasks.

3. Experiments and results

In order to validate the proposed approach we perform a set of experiments focused on three main aspects. First, we evaluate the effect of the U-Net skip connections in the proposed multimodal reconstruction pre-training. We use AMD and glaucoma diagnosis from retinographies as case study for this evaluation. Second, under this same AMD and glaucoma use cases, we compare the proposed self-supervised pre-training method with commonly used fully-supervised baseline approaches, based on initializing the diagnosis network with random weights, and using ImageNet classification pre-training. Finally, our work is compared with (Li et al., 2020) which, to the best of our knowledge, is the only related work in the literature using self-supervised approaches for retinal image analysis. Specifically. In order to provide comparable results, we follow the exact same experimental setting used in Li et al. (2020), to provide AMD and pathological myopia (PM) diagnosis from retinographies. In this case, the proposed multimodal self-supervised pre-training framework is directly applied without bells and whistles. The following sections provide the specific details and the obtained results for these three experimental settings.

3.1. Datasets

3.1.1. Multimodal reconstruction pre-training

For the proposed multimodal self-supervised pre-training, we use 59 retinography–angiography pairs from the public Isfahan MISP database (Alipour, Rabbani, & Akhlaghi, 2012). In this dataset, half of the images correspond to patients diagnosed with diabetic retinopathy, an eye condition that arises as a complication of diabetes (Rahim, Palade, Almakky, & Holzinger, 2019; Rahim, Palade, Jayne, Holzinger, & Shuttleworth, 2015). The other half of the images correspond to healthy individuals. All the images in this dataset are used for training/validation in the multimodal reconstruction.

3.1.2. Retinal CAD

For the diagnosis of glaucoma, we use 800 annotated retinographies from the public REFUGE dataset (Orlando et al., 2019). The prevalence of glaucoma in this dataset is 10%. The dataset includes a default split into two sets of 400 images each, named *Training* and *Validation*. In our experiments, we use the 400 images of *Training* as training data and the 400 images of *Validation* as hold-out test data.

For the diagnosis of AMD, we use 400 annotated retinographies from the public ADAM dataset (Fu et al., 2020). These images correspond to the *Training* split of this dataset, which is also used for the experiments in Li et al. (2020). The prevalence of AMD in this dataset is 23%. Similarly to glaucoma, we randomly split the dataset into two sets of 200 images with the same prevalence of AMD, one for training and the other as hold-out test data.

For the comparison with the state-of-the-art, we include an additional collection of images from the public PALM dataset (Fu et al., 2019). This dataset contains representative samples of retinas with pathological myopia (PM). In particular we use the 400 annotated retinographies from the *Training* split, which were also used in Li et al. (2020). The prevalence of PM in this dataset is 50%.

3.2. Evaluation

In order to quantitatively evaluate the proposed approach, the performance of the neural networks in the target classification tasks is evaluated using Receiver Operator Characteristic (ROC) analysis. This allows to directly evaluate the network predictions, which can be seen as the likelihood of the input samples being pathological, without the necessity of applying any specific decision threshold. In this way, we generate the ROC curves, which plot sensitivity and specificity for different decision thresholds. Additionally, we also compute the Area Under Curve (AUC) for ROC, which is commonly used to summarize the performance of the method into a single value.

3.3. Results

Fig. 5 depicts the results obtained for the diagnosis of AMD and glaucoma using the proposed multimodal self-supervised pre-training. The performance is evaluated by means of AUC-ROC for a varying number of skip connections in the multimodal reconstruction network. In these experiments, the skip connections are added one at a time, starting with the innermost layers and following the order described in Fig. 4. Additionally, for each number of skip connections, Fig. 5 also depicts the performance of the multimodal reconstruction, i.e., the pre-training task, by means of SSIM in the validation set. In order to better appreciate the differences between the considered alternatives, Fig. 6 depicts the complete ROC curves for each experiment in the target classification tasks.

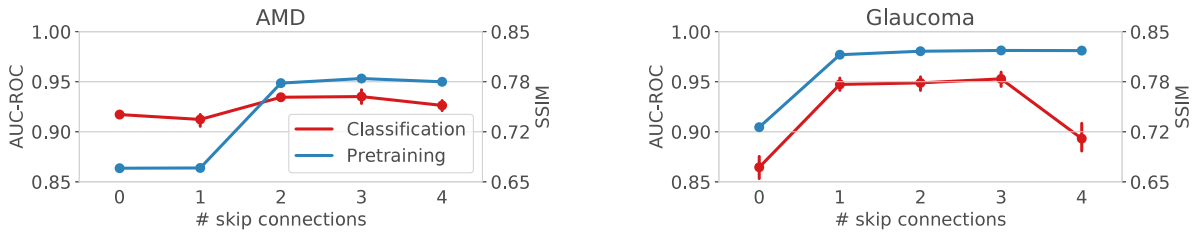

Fig. 5. Performance of the target classification tasks and their corresponding multimodal self-supervised pre-training for a varying number of skip connections. The depicted results for the classification task represent the mean value and standard deviation for 5 repetitions of the experiments.

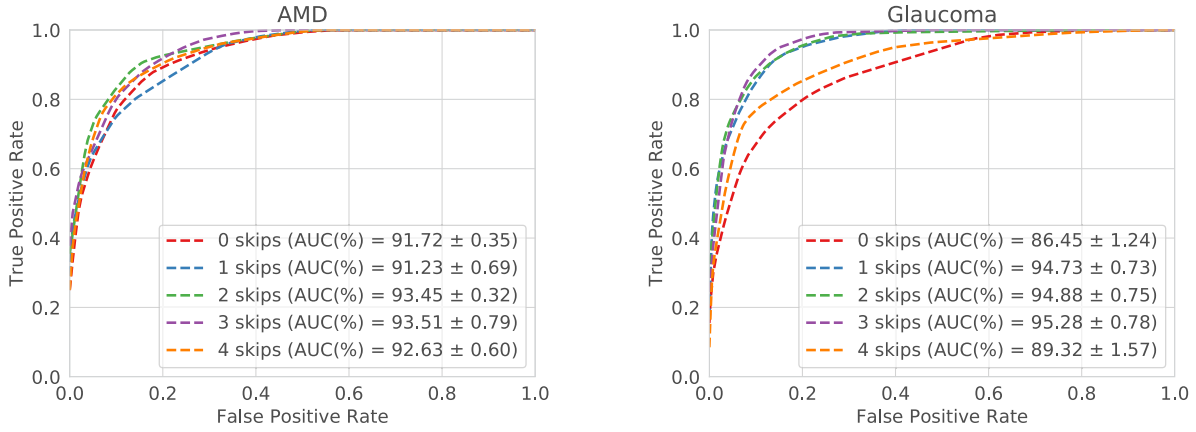

Fig. 6. ROC analysis of the target classification tasks for a varying number of skip connections. The results are obtained from 5 repetitions of the experiments and the depicted ROC curves represent the average performance for each case.

In general, the obtained results show that the proposed approach produces a satisfactory performance for both AMD and glaucoma classification. However, the performance is not equally satisfactory in all the experiments. The best results are achieved with an intermediate number of skip connections. Particularly, $\{2,3\}$ in the case of AMD and $\{1,2,3\}$ in the case of glaucoma. A lower number of skip connections results in a reduced performance for the target classification tasks and, also, for the multimodal reconstruction. The latter is expected due to the importance of the skip connections in the generation of detailed outputs at full resolution. In that sense, the lack of skip connections adversely affects the multimodal reconstruction, which, in turn, seems to compromise the learning of useful representations for the target classification task. Using all the skip connections also results in a reduction in the performance of the classification task, despite that the performance of the multimodal reconstruction is not significantly altered. These results fit with the idea that an extensive use of skip connections in the pre-training network may be detrimental for transfer learning purposes if only the pre-trained encoder is going to be reused.

To better understand the quantitative results, Fig. 7 depicts representative examples of generated angiographies for a varying number of skip connections. It can be observed that the quality of the generated angiographies fits perfectly well with the quantitative multimodal reconstruction results depicted in Fig. 5. In that sense, a minimum number of skip connections seems to be necessary to facilitate the adequate convergence of the multimodal reconstruction task. Nevertheless, it can be observed that, even in the worst cases, the network learns to recognize some important retinal structures and to generate a coarse representation of them. In particular, Fig. 7(e) and (f) show that the network recognizes the center of the macula, whereas Fig. 7(j) shows that the network broadly recognizes the optic disc. However, in these cases, many details such as the vasculature or lesions are missing.

Regarding the comparison between AMD and glaucoma, it can be observed in Figs. 5 and 6 that glaucoma classification is more sensitive to changes in the number of skip connections. It that sense, the classification of AMD keeps an adequate performance for all the experiments, even when no skip connections are used in the pre-training network. However, this is not the case for the classification of glaucoma, which experiments an important boost when the adequate pre-training settings are being used.

3.4. Comparison with fully-supervised approaches

To further analyze the advantages of the proposed approach, we perform a comparison against the most common alternatives in the literature, namely training the classification tasks from scratch and pre-training the networks on the annotated ImageNet dataset (Deng et al., 2009). Regarding the training from scratch, we use the initialization method proposed by He et al. (2015). In the case of the ImageNet pre-training, we use the pre-trained VGG-B network that is provided in the computer vision library of the PyTorch project (Paszke et al., 2017). It should be noticed that this network has been pre-trained in a fully-supervised fashion using more than a million annotated images. In contrast, the proposed multimodal self-supervised pre-training represents a novel alternative that only requires additional unlabeled data, and uses a dataset that is several orders of magnitude lower, counting with only 59 multimodal image pairs. In this regard, an important advantage of the proposed approach is that the size of the pre-training dataset could be increased without any human labeling effort.

Fig. 8 depicts the comparison of the proposed approach against training from scratch and ImageNet pre-training for both AMD and glaucoma diagnosis. This comparison is performed using the best empirical configuration for each of the methods. In particular, the results for the multimodal self-supervised pre-training correspond to the number of skip connections that provides the best performance for each disease (i.e., 2 skip connections for AMD and 3 skip connections for glaucoma). In the case of the ImageNet pre-training, it is common to apply a normalization scheme to the input images based on the statistics of the ImageNet dataset. In this work, we explored fine-tuning on the

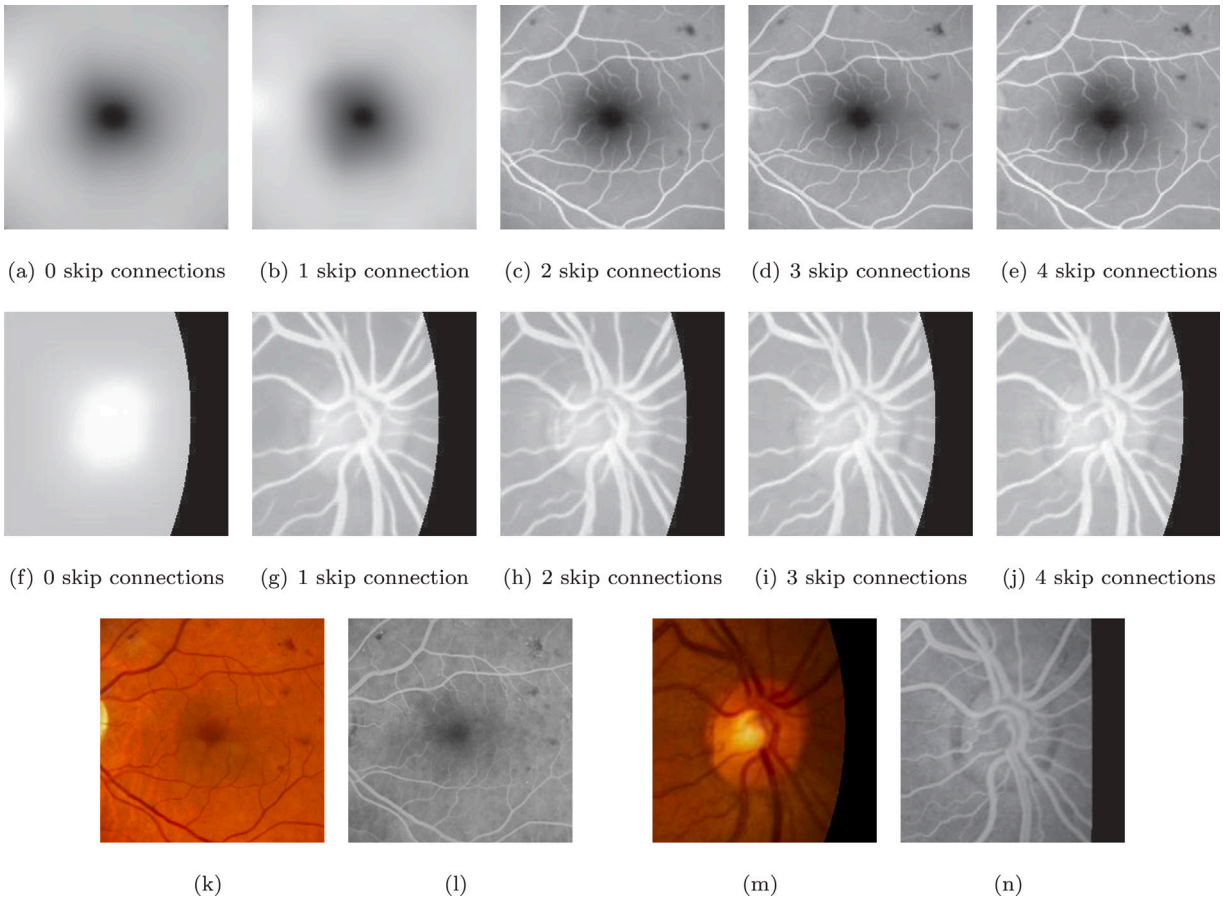

Fig. 7. Representative examples of generated angiographies for a varying number of skip connections. (a–e) Examples corresponding to the macula region (AMD pre-training). (f–j) Examples corresponding to the optic disc region (glaucoma pre-training). (k) Input retinography and (l) target angiography for (a–e). (m) Input retinography and (n) target angiography for (f–j).

application-specific datasets both with and without this normalization. The results presented in Fig. 8 correspond to the configuration that provides the best performance for each disease, which is the default ImageNet normalization for AMD and no normalization for glaucoma. Regarding the obtained results, it is observed that the proposed approach outperforms the training from scratch by a significant margin in both AMD and glaucoma diseases. This evidences that the patterns learned for the multimodal reconstruction are also useful for the detailed analysis of important retinal areas such as the macula or the optic disc. Thus, the obtained results demonstrate that the proposed approach is able to successfully take advantage of the unlabeled multimodal data for transfer learning purposes. Regarding the comparison with the ImageNet pre-training, the proposed self-supervised pre-training achieves a similar performance for the diagnosis of AMD despite not requiring any additional annotated data. Moreover, in the case of glaucoma, the proposed approach even outperforms the ImageNet pre-training by a significant margin. In this regard, it should be noticed that most of the self-supervised alternatives in the state-of-the-art are not able to equal the performance of the ImageNet pre-training (Jing & Tian, 2020). Considering this, the proposed approach offers a remarkable performance.

The results presented in Fig. 8 show that the ImageNet pre-training provides a satisfactory performance improvement for the diagnosis of AMD, while in the case of glaucoma the performance is slightly lower than that of the random initialization. These two ocular diseases require the analysis of different areas of the retina, namely the macula and the optic disc, involving features of different nature. Additionally, AMD is typically diagnosed by detecting the presence of certain local pathological lesions, whereas the diagnosis of glaucoma is typically performed by analyzing the morphology of the optic disc. Thus, a plausible explanation could be the recently demonstrated bias of the ImageNet pre-trained networks towards recognizing textures rather than shapes (Geirhos et al., 2019). In that sense, in our experiments, the ImageNet pre-trained network excels when the detection of subtle abnormalities is required (AMD) but falls behind when the morphological properties become more important (glaucoma).

In practice, we have also observed that the networks pre-trained on ImageNet present a significantly larger generalization gap in comparison to the other alternatives. This indicates that the networks' predictions tend to rely more on patterns that are specific to particular images, instead of those that are common to all the images of the same class (healthy or pathological). Thus, although ImageNet pre-training provides a very rich set of patterns that improves the network's training, this does not necessarily translate to a better performance in unseen images. This issue seems to be aggravated in the case of glaucoma due to the atypical morphological analysis that is required. Additionally, it should be noticed that the performance for the diagnosis of glaucoma also decreases when the proposed multimodal self-supervised pre-training does not properly converge due to the lack of skip connections in the network architecture (compare AUC-ROC values of Figs. 6 and 8). Therefore, in our experiments, it is clear that providing an adequate pre-training for the diagnosis of glaucoma is more challenging than doing the same for the diagnosis of AMD.

3.5. Comparison with state-of-the-art self-supervised approaches

In this section, we perform a comparison of the proposed approach against existing self-supervised approaches in the literature. In

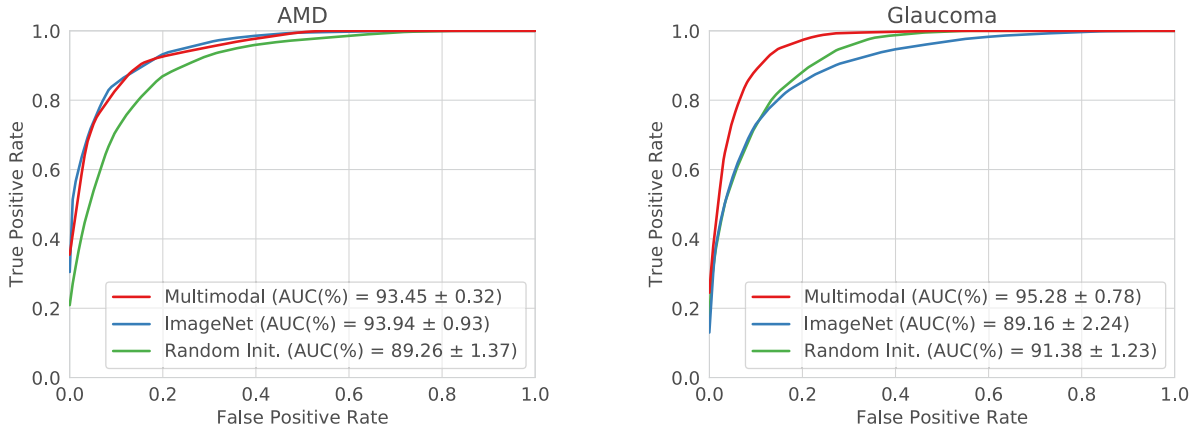

Fig. 8. ROC curves for the target classification tasks using the proposed multimodal self-supervised pre-training (Multimodal), ImageNet classification pre-training (ImageNet) and training from scratch (Random Init.). The results are obtained from 5 repetitions of the experiments and the depicted ROC curves represent the average performance for each alternative.

Table 1
Comparison with state-of-the-art self-supervised approaches for retinal computer-aided diagnosis.

Method	AMD	PM
	AUCROC (%)	AUCROC (%)
Li et al. (2020)	83.17	98.41
Proposed	89.57 ± 3.22	99.48 ± 0.58

particular, only one prior work has recently proposed an alternative self-supervised approach for retinal image analysis (Li et al., 2020). In order to adequately produce a fair comparison with this approach, we perform additional experiments using the same configuration that is adopted in Li et al. (2020). In particular, for these experiments, the whole images are used as input to the network. For this, the images are rescaled to a size of 224 × 224 pixels. The experiments are performed for the diagnosis of AMD and pathological myopia (PM) following a 5-fold cross-validation approach. In this case, we avoid to specifically tailor the methodology for each particular diseases and use the U-Net variant with 2 skip connections for all the experiments.

The comparison with the state-of-the-art is depicted in Table 1. In our case, we provide both the mean and standard deviation of the obtained results. It is observed that the proposed approach clearly outperforms the state-of-the-art alternative in both datasets. The difference in performance is greater for AMD, however, it is significant in both cases, especially considering the reduced standard deviation of our approach in PM. Moreover, besides the remarkable improvement in performance, our approach offers additional important advantages. Particularly, our proposal directly exploits the available paired multimodal data in a single pre-training step, whereas the method proposed in Li et al. (2020) requires two separate training stages and several neural networks in the pre-training phase. Thus, our proposal represents a more efficient and straightforward alternative. Moreover, to achieve the results reported in Li et al. (2020), the authors required to combine the multimodal synthesis augmentation with regular augmentation approaches used in broad domain self-supervised approaches (Ye et al., 2019), so the contribution of the multimodal-based self-supervision is not as clearly exploited in that approach. Instead the contribution of the multimodal information is clearly exploited in our method, without the need of complementary self-supervision tasks. Finally, the proposed approach, which is based on a pixel-level prediction (the multimodal reconstruction) allows the simultaneous pre-training of both image-level and pixel-level target tasks, such as, e.g., classification and segmentation. Thus, in contrast to the previous alternative, the proposed approach is also adequate for all kind of multi-task settings.

4. Conclusions

Nowadays, deep learning algorithms are commonly used in CAD systems. However, the performance of these methods is limited by the availability of sufficient annotated data. In order to mitigate this issue, we propose a self-supervised pre-training for deep learning-based retinal CAD systems consisting in the multimodal reconstruction between complementary imaging modalities. This approach exploits common existent unlabeled multimodal data in the medical domain for learning useful domain-specific representations.

The advantages of the proposed approach are mainly demonstrated in the context of two of the most prevalent impairing ocular diseases: AMD and glaucoma. We performed several experiments to analyze this novel transfer learning paradigm, including the study of important factors regarding the network architectures. In order to demonstrate the relevance of the proposed approach, we performed a comparison against two common fully-supervised approaches, namely training the network from scratch and pre-training on the annotated ImageNet dataset. Additionally, we also provide a comparison against existing self-supervised alternatives in retinal image analysis, including experiments in additional scenarios such as pathological myopia. The obtained results demonstrate that the proposed approach offers a satisfactory performance in all the pathological scenarios. Moreover, the multimodal reconstruction pre-training significantly outperforms both the training from scratch and the state-of-the-art alternatives, while it also demonstrates to be an overall superior approach to ImageNet pre-training.

Finally, given the excellent results that were obtained in all the pathological scenarios, in future work we plan to study the application of the proposed approach in other medical domains where multimodal visual data is also commonly available.

CRediT authorship contribution statement

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

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.

Acknowledgments

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; Ministerio de Ciencia e Innovación, Government of Spain, through the RTI2018-095894-B-I00 and PID2019-108435RB-I00 research projects; Xunta de Galicia and the European Social Fund (ESF) of the EU through the predoctoral grant contract ref. ED481A-2017/328; Consellería de Cultura, Educación e Universidade, Xunta de Galicia, through Grupos de Referencia Competitiva, grant ref. ED431C 2020/24. CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

References

Alipour, S. H. M., Rabbani, H., & Akhlaghi, M. R. (2012). Diabetic retinopathy grading by digital curvelet transform. *Computational and Mathematical Methods in Medicine*, *2012*, http://dx.doi.org/10.1155/2012/761901.

AREDS Research Group (2001). The age-related eye disease study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the age-related eye disease study report number 6. *American Journal of Ophthalmology*, *132*(5), 668–681. http://dx.doi.org/10.1016/S0002-9394(01)01218-1.

Bloice, M. D., Roth, P. M., & Holzinger, A. (2019). Biomedical image augmentation using augmentor. *Bioinformatics*, *35*(21), 4522–4524. http://dx.doi.org/10.1093/bioinformatics/btz259.

Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems (NIPS), Vol. 33*.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2019). Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, *58*, Article 101539. http://dx.doi.org/10.1016/j.media.2019.101539.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*.

Cheplygina, V., de Bruijne, M., & Pluim, J. P. (2019). Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, *54*, 280–296. http://dx.doi.org/10.1016/j.media.2019.03.009.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fu, H., Li, F., Orlando, J. I., Bogunović, H., Sun, X., Liao, J., et al. (2019). PALM: Pathologic myopia challenge. http://dx.doi.org/10.21227/55pk-8z03.

Fu, H., Li, F., Orlando, J. I., Bogunović, H., Sun, X., Liao, J., et al. (2020). ADAM: Automatic detection challenge on age-related macular degeneration. http://dx.doi.org/10.21227/dt4f-rt59.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping.. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *The IEEE International Conference on Computer Vision (ICCV)*.

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2018a). Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. *Procedia Computer Science*, *126*, 97–104. http://dx.doi.org/10.1016/j.procs.2018.07.213.

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

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2020a). Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction. *Applied Soft Computing*, Article 106210. http://dx.doi.org/10.1016/j.asoc.2020.106210.

Hervella, A. S., Rouco, J., Novo, J., & Ortega, M. (2020b). Self-supervised multimodal reconstruction of retinal images over paired datasets. *Expert Systems with Applications*, Article 113674. http://dx.doi.org/10.1016/j.eswa.2020.113674.

Houssein, E. H., Emam, M. M., Ali, A. A., & Suganthan, P. N. (2020). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*, Article 114161. http://dx.doi.org/10.1016/j.eswa.2020.114161.

Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. http://dx.doi.org/10.1109/TPAMI.2020.2992393.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Li, X., Jia, M., Islam, M. T., Yu, L., & Xing, L. (2020). Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 1. http://dx.doi.org/10.1109/TMI.2020.3008871.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. http://dx.doi.org/10.1016/j.media.2017.07.005.

Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*.

Oh, C., Ham, B., Kim, H., Hilton, A., & Sohn, K. (2019). OCEAN: Object-centric arranging network for self-supervised visual representations learning. *Expert Systems with Applications*, *125*, 281–292. http://dx.doi.org/10.1016/j.eswa.2019.01.073.

Orlando, J. I., et al. (2019). REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, Article 101570. http://dx.doi.org/10.1016/j.media.2019.101570.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., & Efros, A. (2016). Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems (NIPS), Vol. 32* (pp. 3347–3357).

Rahim, S. S., Palade, V., Almakky, I., & Holzinger, A. (2019). Detection of diabetic retinopathy and maculopathy in eye fundus images using deep learning and image augmentation. In *Machine Learning and Knowledge Extraction* (pp. 114–127). Cham: Springer International Publishing.

Rahim, S. S., Palade, V., Jayne, C., Holzinger, A., & Shuttleworth, J. (2015). Detection of diabetic retinopathy and maculopathy in eye fundus images using fuzzy image processing. In Y. Guo, K. Friston, F. Aldo, S. Hill, & H. Peng (Eds.), *Brain Informatics and Health* (pp. 379–388). Cham: Springer International Publishing.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., et al. (2018). Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery*, *13*(6), 925–933. http://dx.doi.org/10.1007/s11548-018-1772-0.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

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

Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., et al. (2020). 3D self-supervised methods for medical imaging. In *Advances in Neural Information Processing Systems (NIPS), Vol. 33*.

Tariq, M., Iqbal, S., Ayesha, H., Abbas, I., Ahmad, K. T., & Niazi, M. F. K. (2020). Medical image based breast cancer diagnosis: State of the art and future directions. *Expert Systems with Applications*, Article 114095. http://dx.doi.org/10.1016/j.eswa.2020.114095.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Weinreb, R. N., Aung, T., & Medeiros, F. A. (2014). The pathophysiology and treatment of glaucoma: a review.. *JAMA*, *311*(18), 1901–1911.

Wong, K. C., Syeda-Mahmood, T., & Moradi, M. (2018). Building medical image classifiers with very limited data using segmentation networks. *Medical Image Analysis*, *49*, 105–116. http://dx.doi.org/10.1016/j.media.2018.07.010.

Ye, M., Zhang, X., Yuen, P. C., & Chang, S.-F. (2019). Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *European Conference on Computer Vision (ECCV)*.