Computers in Biology and Medicine 143 (2022) 105302

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Multimodal image encoding pre-training for diabetic retinopathy grading

Álvaro S. Hervella [a,b,*], José Rouco [a,b], Jorge Novo [a,b], Marcos Ortega [a,b]

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

ARTICLE INFO

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

ABSTRACT

Diabetic retinopathy is an increasingly prevalent eye disorder that can lead to severe vision impairment. The severity grading of the disease using retinal images is key to provide an adequate treatment. However, in order to learn the diverse patterns and complex relations that are required for the grading, deep neural networks require very large annotated datasets that are not always available. This has been typically addressed by reusing networks that were pre-trained for natural image classification, hence relying on additional annotated data from a different domain. In contrast, we propose a novel pre-training approach that takes advantage of unlabeled multimodal visual data commonly available in ophthalmology.

The use of multimodal visual data for pre-training purposes has been previously explored by training a network in the prediction of one image modality from another. However, that approach does not ensure a broad understanding of the retinal images, given that the network may exclusively focus on the similarities between modalities while ignoring the differences. Thus, we propose a novel self-supervised pre-training that explicitly teaches the networks to learn the common characteristics between modalities as well as the characteristics that are exclusive to the input modality. This provides a complete comprehension of the input domain and facilitates the training of downstream tasks that require a broad understanding of the retinal images, such as the grading of diabetic retinopathy.

To validate and analyze the proposed approach, we performed an exhaustive experimentation on different public datasets. The transfer learning performance for the grading of diabetic retinopathy is evaluated under different settings while also comparing against previous state-of-the-art pre-training approaches. Additionally, a comparison against relevant state-of-the-art works for the detection and grading of diabetic retinopathy is also provided. The results show a satisfactory performance of the proposed approach, which outperforms previous pre-training alternatives in the grading of diabetic retinopathy.

1. Introduction

Diabetic Retinopathy (DR) is a complication of the diabetes affecting the retina, representing one of the leading causes of visual disability worldwide [1,2]. The early detection of the disease and the accurate grading of its severity are important steps towards providing the most adequate treatment and avoiding permanent vision loss [3,4]. However, the complex and diverse effects of this disease in the retina make the grading a very tedious and challenging task [5]. Simultaneously, the ever-increasing global prevalence of diabetes also makes the diagnosis of DR an important challenge in terms of health resources [2,6]. This is motivating the development of automated methods for the detection and severity grading of DR [7], which will facilitate the efficient screening of the population at risk [4].

In clinical practice, the detection and severity grading of DR is typically performed through the visual examination of the eye fundus using imaging techniques such as color retinography (or color fundus imaging) [3,8]. This retinal imaging modality is an affordable and widely available technique that allows the study of the retinal anatomy and the detection of pathological structures related to DR [8]. In this regard, the severity of DR is given by the presence of different lesions in the retina, being necessary to consider both the type and the number of lesions for the grading of DR [9]. Given the complexity of the disease, different severity scales have been clinically proposed and are being used in different countries [9]. For instance, Table 1 depicts the International Clinical Diabetic Retinopathy (ICDR) severity scale [10], which classifies DR into 5 different severity grades and is commonly used in recent literature on the grading of DR. This severity scale has been used

* Corresponding author. 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.compbiomed.2022.105302
Received 30 September 2021; Received in revised form 11 January 2022; Accepted 26 January 2022
Available online 16 February 2022

Table 1
International Clinical Diabetic Retinopathy disease severity scale for the grading of DR [10]. NPDR denotes Non Proliferative DR whereas PDR denotes Proliferative DR.

Grade	Clinical findings on color retinography
0 - No apparent DR	No abnormalities
1 - Mild NPDR	Microaneurysms only
2 - Moderate NPDR	More than just microaneurysms but less than severe NPDR
3 - Severe NPDR	Any of the following but no signs of PDR: ● More than 20 intraretinal hemorrhages in each of four quadrants ● Definite venous beading in two or more quadrants ● Prominent intraretinal microvascular abnormalities in one or more quadrants
4 - PDR	One or more of the following: ● Neovascularization ● Vitreous/preretinal hemorrhage

for the grading of publicly available databases such as IDRiD [11], DDR [12], or Messidor-2 [5]. In contrast, Table 2 depicts an alternative scale that classifies DR into 4 different severity grades and was used for the grading of the well known Messidor database [13]. Fig. 1 depicts representative examples of images with different grades of DR according to the severity scale of Table 1. These examples include color retinography images of both left and right eyes.

The diversity of lesions and the complex relations among the different severity grades make the grading of DR especially challenging. In fact, it has been shown that even clinical specialists tend to disagree in the most complex cases [5]. Thus, automated grading methods can also be useful by helping the clinicians to provide a more reliable diagnosis. In that regard, numerous deep learning approaches have been proposed for addressing the detection of DR as a binary classification, i.e. grouping the different grades into only two different classes (typically 0, 1 vs 2,3,4) [9,14]. However, the complete grading of DR represents a more challenging and less explored objective, for which there is an increasing interest [9].

The common approach for the automated grading of DR from color retinography is the use of Deep Neural Networks (DNNs) [7,9]. Previous works have been typically focused on improving aspects such as the network architecture (e.g. using attention mechanisms [15]) or the formulation of the training objective (e.g. using novel regularization schemes [16]). However, all the works follow the same approach to alleviate the scarcity of annotated data, particularly using neural networks previously pre-trained on an additional annotated dataset of natural images (i.e. the ImageNet [17] dataset) [9].

The limited amount of annotated data that is usually available in medical imaging is a long-standing issue for the application of DNNs in the field [9,18]. In this context, the use of neural networks pre-trained on the ImageNet dataset has been the go-to approach during several years [9,19]. However, while ImageNet pre-training allows to achieve successful results in numerous applications, it is still a fully-supervised approach that relies on the availability of large amounts of annotated natural images. Thus, this approach does not really solve the fundamental issue of the dependency of DNNs on large amounts of annotated data. Instead, it only replaces the lacking annotations in one application domain by additional annotated data from another.

Table 2
Severity scale for the grading of DR on the publicly available Messidor database [10].

Grade	Clinical findings on color retinography
0 - No apparent DR	No abnormalities
1 - Mild DR	5 or less microaneurysms but no hemorrhages
2 - Moderate DR	Any of the followings but no neovascularization: ● Between 6 and 14 microaneurysms ● 4 or less hemorrhages
3 - Severe DR	Any of the followings: ● Neovascularization ● 15 or more microaneurysms ● 5 or more hemorrhages

Recently, self-supervised learning has arisen as a promising alternative to the traditional supervised pre-training approaches [20]. Self-supervised learning is based on the use of pretext tasks for which the supervisory signals are obtained without involving manual annotations. In this case, the networks are typically trained in the prediction of hidden portions of the data or hidden relations among different data samples. For instance, a DNN can learn about the input domain in a self-supervised fashion by restoring noisy samples [21], colorizing gray-scale samples [22], or performing instance discrimination via contrastive learning [23]. An advantage of these approaches is that the pre-training can be easily performed on the same application domain of the final target task, given that no manual annotations are required. Thus, these approaches can potentially provide more useful high level representations than traditional supervised alternatives that are performed in natural images.

A particular free source of supervision that can be found in medical imaging is the existence of complementary imaging modalities, i.e. different imaging techniques that depict complementary visualizations of the same organs or tissues [24]. In modern clinical practice, it is common the use of complementary imaging techniques for the assessment of the most complex cases, which eases the gathering of multimodal collections of medical images (i.e. collections of multimodal visual data) [25,26]. Traditionally, these multimodal visual data have only been used when labels for the images are also available, e.g. developing algorithms that make their prediction based on a multimodal input [27,28]. However, the differences and similarities among complementary modalities represent a potentially rich source of supervision in itself, which can be taken advantage of for representation and transfer learning purposes. In this regard, the prediction of a target image modality from another input modality has been explored in recent works [24,29]. In this case, the aim is the pre-training of a neural network for different downstream tasks performed on the same input modality [30,31]. However, while the supervision provided by this multimodal prediction explicitly teaches the network to recognize the similarities between the modalities, there is no explicit incentive to recognize the differences between them. Thus, some relevant knowledge about the input image modality may be lacking in the pre-trained networks.

In general, downstream tasks that require a broad understanding of the retinal images, such as the grading of DR, would benefit from pre-training approaches that ensure a complete comprehension of the input modality. To that aim, we propose a novel self-supervised multimodal pre-training approach that explicitly teaches the network to recognize the common characteristics between modalities as well as the characteristics exclusive to the input modality. In this way, the pre-trained networks will have a complete comprehension of the input domain modality, including all the anatomical and pathological structures that are present in the images.

The proposed pre-training approach, denoted as Multimodal Image Encoding (MIE), is applied on top of a standard convolutional encoder commonly used for image classification. The learning of the common and exclusive characteristics regarding the input modality is ensured by providing two complementary supervisory signals for the training of the network, one due to the multimodal prediction of the target modality and another due to the reconstruction of the input modality. Likewise, the learning of rich representations from these two supervisory signals is ensured by a proposed network design that facilitates the disentanglement of the common and exclusive features at the output of the convolutional encoder being pre-trained.

In this work, MIE is applied as self-supervised pre-training for the grading of DR from color retinography. For that purpose, color retinography is used as input modality and fluorescein angiography [25] as target modality during the training of MIE. Retinography-angiography

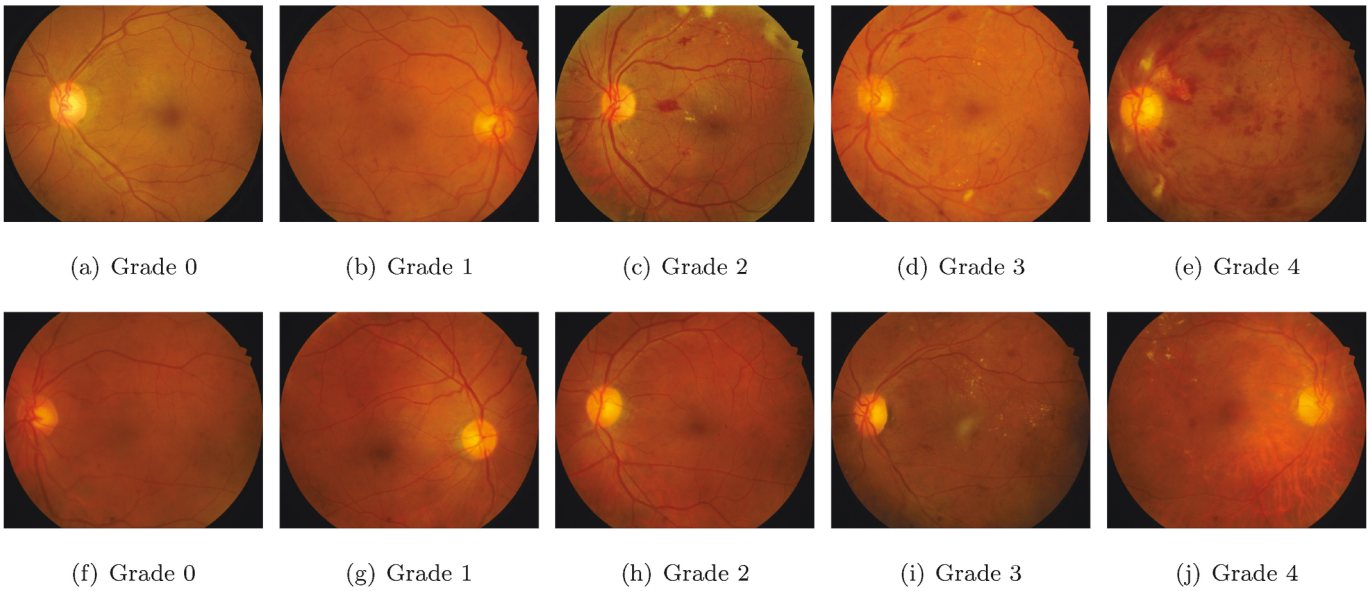

Fig. 1. Examples of color retinography images for different grades of DR according to the severity scale depicted in Table 1 [10]. These examples are taken from the test set of the public IDRiD dataset [11] and include both left and right eyes. In particular, ((a),(c),(d),(e),(f),(h),(i)) represent left eyes whereas ((b),(g),(j)) represent right eyes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

pairs of unlabeled images can be easily obtained due to the common use of this multimodal pair in clinical practice for decades [25] worldwide. Nevertheless, after the self-supervised multimodal pre-training, the networks only require the input retinography for the subsequent grading of DR. The main contributions of this work as summarized as follow:

- A novel self-supervised pre-training approach for learning from multimodal visual data in medical imaging is proposed. To the best of our knowledge, this is the first work that proposes a novel self-supervised alternative to ImageNet pre-training for the grading of DR.
- The proposed network design disentangles the common and exclusive features regarding the input modality and allows the learning of rich representations from the unlabeled multimodal visual data. Additionally, the proposed approach can be applied on top of a standard convolutional encoder for image classification.
- The transfer learning performance for the detection and grading of DR is evaluated on several datasets and under different experimental settings. All the experiments are performed on public datasets, for both the pre-training (Isfahan MISP [26]) and the grading of DR (Messidor [13], IDRiD [11], EyePACS-Kaggle [32]).
- An exhaustive comparison against different alternatives, including ImageNet pre-training and the previous state-of-the-art multimodal approach is performed. Additionally, a comparison against relevant state-of-the-art works for the detection and grading of DR is also provided.

1.1. State-of-the-art for DR grading

In this section, we provide a more detailed discussion about relevant works on the detection or grading of DR. The automated diagnosis of DR has been previously addressed in numerous works using classical approaches for image processing and analysis [7,33]. In that regard, a classical pipeline for the automated diagnosis of DR usually contains the following steps: image pre-processing, extraction of anatomical structures, extraction of relevant features, and feature-based image classification [7,33].

The objective of the pre-processing stage [7] is to facilitate the subsequent extraction of anatomical structures and relevant features for the diagnosis. For that purpose, illumination correction [34], contrast enhancement [34,35], and noise or artifact reduction [35] have been commonly performed [7,34,36]. Meanwhile, the subsequent extraction of anatomical structures is usually performed with two different objectives. Firstly, this can be used as an intermediate step to facilitate the extraction of other anatomical structures or relevant lesions [7,34,37]. For instance, the localization or segmentation of the optic disc is typically used as an intermediate step in the detection of bright lesions such as exudates [34], whereas the segmentation of the blood vessels is useful for the detection of red lesions such as hemorrhages or microaneurysms [37]. Secondly, some anatomical structures can be directly used for the extraction of anatomical features that are relevant for the diagnosis of DR, such as vessel areas [33,35] or vessel bifurcation points [34,35].

Besides the vascular features, the main clinical characteristics that are useful for the classification of the images are those related to the presence of lesions such as microaneurysms, hemorrhages, or exudates [33,37]. In this regard, the detection of these lesions can be directly used for the classification of the images using clinical severity scales (such as the ones depicted in Tables 1 and 2) as a fixed set of rules for the grading [37]. Nevertheless, the presence and size of the lesions can be also used as individual features for the later classification of the images using machine learning techniques [34,35]. Additionally, non-clinical features have also been used in several works, including e.g. texture features or entropy measures [34].

With regards to the feature-based classification stage, different machine learning techniques have been used, such as Support Vector Machines (SVM) [7,34,35], Decision Trees [34], Neural Networks [7,34], or Random Forests [7,33]. In that regard, currently, DNNs represent the most common approach for the detection and grading of DR. In contrast to previous machine learning alternatives, DNNs avoid the use ad-hoc image features and allow to directly classify the images in a single step. However, the use of pre-processing techniques to facilitate the recognition of the main characteristics in the images is still widely extended [9].

Regarding the use of DNNs, standard classification networks, such as VGG [38], ResNet [39], or EfficientNet [40] are commonly used [9,41]. However, some specific variations of these networks have also been proposed. For instance, Wu et al. [42] propose a dual design consisting of coarse and fine subnetworks that perform binary and multi-class classification, respectively. Other common variations are the addition

of attention layers [42] or novel attention blocks that can be placed on top of standard networks [15,43]. These blocks are usually inspired by the Convolutional Block Attention Modules [44] that perform channel and spatial-wise attention. In particular, He et al. [15] propose to perform attention over groups of features that aim at representing the different categories that must be predicted. Also, in the context of the joint diagnosis of DR and Diabetic Macular Edema, Li et al. [43] propose to perform cross-disease attention over diseases-specific features.

Regarding the training objective, the most common approach is the use of cross-entropy loss for multi-class classification [9]. However, this area has also been the focus of several proposals, particularly taking advantage of the fact that the classes can be ordered by their severity. In this regard, de la Torre et al. [45] propose the use of Quadratic Weighted Kappa, which is commonly used for the evaluation, as loss function. This loss assigns a greater penalty when the predicted class is more distant from the target label. In contrast, Araújo et al. [46] take advantage of the order among classes to model each individual prediction as a Gaussian distribution, hence also giving an estimate of the uncertainty in the form of the predicted variance. Differently, Galdran et al. [16] propose a regularization approach that allows to set the desired cost associated to each possible error in a confusion matrix. Also for regularization purposes, Tu et al. [47] propose the prediction of lesion maps at a reduced scale together with the grading.

Regarding the initialization or pre-training of the networks, all the previous works rely on the classical fully-supervised ImageNet pre-training [9]. Thus, our work is the first to explore a novel self-supervised alternative using unlabeled data from the same domain. In particular, our proposal takes advantage of unlabeled multimodal image pairs that are common in ophthalmology [25,26].

The remainder of the manuscript is structured as follows. The materials and methods that are used in this work are described in Section 2. This section includes (Section 2.1) the methodology for the proposed pre-training as well as (Section 2.2) the methodology for the grading of DR. The experiments and results are presented in Section 3. A discussion of the obtained results is provided in Section 4 and, finally, conclusions are drawn in Section 5.

2. Materials and methods

This work presents a novel self-supervised pre-training as well as a complete methodology for the grading of DR in color retinography using DNNs. The proposed methodology for the grading of DR can be split into two different parts, which are depicted in the diagram of Fig. 2.First, a neural network is pre-trained using the proposed MIE pre-training using unlabeled multimodal retinal images. In this regard, the proposed pre-training uses unlabeled image pairs consisting of fluorescein angiography and color retinography, hence it does not requires any labeled data. Then, the pre-trained neural network is fine-tuned for the grading of DR. This fine-tuning is performed using labeled color retinography images and standard supervised approaches. Finally, after the pre-training and fine-tuning phases, the neural network is able to make a prediction on the grading of DR using only a color retinography image as input.

2.1. Multimodal image encoding

In contrast to previous approaches that explored the use of multimodal images for self-supervised learning [24], MIE explicitly aims at learning both the features that are common between modalities as well as the features that are exclusive to the input image modality. This results in a large variety of learned patterns that completely describe the desired image modality and can be taken advantage of for any downstream tasks performed on the same modality.

In medical imaging, it is typical to use complementary image modalities that not only change the appearance of the different organs and tissues, but also allow the visualization of completely different structures or properties of the tissues. This is typically achieved by the use of either completely different image acquisition technologies or injected contrasts dyes [25]. Thus, considering two complementary image modalities of this kind, $\mathcal{A}$ and $\mathcal{B}$, it is expected that, while most of the content is common between modalities, some relevant structures may only be appreciated in one of the modalities, e.g. $\mathcal{A}$. In this case, it could be possible to train a neural network in the prediction of modality $\mathcal{B}$ from modality $\mathcal{A}$, such as in Ref. [24]. However, given that some relevant structures or properties of the input modality $\mathcal{A}$ are missing on the target modality $\mathcal{B}$, the network may ignore these characteristics exclusive to $\mathcal{A}$ and focus on learning just the characteristics that are common between modalities. Thus, this kind of approach can potentially discard valuable patterns useful for representation learning purposes and transfer learning towards different downstream tasks.

A conceptual diagram of the proposed pre-training approach is depicted in Fig. 3. The aim of MIE is to ensure that the network learns all the characteristics of the input modality $\mathcal{A}$ during the pre-training phase. For that purpose, MIE defines a new domain $\mathcal{Z}$ that aims at encoding the characteristics of modality $\mathcal{A}$ that are missing in modality $\mathcal{B}$. Then, given any input image $x_{\mathcal{A}} \in \mathcal{A}$, the network learns to generate two complementary representations, $x_{\mathcal{B}} \in \mathcal{B}$ and $x_{\mathcal{Z}} \in \mathcal{Z}$, which together provide a complete representation of the input image contents. In particular, $x_{\mathcal{B}}$ is the predicted representation of the input image in the target modality $\mathcal{B}$.

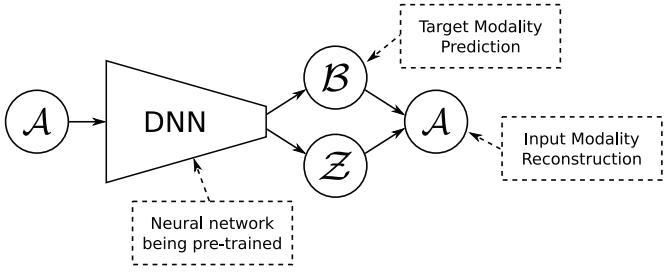

Fig. 3. Conceptual diagram of the proposed MIE pre-training. $\mathcal{A}$ represents the input modality, $\mathcal{B}$ represents the target modality, and $\mathcal{Z}$ represents a new domain containing the characteristics of $\mathcal{A}$ that do not belong to $\mathcal{B}$.

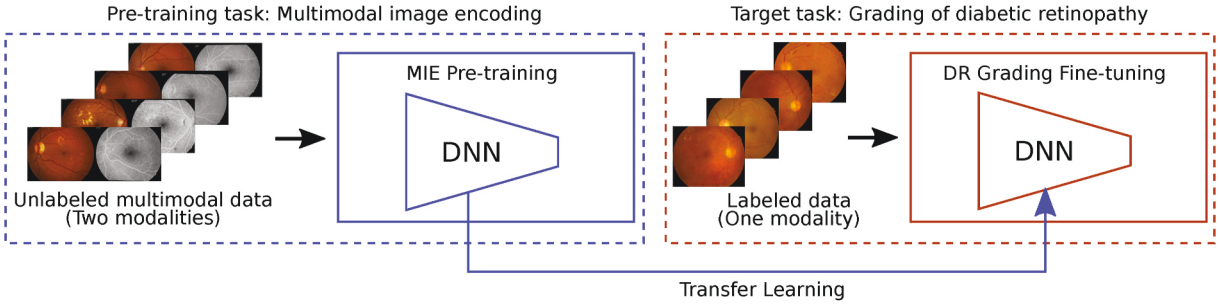

Fig. 2. Diagram of the proposed methodology for the grading of DR.

Meanwhile, $x_{\mathcal{Z}}$ is a complementary representation that encodes the characteristics of the input image that cannot be represented in $x_{\mathcal{B}}$ because they do not belong to the target modality $\mathcal{B}$. In order to provide the training feedback for simultaneously learning the representations $x_{\mathcal{B}}$ and $x_{\mathcal{Z}}$, MIE presents two complementary training objectives. One is the prediction of the target image modality from the input image $x_{\mathcal{A}}$. The other is the reconstruction of the input modality from the intermediate representations $x_{\mathcal{B}}$ and $x_{\mathcal{Z}}$. In this work, the objective is to facilitate the grading of DR from retinography using DNNs, hence retinography is used as the input modality $\mathcal{A}$. For the target modality $\mathcal{B}$, we use fluorescein angiography, an alternative modality that requires the injection of a contrast dye to the patients [25]. The use of this contrast dye provides additional information about the retinal vasculature and related lesions, which is useful for the diagnosis of DR [48]. However, as a counterpart, some other structures or characteristics are harder to appreciate in this modality. Additionally, fluorescein angiography presents the drawback of being an invasive image modality with potential side effects due to the required contrast dye. However, in this work, fluorescein angiography is only used in the pre-training phase, taking advantage of publicly available multimodal image collections. After the pre-training phase, the neural networks are fine-tuned in the grading of DR using only color retinography. The main visual differences between retinography and angiography are depicted in Fig. 4. Additionally, it can be seen in this example that there is a relative displacement of the retinal structures between the two image modalities. This is because retinography and fluorescein angiography are not captured at the same time, hence there is movement of the patient between the capture of one image and the other. In order to fully take advantage of the unlabeled multimodal data, MIE uses retinography and fluorescein angiography image pairs that are automatically aligned with the methodology proposed in Ref. [49]. The aligned image pairs allow the use of pixel-wise metrics as loss function for both the target modality prediction and the input modality reconstruction. In this regard, the use of pixel-level training feedback is an advantage that reduces the necessity of training data in comparison to the image-level counterpart [29].

2.1.1. Network architecture

In order to ensure the learning of relevant image patterns and avoid trivial solutions, we propose a specific network design for MIE that is depicted in Fig. 5.

In this regard, MIE can be split into two main transformations, $F: \mathcal{A} \rightarrow \mathcal{B}, \mathcal{Z}$ and $G: \mathcal{B}, \mathcal{Z} \rightarrow \mathcal{A}$. The former addresses the simultaneous generation of the target modality $\mathcal{B}$ and the complementary information $\mathcal{Z}$, whereas the latter addresses the reconstruction of the input modality $\mathcal{A}$ from the combination of $\mathcal{B}$ and $\mathcal{Z}$. In this case, we use an encoder-decoder architecture with skip connections for both F and G. Thus, in total, the proposed network design presents 4 distinct components: the encoder F_E, the decoder F_D, the encoder G_E, and the decoder G_D. Among these components, F_E represents the classification encoder that is pre-trained and will be later fine-tuned for the grading of DR. Meanwhile, F_D, G_E, and G_D are auxiliary components necessary for performing the proposed pre-training approach. During the MIE pre-training, the encoder F_E learns to recognize relevant patterns from the input image modality. In order to learn a rich set of features that completely describe the input image contents, the output of the encoder is connected to two branches. The first branch connects to the decoder F_D that generates the prediction $\widehat{x}_{\mathcal{B}}$ in the target image modality space. This branch is used as a means of learning the common features between modalities. In contrast, the second branch encodes the features that are exclusive to the input image modality, which are necessary for the complete reconstruction of the input image. This reconstruction is performed by the encoder G_E and the decoder G_D that merge back together the common and exclusive features. In particular, both the encoded features $\widehat{x}_{\mathcal{Z}}$ and the high level representation extracted by G_E are fed to the decoder G_D that generates the reconstructed image $\widehat{x}_{\mathcal{A}}$. In summary, in the proposed network design the output of the encoder F_E is connected to two different branches, producing two possible paths for the forward flow of information through the network:

- 1st branch: the information flows through the path $F_E \rightarrow F_D \rightarrow G_E \rightarrow G_D$. The aim of this branch is to learn the features that are common between modalities.
- 2nd branch: the information flows through the path $F_E \rightarrow G_D$. The aim of this branch is to learn the features that are exclusive to the input modality.

Despite the differences between retinography and angiography, the general structure of the images is shared between modalities. Thus, the common features must encode an important part of the image contents together with their precise spatial distribution. To facilitate the generation of structurally accurate predictions at full resolution, $\widehat{x}_{\mathcal{A}}$ and $\widehat{x}_{\mathcal{B}}$, skip connections are used for the first network branch containing the common features. This significantly reduces the information bottleneck between each encoder and decoder. However, in this case, the network is forced to learn relevant patterns from the data due to the inherent complexity of the required multimodal transformation. In the case of the second branch containing the exclusive features, the learning of relevant features is directly forced by the spatial bottleneck between F_E and G_D. Thus, in this case, no skip connections are used between F_E and G_D. In this regard, our experiments empirically demonstrate that the proposed design provides and adequate bottleneck that is enough to avoid an identity mapping through the second branch of the network. Thus, the features exclusive to the input image modality are successfully learned by the network. Additionally, to ensure that this narrower branch is not ignored during the training, a gradient block operation is introduced between F_D and G_E. This operation allows the forward flow of information from F_D to G_E but blocks the backward flow of training feedback. This avoids conflicting feedback through F_D, ensuring that the input image is not bypassed through the skip connections.

Regarding the characteristics of the different encoders and decoders, we adopt the common VGG [19,38] design for all of them. The particular layers of each encoder and decoder are depicted in the diagram of Fig. 6.

In particular, the convolutions present 3×3 kernels and ReLU activation functions, the downsampling is performed with strided max pooling operations, and the upsampling with transpose convolutions. In this case, we adopt a standard VGG-B [38] convolutional encoder for the main encoder in the network, F_E, which will be reused for the grading of DR. To ensure that most of the relevant features are learned in this encoder, the other parts of the network present a relatively lower capacity. Particularly, while we keep the same number of downsampling and upsampling steps, both the number of layers and number of channels is reduced by half in F_D, G_E, and G_D.

2.1.2. Network training

As depicted in Fig. 5, during the training the network receives complementary feedback from two different sources. One is the prediction of the target image modality and the other is the reconstruction of the input modality. Regarding the multimodal prediction, we adopt the approach proposed in Ref. [24] using the negative Structural Similarity (SSIM) as loss function. Particularly, SSIM [50] is a similarity metric that considers the intensity, contrast, and structural differences between images. The SSIM value for a pair of pixels (x, y) is obtained using a set local statistics as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1) + (2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{1}$$

where μ_x and μ_y are the local averages for x and y, respectively, σ_x and σ_y are the local standard deviations for x and y, respectively, and σ_{xy} is the local covariance between x and y. The local statistics are computed for each pixel by weighting its neighborhood with a Gaussian window of σ

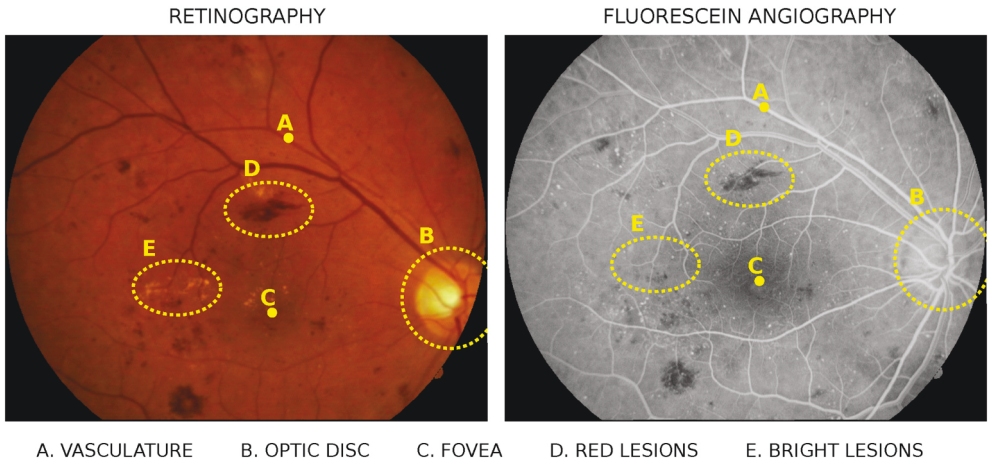

Fig. 4. Representative example of color retinography and fluorescein angiography for the same eye. The main anatomical structures in the retina and relevant lesions are highlighted. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

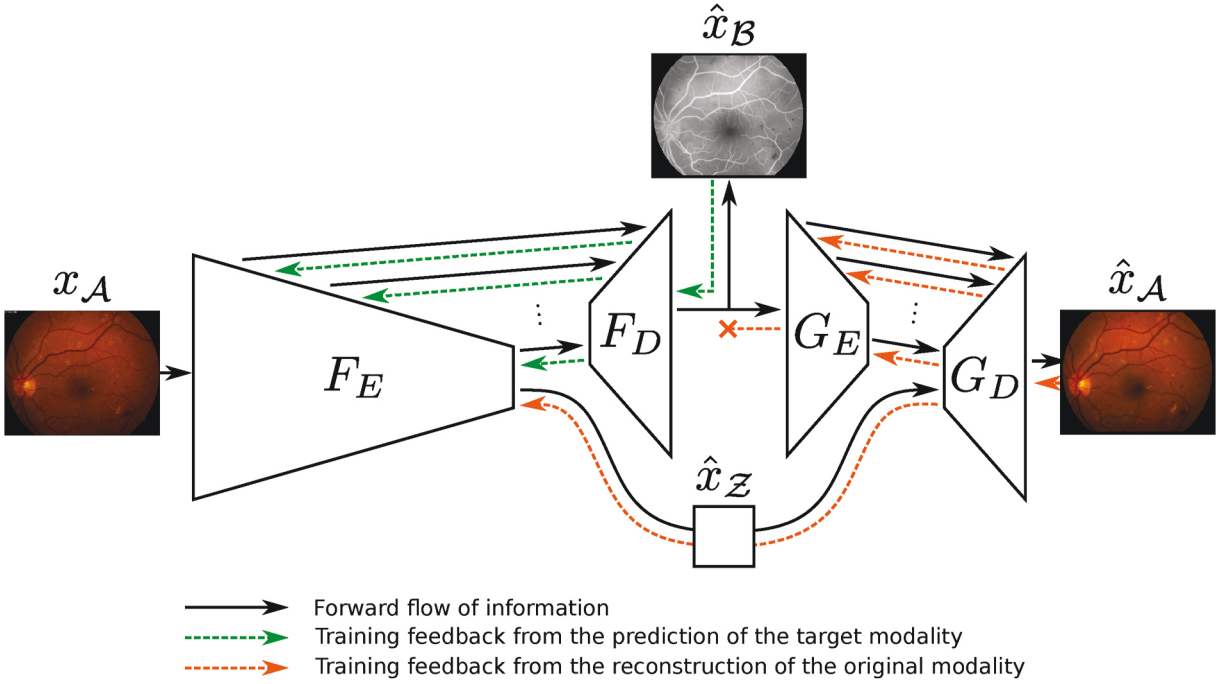

Fig. 5. Proposed network architecture for the pre-training of a convolutional encoder using MIE.

= 1.5 [50]. c_1 and c_2 are small constants to avoid instability when the denominator is close to zero [50].

Then, the loss for the prediction of the target image modality is defined as:

$$\mathcal{L}_B = -\frac{1}{N}\sum_n^N SSIM(\hat{\mathbf{x}}_{B,n}, \mathbf{x}_{B,n}) \tag{2}$$

where N denotes the number of pixels in the images.

Given that both the input image reconstruction and the multimodal prediction provide a training target of similar characteristics, we also use the negative SSIM as loss function for the input image reconstruction. Thus, this loss is defined as:

$$\mathcal{L}_A = -\frac{1}{NC}\sum_c^C \sum_n^N SSIM(\hat{\mathbf{x}}_{A,n,c}, \mathbf{x}_{A,n,c}) \tag{3}$$

where C denotes the number of channels in the color images.

Finally, the training is performed by the optimization of the joint training loss defined as:

$$\mathcal{L}_{MIE} = \mathcal{L}_A + \mathcal{L}_B \tag{4}$$

2.1.3. Datasets

The pre-training is performed using a public multimodal dataset of retinal images provided by Isfahan MISP [26].[1] This dataset is composed of 59 multimodal image pairs consisting of a color retinography and a fluorescein angiography image of the same eye. Half of these image pairs correspond to patients diagnosed with DR whereas the other half correspond to healthy individuals. However, given that the proposed MIE pre-training only requires unlabeled data, the labels regarding the presence of DR are not used. The retinography-angiography pairs are automatically aligned following the methodology proposed in Ref. [49].

[1] https://misp.mui.ac.ir/en/node/1498.

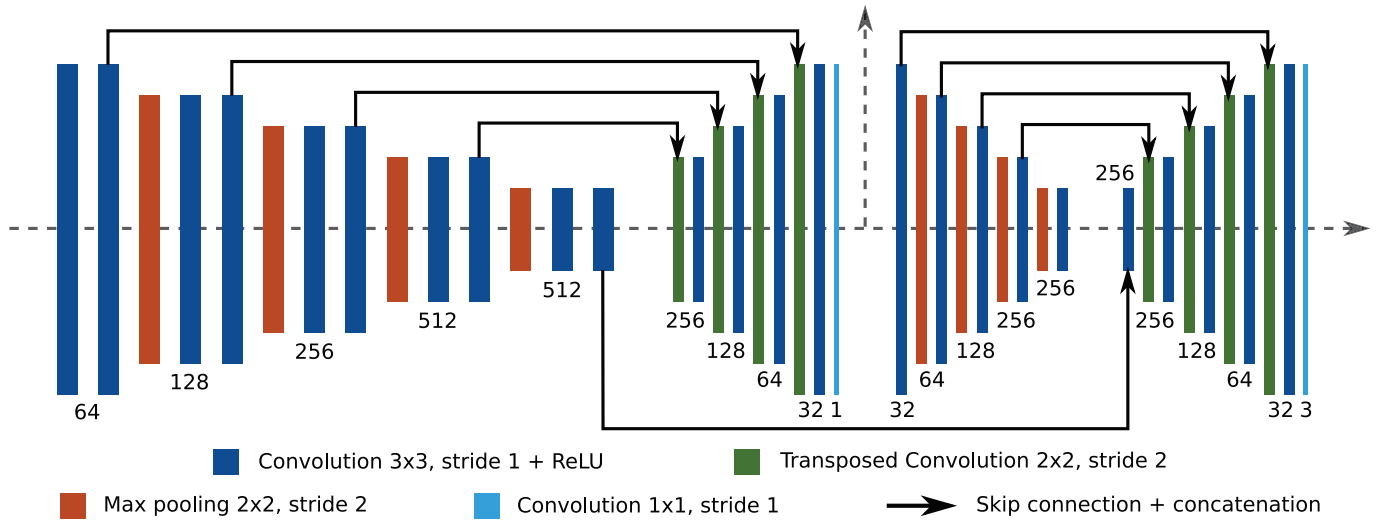

Fig. 6. Detailed diagram of the proposed network design for MIE. Each block represents the output of a layer in the network and the numbers above or below the blocks denote the number of output channels for the corresponding layers. The number of input channels for each layer is the same as the number of output channels of the previous layer or the summation of output channels if there is a concatenation operation. From left to right: Encoder F_E, Decoder F_D, Encoder G_E, and Decoder G_D.

2.1.4. Training details

The training is performed on 50 image pairs and the remainder 9 pairs are used for validation purposes. The neural network is randomly initialized following the approach proposed by He et al. [51]. Then, the training is performed using the Adam optimization algorithm [52] with the standard decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The training is performed at a constant learning rate $\alpha = 1e-4$ for a total of 5000 epochs. The batch size is one image and the samples are resized to a fixed width of 400 pixels. In order to avoid overfitting when training with a small number of samples, we use a data augmentation strategy including both spatial and color augmentations. In particular, the spatial augmentations are affine transformations including scaling, rotation, and shearing, whereas the color augmentations are channel-wise transformations performed in HSV color space, similar to those proposed in Ref. [24].

2.2. Diabetic retinopathy grading

For the grading of DR we follow the most common approach in the literature [9]. In this regard, the grading of DR is formulated as a multi-class classification where the likelihood for several mutually exclusive classes must be predicted. These classes represent the different grades of the pathology, which can be 4 or 5 depending on the clinical criteria being applied [5,13]. This multi-class classification is trained using the cross-entropy between the predicted likelihoods and the ground truth labels as loss function.

Regarding the network architecture, the classification is performed using a fully convolutional encoder followed by a global average pooling and a fully connected layer. The global pooling produces a fixed set of features regardless of the input image size and then the fully connected layers generate the final prediction for each class. Additionally, these predictions are normalized by applying a softmax operation to the output. In our proposal, the encoder is pre-trained using the proposed MIE approach.

2.2.1. Datasets

Regarding the grading of DR, the main experiments and comparisons in this work are performed on the public IDRiD [11] and Messidor [13] datasets. For the comparison with the state-of-the-art, an additional cross-dataset experiment is performed on the public EyePACS-Kaggle [32] and Messidor-2 [53] datasets.

IDRiD: This dataset consists of 516 retinal images from clinical exams performed in India. The image acquisition was performed with pupil dilation using a Kowa VX-10 alpha fundus camera. The images are categorized into the 5 severity grades. The ground truth was annotated by two medical experts following the ICDR severity scale [10], which is depicted in Table 1. The dataset presents a default split of 403 images for training and 113 for test. This default training/test split is used in all the experiments.

Messidor: This dataset consists of 1200 retinal images acquired by 3 different ophthalmology departments in France. The image acquisition was performed with a 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph. The images are categorized into 4 severity grades. The ground truth was annotated by the medical experts of the different ophthalmic departments that provided the images, following the severity scale that is depicted in Table 2. Additionally, 800 images were acquired with pupil dilation and 400 without dilation. This dataset does not present any default split and, therefore, we perform all the experiments using a 5-fold cross-validation approach.

EyePACS-Kaggle (*training set*): This dataset consists of 35,126 retinal images acquired with different cameras. The dataset aims to promote the development of robust algorithms in "real-world" data, hence including noise in both images and labels. The images are categorized into 5 severity grades. This dataset is used as training set for the cross-dataset experiment in the state-of-the-art comparison.

Messidor-2: This dataset is an extended and unlabeled version of Messidor consisting of 1748 images. The additional images are acquired under the same conditions as those from the original dataset. The labels for 1744 images are provided by Ref. [5]. In this case, the images are categorized into 5 severity grades. The ground truth was annotated by three medical experts following the ICDR severity scale [10], which is depicted in Table 1. This dataset is used as test set for the cross-dataset experiment in the state-of-the-art comparison.

2.2.2. Training details

In order to reflect all the common practices in the literature [19] and evaluate the proposed approach in different scenarios, we perform two different types of experiments. In the first type, the pre-trained encoder is used as a fixed feature extractor and only the final fully connected layer is fine-tuned for the grading of DR. In contrast, in the second type, the whole network is fine-tuned for the grading of DR.

- Fixed feature extractor: First, the features for each training image are extracted offline using the pre-trained encoder followed by the global average pooling. Then, a linear classifier (i.e. the last fully connected layer) is trained on top of the extracted features following a multinomial logistic regression approach. In this case, the optimization is performed until convergence in the training set using the L-BFGS-B algorithm with $L2$ regularization [54].
- Full network fine-tuning: The optimization is performed using the Adam algorithm [52] with batch size of 8 images and the standard decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to $\alpha = 1e-4$ and it is reduced by a factor of 10 each time that the validation loss does not improve for 10 epochs, up to a minimum of $\alpha = 1e-6$. Then, the training is stopped. In order to avoid overfitting, we use a data augmentation strategy including both spatial and color augmentations. In particular, the spatial augmentations include random rotations and mirroring, whereas the color augmentations are channel-wise transformations performed in HSV color space, similar to those proposed in Ref. [24].

The images are resized to a fixed retinal width of 400 pixels in all the experiments.

2.3. Evaluation metrics

The performance is evaluated using the same metrics that are common in previous works focusing on the detection and grading of DR [9]. In that regard, the networks are evaluated for the grading of DR using Accuracy, Average Classification Accuracy (ACA), and Quadratic Weighted Kappa (QWK). Additionally, to assess the performance in each individual class, the normalized confusion matrices are also computed.

ACA represents a class-balanced version of Accuracy that is computed by individually measuring the accuracy for each class and then computing the average value. Thus, all the classes present the same importance in the final score regardless of the number of samples in each of them. This metric is particularly useful for the grading of DR due to the different number of samples per severity grade that is present in all the datasets.

QWK is a quadratically weighted version of the Cohen's Kappa score [55], which expresses the level of agreement corrected by the probability of agreement by chance. In this regard, QWK = 1 denotes complete agreement between predictions and ground truth whereas QWK = 0 denotes that the same results could be achieve by random classification. In ordinal classification (i.e. when the classes present an inherent order), the quadratic weighting assigns a larger weight to the samples with larger disagreement. In particular, the weight is proportional to the square of the ordinal distance (OD) between the prediction and the ground truth. This metric is commonly used in the grading of DR because it takes into account that the classes are ordered by severity and that a closer erroneous prediction is preferable than a farther one [16,46]. For instance, for an image with ground truth grade 1, a prediction of grade 2 (OD = 1), despite erroneous, is preferable than a prediction of grade 4 (OD = 3).

Besides the previous evaluation, the networks trained for multi-class classification are also evaluated for two binary classification problems that are common in the literature, DR detection (0 vs 1,2, …) and referable DR (rDR) detection (0,1 vs 2, …) [9,14]. This evaluation is performed using Receiver Operator Characteristic (ROC) curves, which plot True Positive Rate (TPR) against False Positive Rate (FPR) for different decision thresholds. Additionally, we also compute the Area Under Curve (AUC) for ROC, which is commonly used to summarized the performance into a single value.

2.4. Alternative approaches

The evaluation of MIE for the grading of DR includes several comparisons against other relevant methods for the pre-training or initialization of the neural network. In that regard, all the alternatives in our experimentation follow the same fine-tuning methodology that is described in Section 2.2 and the difference among them is the pre-training phase. In particular, we consider the following alternatives for the pre-training phase:

- Random: In this case, no pre-training is performed and the neural network is randomly initialized following the formula proposed by He et al. [51].
- ImageNet: The neural network is pre-trained on the ImageNet dataset [17] using a fully-supervised image classification task. In particular, we use a pre-trained network that is provided in the computer vision library of the PyTorch project [56]. In contrast with the proposed MIE pre-training, this alternative requires labeled data and it is performed on natural images (including e.g. people, vehicles, etc.) instead of retinal images.
- MR [24]: The neural network is pre-trained on an unlabeled multimodal dataset of retinal images using a previous state-of-the-art methodology. This approach can be seen as a subcase of MIE that only uses the training feedback from the prediction of the target modality ($\mathcal{L}_B$ in Eq. (4)). Likewise, MR uses a simpler architecture only consisting of the encoder F_E and the decoder F_D (see Fig. 5). To produce a fair comparison, the MR pre-training is performed using the same training configuration that is described in Section 2.1.4 and the same multimodal dataset that is described in Section 2.1.3.
- MIE: The neural network is pre-trained on an unlabeled multimodal dataset of retinal images using the proposed MIE pre-training that is described in Section 2.1.

2.5. Implementation details

The proposed methodology and the experiments in this work were implemented in Python 3 using the following opens source libraries: PyTorch for the parts specific to deep learning, Scikit-Image for the image loading and processing, and Scikit-Learn for the evaluation metrics and training of linear classifiers. The training of the neural networks was performed in GPU using an NVIDIA GTX 1070 with memory size of 8 GB.

3. Experimental results

3.1. Multimodal image encoding

In order to comprehensively evaluate the proposed methodology, first we perform a qualitative analysis of the results obtained in the proposed pre-training. In this regard, Fig. 7 depicts predicted and reconstructed images for some representative examples of the validation set. Additionally, besides the predicted angiography and reconstructed retinography, we include two alternative reconstructions that allow to study which information MIE has encoded through each of the two branches in the network. In particular, Fig. 7 depicts the following images:

- $\mathbf{x_A}$: The original retinography that is used as input to the network.
- $\widehat{\mathbf{x}}_\mathbf{B}$: The predicted angiography.
- $\widehat{\mathbf{x}}_\mathbf{A}$: The reconstructed retinography.
- $\widehat{\mathbf{x}}_A^1$: The reconstructed retinography when only the information provided by the first branch of the network is used. This branch aims at learning the common features between modalities.
- $\widehat{\mathbf{x}}_A^2$: The reconstructed retinography when only the information provided by the second branch of the network is used. This branch aims at learning the features that are exclusive to the input modality.

All these examples, including the alternative reconstructions, are generated using the same pre-trained network. In order to generate the

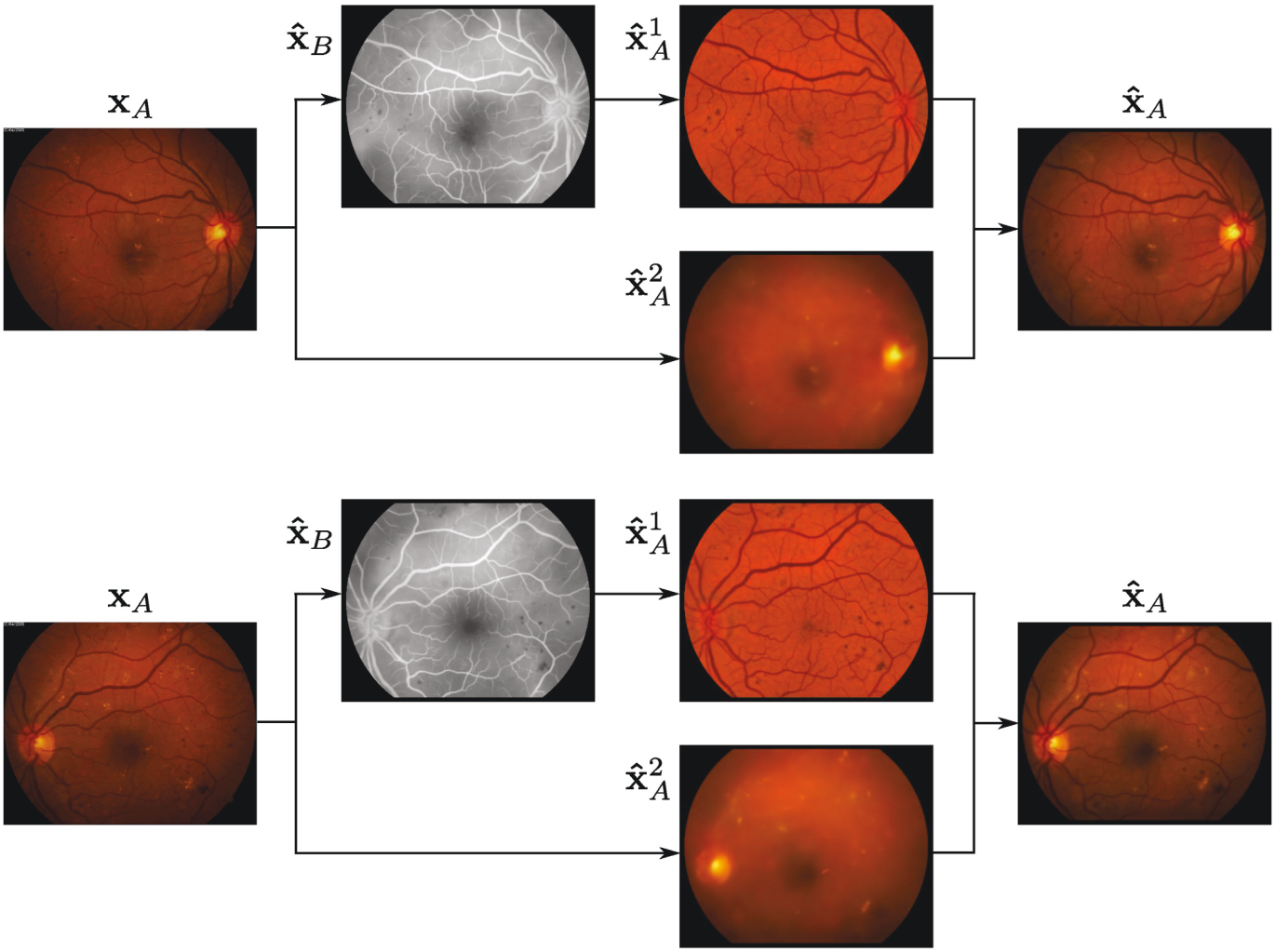

Fig. 7. Examples of predicted and reconstructed images using MIE. $\mathbf{x_A}$ represents the input retinography, $\hat{\mathbf{x}}_\mathbf{B}$ the predicted angiography, and $\hat{\mathbf{x}}_\mathbf{A}$ the reconstructed retinography. Additionally, $\hat{\mathbf{x}}_A^1$ and $\hat{\mathbf{x}}_A^2$ represent alternative reconstructions that are generated by using only the features of the first or second branch in the network, respectively.

alternative reconstructions $\widehat{\mathbf{x}_A^1}$ and $\widehat{\mathbf{x}_A^2}$ using only the information of the first and the second branch, respectively, the features of the opposite branch are masked out in the forward pass through the network at test time. In particular, the features of the opposite branch are multiplied by a tensor of zeros before being fed to the final decoder G_D in the network (see Fig. 5).

The results show that MIE successfully generates angiographies and retinographies that are consistent with the content of the input images. The reconstructed retinographies are very similar to the input images (0.93 SSIM [50] for the images in the validation set), which indicates that the network has successfully learned a rich set of features that allow to completely describe the image contents in this modality. The main difference that can be perceived between the reconstructed and the original images is the lack of fine structural detail for some of the reconstructed bright lesions. In this regard, it must be noticed that the bright lesions are exclusive to the retinography and, therefore, they are not depicted in the predicted angiography. Thus, the information related to these lesions must be propagated trough the bottleneck of the second branch, which explains the loss of fine structural detail in some cases.

The alternative reconstructions, $\hat{\mathbf{x}}_A^1$ and $\hat{\mathbf{x}}_A^2$, confirm that, as intended, the network encodes different information in each of the two branches. In particular, $\hat{\mathbf{x}}_A^1$ depicts only those characteristics of the retinography that can be inferred from an angiography, such as the vasculature or the red lesions In contrast, $\hat{\mathbf{x}}_A^2$ mainly depicts the characteristics that are exclusive to the retinography, such as the color for different parts of the image, the bright lesions, or the detail within the optic disc. Thus, as intended, the first branch encodes the features that are common between modalities whereas the second branch encodes the features that are exclusive to the input modality. These two branches together allow the learning of a rich set of features that completely describe the input image modality in the encoder that is being pre-trained.

3.2. Pre-trained network as feature extractor

To study the quality of the representations learned by the proposed approach, we perform a set of experiments using the pre-trained encoder as fixed feature extractor. In this case, as described in Section 2.2.2, a linear classifier is trained on top of the features extracted by the encoder. For the IDRiD dataset we use the default training/test split and for the Messidor dataset we use a 5-fold cross-validation approach. Tables 3 and 4 depict the results obtained for the IDRID and Messidor datasets, respectively, using the different studied approaches (Random, ImageNet, MR, and MIE).

These results show that there is a large performance gap between Random and the different pre-training alternatives. This indicates that all the studied alternatives provide features that are useful for the

Table 3
Results for DR grading on IDRiD using the pre-trained networks as fixed feature extractors. Bold denotes the best result for each evaluation metric.

Method	Pre-training	QWK	ACA	Accuracy	DR AUC	rDR AUC
Random	None	0.00	20.00	33.01	35.81	36.58
ImageNet	Supervised	58.66	43.49	53.40	87.08	88.50
MR	Self-supervised	66.42	43.92	60.19	87.38	91.51
MIE (Proposed)	Self-supervised	**71.05**	**46.30**	**61.17**	**91.90**	**94.59**

Table 4
Results for DR grading on Messidor using the pre-trained networks as fixed feature extractors. Bold denotes the best result for each evaluation metric.

Method	Pre-training	QWK	ACA	Accuracy	DR AUC	rDR AUC
Random	None	0.00 ± 0.00	25.0 ± 0.00	46.08 ± 0.10	62.14 ± 2.90	61.64 ± 3.98
ImageNet	Supervised	**71.61 ± 2.18**	**54.76 ± 1.96**	63.69 ± 1.89	**85.53 ± 2.28**	**89.6 ± 2.47**
MR	Self-supervised	63.87 ± 2.65	47.69 ± 1.26	62.51 ± 1.43	81.37 ± 1.89	86.39 ± 2.61
MIE (Proposed)	Self-supervised	70.77 ± 1.86	54.18 ± 3.20	**66.05 ± 2.74**	84.39 ± 2.60	89.13 ± 2.10

classification of the images, always improving the results obtained before any pre-training. However, not all the pre-training methods provide the same performance. In this regard, MIE represents the best alternative, given that it is the best approach in IDRID and a close second best in Messidor. Additionally, MIE also provides the highest Accuracy on Messidor despite the similar ACA between ImageNet and MIE. This difference between Accuracy and ACA is explained by the different number of samples among classes on the Messidor dataset. In that regard, the obtained results indicate that MIE provides a better performance in the most numerous classes of this dataset. In contrast, while ImageNet pre-training obtained the best performance in Messidor, it was the worse pre-training in IDRiD. This shows that the features learned using this approach do not adapt equally well to all the possible scenarios in retinal imaging. In this regard, despite of being a limited domain, retinal images still present a high variability due to the different acquisition devices and procedures as well as the individuals ethnicity. Additionally, the distribution of grades and even the grading criteria can vary among datasets. In this case, IDRiD and Messidor datasets present key differences in all these aspects, as described in Section 2.2.1.

3.3. Transfer learning with full network fine-tuning

The full potential of MIE for transfer learning is studied by performing a full network fine-tuning in the grading of DR. For the IDRiD dataset we use the default training/test split and for the Messidor dataset we use a 5-fold cross-validation approach. In these experiments, 25% of the training set is used as validation subset to apply the learning rate schedule and early stopping. Given the small size of the IDRiD dataset, in order to account by the variability in the selection of the validation samples, we perform 4 training repetitions following a 4-fold procedure for the training/validation splits. Tables 5 and 6 depict the results of these experiments for the IDRID and Messidor datasets, respectively, using the different studied approaches (Random, ImageNet, MR, and MIE). Additionally, Figs. 8 and 9 depict the average confusion matrix for each method and dataset.

With regards to the detection of DR and rDR, Figs. 10 and 11 depict the ROC curves for the different pre-training approaches on the IDRiD and Messidor datasets, respectively.

Similarly to the previous experiments, the results show a large performance gap between Random and the different pre-training alternatives. Thus, the studied pre-training approaches not only provide useful features readily available for the classification but also facilitate the complete training of the network using task-specific annotated datasets. In this case, MIE represents the best alternative for both datasets, outperforming MR as well as ImageNet pre-training. In the IDRiD dataset, MR provides a slightly higher mean rDR AUC than MIE. In this regard, a higher rDR AUC indicates a better separation between grades \{0,1\} and grades \{2,3,4\}, which corresponds with the detection of rDR. However, the difference in rDR AUC is small and, overall, the results indicate a better performance of MIE.

The comparisons shows that MIE leverages the domain-specific knowledge provided by the unlabeled multimodal data to a greater extent than MR. Additionally, in comparison to the supervised pre-training on a different domain (i.e. the natural images of ImageNet), pre-training on the same domain using MIE results in better transfer learning performance even when the number of pre-training samples is several orders of magnitude lower.

With regards to ImageNet pre-training, in this case a similar performance, in relative terms, is achieved for IDRiD and Messidor. The performance gap between datasets (in relative terms) that was observed in Section 3.2 is reduced when the ImageNet pre-trained networks are fine-tuned for the specific characteristics of each dataset.

Regarding the performance for each class, all the pre-training alternatives produce a similar pattern in the confusion matrices. In that sense, in the IDRiD dataset, MIE outperforms the other pre-training alternatives in 3 out of 5 classes, whereas, in the Messidor dataset, MIE consistently outperforms the other alternatives across all the classes.

3.4. State-of-the-art comparison

In this section, we provide a comparison of our proposal against relevant state-of-the-art methods for the grading of DR. The proposed methodology for these comparisons, MIE-DR, corresponds to performing a complete network fine-tuning for the grading of DR after the pre-training using MIE. For each dataset, the comparison is performed using the same evaluation metrics that are used in previous works.

First, Tables 7 and 8 depict the state-of-the-art comparisons for IDRiD and Messidor, respectively. In this case, the results of MIE correspond to the same experiments that are performed in Section 3.3.

Then, in order to compare with other relevant works in the literature,

Table 5
Results after performing transfer learning for DR grading on IDRiD. Bold denotes the best result in terms of mean value for each evaluation metric.

Method	Pre-training	QWK	ACA	Accuracy	DR AUC	rDR AUC
Random	None	14.97 ± 6.89	22.89 ± 2.12	36.41 ± 3.53	68.58 ± 1.85	65.77 ± 1.79
ImageNet	Supervised	71.93 ± 4.77	49.97 ± 1.9	64.08 ± 2.66	91.35 ± 1.84	93.38 ± 1.0
MR	Self-supervised	72.97 ± 1.69	46.24 ± 2.9	59.95 ± 1.26	91.74 ± 0.67	**94.46 ± 0.46**
MIE (Proposed)	Self-supervised	**76.44 ± 1.54**	**51.59 ± 2.36**	**65.05 ± 1.19**	**93.38 ± 0.37**	94.36 ± 0.21

Table 6
Results after performing transfer learning for DR grading on Messidor. Bold denotes the best result in terms of mean value for each evaluation metric.

Method	Pre-training	QWK	ACA	Accuracy	DR AUC	rDR AUC
Random	None	12.08 ± 14.91	29.03 ± 5.1	48.1 ± 2.74	64.42 ± 2.7	63.27 ± 5.12
ImageNet	Supervised	87.77 ± 1.97	68.05 ± 3.76	75.74 ± 3.86	93.91 ± 1.13	96.92 ± 0.98
MR	Self-supervised	88.4 ± 1.68	67.39 ± 2.62	74.73 ± 2.33	93.23 ± 1.03	97.1 ± 0.81
MIE (Proposed)	Self-supervised	**89.7 ± 0.87**	**72.55 ± 1.75**	**79.44 ± 0.59**	**94.22 ± 0.92**	**97.4 ± 0.47**

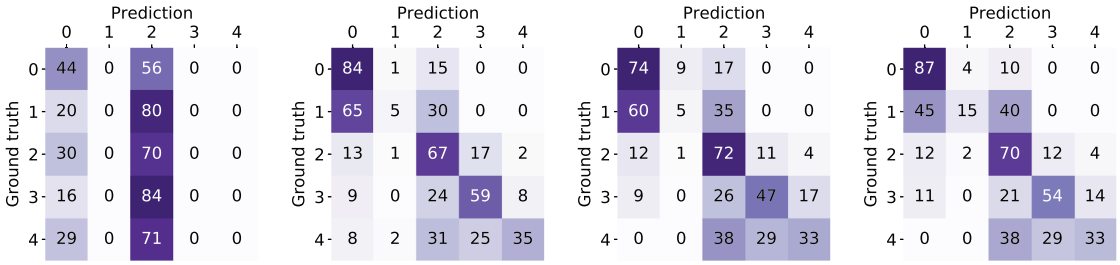

Fig. 8. Average normalized confusion matrices after performing transfer learning for DR grading on IDRiD.

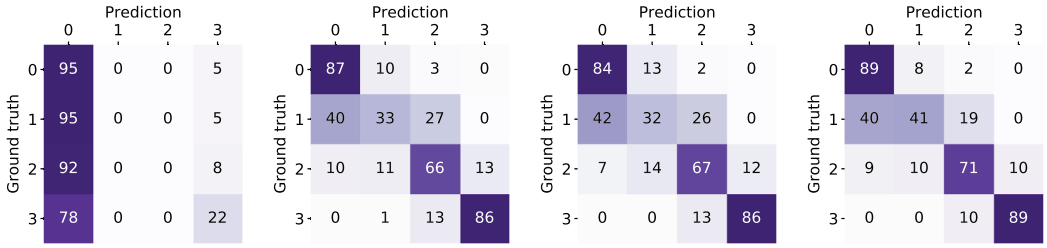

Fig. 9. Average normalized confusion matrices after performing transfer learning for DR grading on Messidor.

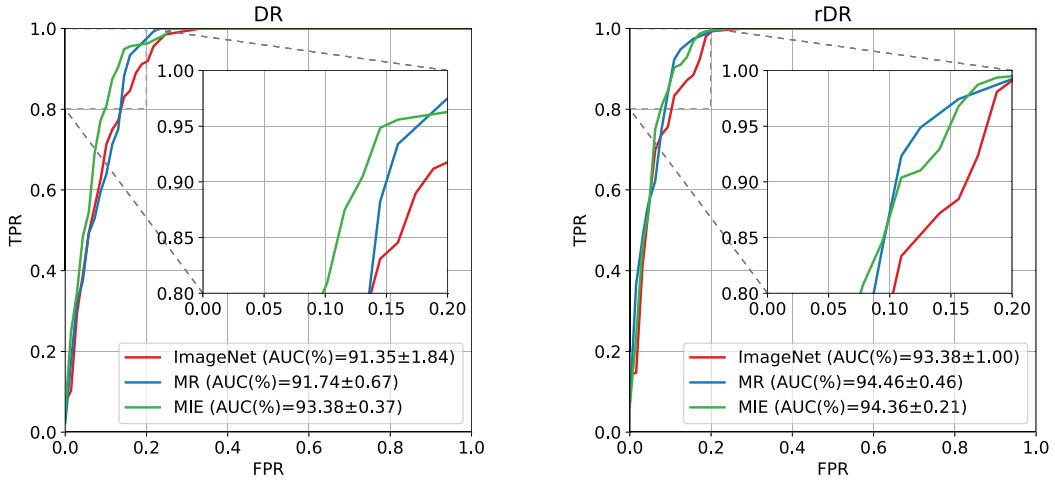

Fig. 10. Mean ROC curves for the detection of DR and rDR on IDRiD.

we perform an additional cross-dataset experiment by training the networks on the EyePACS-Kaggle dataset and evaluating the performance on Messidor-2. In this case, given the very large size of the training dataset, only 10% of the training data is used for validation and a single training experiment is performed. Table 9 depicts the state-of-the-art comparison for this cross-dataset experiment.

The provided comparisons show that MIE-DR is competitive in all the scenarios, always achieving the best performance for the grading of DR. In this regard, the proposed approach is advantageous even when a large annotated training dataset is available for the grading of DR, such is the

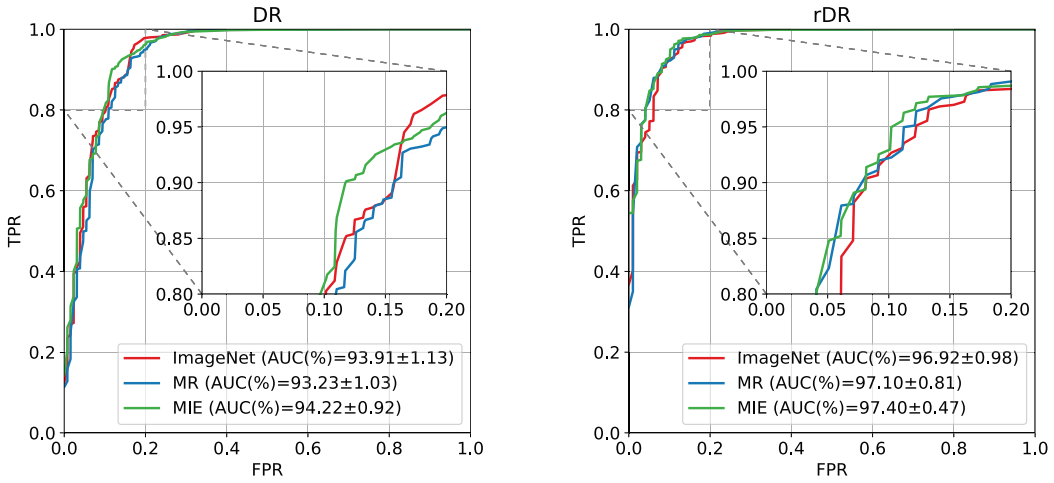

Fig. 11. Mean ROC curves for the detection of DR and rDR on Messidor.

Table 7
State-of-the-art comparison for DR grading on the IDRiD dataset. Bold denotes the best result.

Method	Accuracy
Wu et al. [42]	60.20
Tu et al. [47]	**65.05**
MIE-DR (Proposed)	**65.05 ± 1.19**

Table 8
State-of-the-art comparison for DR grading on the Messidor dataset. Bold denotes the best result for each evaluation metric.

Method	ACA	DR AUC	rDR AUC
Cao et al. [57]	-	93.90 ± 0.90	-
Li et al. [43]	-	-	96.30
Martinez-Murcia et al. [58]	56.40 ± 13.00	81.00 ± 7.00	93.00 ± 2.00
Hua et al. [27]	-	-	**99.40 ± 1.10**
MIE-DR (Proposed)	**72.55 ± 1.75**	**94.22 ± 0.92**	97.4 ± 0.47

Table 9
State-of-the-art comparison for DR grading in a cross-dataset setting. Training on EyePACS-Kaggle and evaluation on Messidor-2. Bold denotes the best result for each evaluation metric.

Method	QWK	ACA
Araújo et al. [46]	71.00	59.60
Galdran et al. [16]	79.79 ± 1.03	63.41 ± 1.99
de la Torre et al. [45]	83.20	-
MIE-DR (Proposed)	**83.43**	**67.03**

case of the cross-dataset experiment in Table 9. Moreover, in this case, Additionally, Fig. 8 shows that MIE-DR is also competitive for the classification of DR (0 vs 1,2,3) or rDR (0,1 vs 2,3) against works that specifically address these tasks as a binary classification. Conclusively, these results show that MIE is a viable and effective pre-training alternative for the grading of DR. In this regard, previous works are typically based on the used of ImageNet pre-training networks, requiring additional annotated data. In contrast, the proposed MIE pre-training is performed using only unlabeled images. In our proposal, the lack of annotations is successfully compensated by taking advantage of a small set of multimodal images corresponding to the final application domain, i.e. retinal imaging.

4. Discussion

The automated grading of DR represents one of the most challenging tasks in the field of retinal image analysis. This task requires to take into consideration different lesions and retinal anomalies, in order to then produce a high level analysis of these findings to estimate the adequate grade of the disease. In our experiments, the difficulty of the grading of DR is demonstrated by the low performance that is achieved when training a DNN from scratch (Tables 5 and 6 and Figs. 8 and 9). In that case, following a standard training methodology, the networks fail to converge to a satisfactory solution.

In the literature, the typical approach to train a DNN in the grading of DR is the reuse of networks pre-trained for ImageNet classification. In this regard, previous works have proposed different methodologies, including modifications to the network architecture and the training objective, but always building upon the ImageNet pre-training. That approach presents the disadvantages of requiring additional annotated data and being performed on a different application domain. In contrast to all the previous works, the novelty of our methodology, MIE-DR, is precisely on the pre-training phase. In particular, we propose a novel self-supervised pre-training, MIE, that is performed on unlabeled multimodal data of the final application domain, i.e. retinal images.

In order to demonstrate the advantages of MIE as pre-training, we evaluate the transfer learning performance under two different experimental settings using a standard network architecture and training methodology. These results show that, in general, MIE is superior to existing alternatives, such as ImageNet and MR pre-training. In particular, MIE provides satisfactory results on IDRiD and Messidor for both the linear classification and full network fine-tuning evaluations. In contrast, ImageNet pre-training achieves the best results in a particular scenario for linear classification but it is not robust enough to changes in the data distribution neither it takes the same advantage of the full network fine-tuning. This outcome is also supported by the state-of-the comparisons, where MIE-DR is highly competitive and outperforms previous works based on ImageNet pre-trained networks.

With regards to MR pre-training, this previous multimodal approach provides a performance that is competitive with ImageNet pre-training in several scenarios. However, in general, it is outperformed by MIE. This shows that MIE allows to further take advantage of the multimodal data by explicitly learning the features that are common between modalities as well as the features exclusive to the input modality. This translates into better high level representations of the input modality, subsequently improving the transfer learning performance, either using linear classification or full network fine-tuning. Additionally, the success of MIE disentangling the common and exclusive features for the input

image modality is also evident by the visual analysis of the reconstructed images in Fig. 7.

Despite the satisfactory results, MIE-DR also presents some limitations that are common to previous works on the grading of DR, such as the uneven performance among classes (Figs. 8 and 9). This is mainly due to the unbalanced distribution of samples in the training data and the existence of particular challenging cases, e.g. the distinction between grades 0 and 1 that depends on the presence of tiny microaneurysms. In this regard, the focus of this work is exclusively on providing a novel pre-training alternative using unlabeled multimodal data. This is the first work addressing the grading of DR that focus on this important aspect of the methodology. However, achieving a balanced performance among classes without compromising the overall results is an important objective that should be considered in future works. Additionally, aiming at improving the classification of the most challenging cases, future works could explore multi-task learning approaches by simultaneously performing the segmentation of retinal lesions together with the grading of DR. In this regard, the generation of lesions maps could also be used as a means of improving the interpretability of the learned models. Finally, these future works could take advantage of the results presented in this work and build novel methodologies upon MIE pre-trained networks, for both the grading of DR and the segmentation of retinal lesions.

Another important future research direction is to adapt the idea of MIE to other relevant image modalities such as, e.g., OCT or OCT-A, or even different image modalities from other medical domains. In that regard, it would be worth exploring the adaption of the proposed idea to 3D multimodal data, which is also common in medicine.

5. Conclusions

The grading of DR is usually approached by reusing networks pre-trained for supervised image classification on the ImageNet dataset. In this context, we proposed MIE as a novel self-supervised alternative that takes advantage of unlabeled multimodal images pairs for pre-training the networks. In contrast to previous multimodal pre-training methods, our proposal explicitly teaches the networks to recognize the common characteristics between modalities as well as the characteristics exclusive to the input modality. This ensures a complete understanding of the retina in the same image modality that will be later used for the grading of DR.

The proposed methodology is evaluated on several public datasets under different experimental settings. First, the performed analyses indicate that MIE successfully learns and disentangles the common and exclusive characteristics between modalities. Second, the transfer learning results show that, overall, MIE outperforms previous multimodal pre-training methods as well as the commonly used ImageNet pre-training. This is corroborated by the state-of-the-art comparison where our proposal is competitive and usually outperforms previous works. These satisfactory results open the door to further developments in the detection and grading of DR, building upon neural networks that are successfully pre-trained on unlabeled multimodal data.

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; Axencia Galega de Innovación (GAIN), Xunta de Galicia, ref. IN845D 2020/38; 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%). Funding for open access charge: Universidade da Coruña/CISUG.

References

[1] D.S.W. Ting, G.C.M. Cheung, T.Y. Wong, Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review, Clin. Exp. Ophthalmol. 44 (4) (2016) 260–277, https://doi.org/10.1111/ceo.12696.

[2] Z. L. Teo, Y.-C. Tham, M. Yu, M. L. Chee, T. H. Rim, N. Cheung, M. M. Bikbov, Y. X. Wang, Y. Tang, Y. Lu, I. Y. Wong, D. S. W. Ting, G. S. W. Tan, J. B. Jonas, C. Sabanayagam, T. Y. Wong, C.-Y. Cheng, Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis, Ophthalmology ISSN 0161-6420, doi: https://doi.org/10.1016/j.ophtha.2021.04.027.

[3] L.Z. Heng, O. Comyn, T. Peto, C. Tadros, E. Ng, S. Sivaprasad, P.G. Hykin, Diabetic retinopathy: pathogenesis, clinical grading, management and future developments, Diabet. Med. 30 (2013) 640–650, https://doi.org/10.1111/dme.12089. ISSN 1464-5491.

[4] S. Vujosevic, S.J. Aldington, P. Silva, C. Hernández, P. Scanlon, T. Peto, R. Simó, Screening for diabetic retinopathy: new perspectives and challenges, Lancet Diabetes Endocrinol. 8 (4) (2020) 337–347, https://doi.org/10.1016/S2213-8587(19)30411-5. ISSN 2213-8587.

[5] J. Krause, V. Gulshan, E. Rahimy, P. Karth, K. Widner, G.S. Corrado, L. Peng, D. R. Webster, Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy, Ophthalmology 125 (8) (2018) 1264–1272, https://doi.org/10.1016/j.ophtha.2018.01.034. ISSN 0161-6420.

[6] N. Cho, J. Shaw, S. Karunga, Y. Huan, J. da Rocha Fernandes, A. Ohlrogge, B. Malanda, IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045, Diabetes Res. Clin. Pract. 138 (2018) 271–281, https://doi.org/10.1016/j.diabres.2018.02.023. ISSN 0168-8227.

[7] S. Stolte, R. Fang, A survey on medical image analysis in diabetic retinopathy, Med. Image Anal. 64 (2020) 101742, https://doi.org/10.1016/j.media.2020.101742. ISSN 1361-8415.

[8] G. Lim, V. Bellemo, Y. Xie, X. Q. Lee, M. Y. T. Yip, D. S. W. Ting, Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: a review, Eye Vis. 7, ISSN 2326-0254, doi:https://doi.org/10.1186/s40662-020-00182-7.

[9] N. Tsiknakis, D. Theodoropoulos, G. Manikis, E. Ktistakis, O. Boutsora, A. Berto, F. Scarpa, A. Scarpa, D.I. Fotiadis, K. Marias, Deep learning for diabetic retinopathy detection and classification based on fundus images: a review, Comput. Biol. Med. 135 (2021) 104599, https://doi.org/10.1016/j.compbiomed.2021.104599. ISSN 0010-4825.

[10] C. Wilkinson, F.L. Ferris, R.E. Klein, P.P. Lee, C.D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J.T. Verdaguer, Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales, Ophthalmology 110 (9) (2003) 1677–1682, https://doi.org/10.1016/S0161-6420(03)00475-5. ISSN 0161-6420.

[11] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao, T. Wu, J. Xiao, F. Wang, B. Yin, Y. Wang, G. Danala, L. He, Y. H. Choi, Y.C. Lee, S.-H. Jung, Z. Li, X. Sui, J. Wu, X. Li, T. Zhou, J. Toth, A. Baran, A. Kori, S.S. Chennamsetty, M. Safwan, V. Alex, X. Lyu, L. Cheng, Q. Chu, P. Li, X. Ji, S. Zhang, Y. Shen, L. Dai, O. Saha, R. Sathish, T. Melo, T. Araújo, B. Harangi, B. Sheng, R. Fang, D. Sheet, A. Hajdu, Y. Zheng, A.M. Mendonça, S. Zhang, A. Campilho, B. Zheng, D. Shen, L. Giancardo, G. Quellec, F. Mériaudeau, IDRiD: diabetic retinopathy - segmentation and grading challenge, Med. Image Anal. 59 (2020) 101561, https://doi.org/10.1016/j.media.2019.101561. ISSN 1361-8415.

[12] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, H. Kang, Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening, Inf. Sci. 501 (2019) 511–522, https://doi.org/10.1016/j.ins.2019.06.011. ISSN 0020-0255.

[13] 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 publicly distributed image database: the MESSIDOR database, Image Anal. Stereol. 33 (3) (2014) 231, https://doi.org/10.5566/ias.1155.

[14] M.M. Islam, H.-C. Yang, T.N. Poly, W.-S. Jian, Y.-C. (Jack) Li, Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: a systematic review and meta-analysis, Comput. Methods Progr. Biomed. 191 (2020) 105320, https://doi.org/10.1016/j.cmpb.2020.105320. ISSN 0169-2607.

[15] A. He, T. Li, N. Li, K. Wang, H. Fu, CABNet: category Attention block for imbalanced diabetic retinopathy grading, IEEE Trans. Med. Imag. 40 (1) (2021) 143–153, https://doi.org/10.1109/TMI.2020.3023463.

[16] A. Galdran, J. Dolz, H. Chakor, H. Lombaert, I. Ben Ayed, Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images, in: A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M.A. Zuluaga, S.K. Zhou, D. Racoceanu, L. Joskowicz (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2020, 2020.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F.-F. Li, ImageNet: a large-scale hierarchical image database, in: The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2009.
[18] V. Cheplygina, M. de Bruijne, J.P. Pluim, Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, Med. Image Anal. 54 (2019) 280–296, https://doi.org/10.1016/j.media.2019.03.009. ISSN 1361-8415.
[19] M.A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Comput. Biol. Med. 128 (2021) 104115, https://doi.org/10.1016/j.compbiomed.2020.104115. ISSN 0010-4825.
[20] L. Jing, Y. Tian, Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, p. 1, https://doi.org/10.1109/TPAMI.2020.2992393, 1.
[21] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, D. Rueckert, Self-supervised learning for medical image analysis using image context restoration, Med. Image Anal. 58 (2019) 101539, https://doi.org/10.1016/j.media.2019.101539. ISSN 1361-8415.
[22] R. Zhang, P. Isola, A.A. Efros, Colorful image colorization, in: European Conference on Computer Vision, ECCV, 2016.
[23] T. Chen, S. Kornblith, M. Norouzi, G.E. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, ICML, 2020.
[24] A.S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised multimodal reconstruction of retinal images over paired datasets, Expert Syst. Appl. (2020) 113674doi, https://doi.org/10.1016/j.eswa.2020.113674.
[25] 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, https://doi.org/10.1111/ceo.12711.
[26] S.H.M. Alipour, H. Rabbani, M.R. Akhlaghi, Diabetic retinopathy grading by digital curvelet transform, Comput. Math. Methods Med. (2012), https://doi.org/10.1155/2012/761901.
[27] C.H. Hua, K. Kim, T. Huynh-The, J.I. You, S.Y. Yu, T. Le-Tien, S.H. Bae, S. Lee, Convolutional network with twofold feature augmentation for diabetic retinopathy recognition from multi-modal images, IEEE J. Biomed. Health Inform. (2020) 1, https://doi.org/10.1109/JBHI.2020.3041848, 1.
[28] Z. Zhou, Z. He, M. Shi, J. Du, D. Chen, 3D dense connectivity network with atrous convolutional feature pyramid for brain tumor segmentation in magnetic resonance imaging of human heads, Comput. Biol. Med. 121 (2020) 103766, https://doi.org/10.1016/j.compbiomed.2020.103766. ISSN 0010-4825.
[29] Á.S. Hervella, J. Rouco, J. Novo, M. Ortega, Deep multimodal reconstruction of retinal images using paired or unpaired data, in: 2019 International Joint Conference on Neural Networks, IJCNN, 2019, pp. 1–8, https://doi.org/10.1109/IJCNN.2019.8852082.
[30] A.S. Hervella, J. Rouco, J. Novo, M. Ortega, Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction, Appl. Soft Comput. (2020) 106210, https://doi.org/10.1016/j.asoc.2020.106210. ISSN 1568-4946.
[31] Álvaro S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis, Expert Syst. Appl. 185 (2021) 115598, https://doi.org/10.1016/j.eswa.2021.115598. ISSN 0957-4174.
[32] J. Cuadros, G. Bresnick, EyePACS: an adaptable telemedicine system for diabetic retinopathy screening, J. Diabet. Sci. Technol. 3 (3) (2009) 509–516, https://doi.org/10.1177/193229680900300315.
[33] N. Salamat, M.M.S. Missen, A. Rashid, Diabetic retinopathy techniques in retinal images: a review, Artif. Intell. Med. 97 (2019) 168–188, https://doi.org/10.1016/j.artmed.2018.10.009. ISSN 0933-3657.
[34] M. Mookiah, U.R. Acharya, R.J. Martis, C.K. Chua, C. Lim, E. Ng, A. Laude, Evolutionary algorithm based classifier parameter tuning for automatic diabetic retinopathy grading: a hybrid feature extraction approach, Knowl. Base Syst. 39 (2013) 9–22, https://doi.org/10.1016/j.knosys.2012.09.008. ISSN 0950-7051.
[35] E. AbdelMaksoud, S. Barakat, M. Elmogy, A comprehensive diagnosis system for early signs and different diabetic retinopathy grades using fundus retinal images based on pathological changes detection, Comput. Biol. Med. 126 (2020) 104039, https://doi.org/10.1016/j.compbiomed.2020.104039. ISSN 0010-4825.
[36] T. Walter, J.-C. Klein, P. Massin, A. Erginay, A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina, IEEE Trans. Med. Imag. 21 (10) (2002) 1236–1243, https://doi.org/10.1109/TMI.2002.806290.
[37] M. Usman Akram, S. Khalid, A. Tariq, S.A. Khan, F. Azam, Detection and classification of retinal lesions for grading of diabetic retinopathy, Comput. Biol. Med. 45 (2014) 161–171, https://doi.org/10.1016/j.compbiomed.2013.11.014. ISSN 0010-4825.
[38] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, ICLR, 2015.
[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[40] M. Tan, Q. Le, EfficientNet: rethinking model scaling for convolutional neural networks, in: 36th International Conference on Machine Learning, ICML, 2019, pp. 10691–10700.
[41] A. Sugeno, Y. Ishikawa, T. Ohshima, R. Muramatsu, Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning, Comput. Biol. Med. 137 (2021) 104795, https://doi.org/10.1016/j.compbiomed.2021.104795. ISSN 0010-4825.
[42] Z. Wu, G. Shi, Y. Chen, F. Shi, X. Chen, G. Coatrieux, J. Yang, L. Luo, S. Li, Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network, Artif. Intell. Med. 108 (2020) 101936, https://doi.org/10.1016/j.artmed.2020.101936. ISSN 0933-3657.
[43] X. Li, X. Hu, L. Yu, L. Zhu, C.W. Fu, P.A. Heng, CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading, IEEE Trans. Med. Imag. 39 (5) (2020) 1483–1493, https://doi.org/10.1109/TMI.2019.2951844.
[44] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block Attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018.
[45] J. de la Torre, A. Valls, D. Puig, A deep learning interpretable classifier for diabetic retinopathy disease grading, Neurocomputing 396 (2020) 465–476, https://doi.org/10.1016/j.neucom.2018.07.102. ISSN 0925-2312.
[46] T. Araújo, G. Aresta, L. Mendonça, S. Penas, C. Maia, Ângela Carneiro, A. M. Mendonça, A. Campilho, DR|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images, Med. Image Anal. 63 (2020) 101715, https://doi.org/10.1016/j.media.2020.101715. ISSN 1361-8415.
[47] Z. Tu, S. Gao, K. Zhou, X. Chen, H. Fu, Z. Gu, J. Cheng, Z. Yu, J. Liu, SUNet: a lesion regularized model for simultaneous diabetic retinopathy and diabetic macular edema grading, in: 2020 IEEE 17th International Symposium on Biomedical Imaging, ISBI, 2020, pp. 1378–1382, https://doi.org/10.1109/ISBI45749.2020.9098673.
[48] M. Tavakoli, R.P. Shahri, H. Pourreza, A. Mehdizadeh, T. Banaee, M.H. Bahreini Toosi, A complementary method for automated detection of microaneurysms in fluorescein angiography fundus images to assess diabetic retinopathy, Pattern Recogn. 46 (10) (2013) 2740–2753, https://doi.org/10.1016/j.patcog.2013.03.011. ISSN 0031-3203.
[49] A.S. Hervella, J. Rouco, J. Novo, M. Ortega, Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement, Procedia Comput. Sci. 126 (2018) 97–104, https://doi.org/10.1016/j.procs.2018.07.213. ISSN 1877-0509.
[50] 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.
[51] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, in: The IEEE International Conference on Computer Vision, ICCV, 2015.
[52] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations, ICLR, 2015.
[53] M.D. Abràmoff, J.C. Folk, D.P. Han, J.D. Walker, D.F. Williams, S.R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang, M. Lamard, D.C. Moga, G. Quellec, M. Niemeijer, Automated analysis of retinal images for detection of referable diabetic retinopathy, JAMA Ophthalmology 131 (3) (2013) 351–357, https://doi.org/10.1001/jamaophthalmol.2013.1743. ISSN 2168-6165.
[54] C. Zhu, R.H. Byrd, P. Lu, J. Nocedal, Algorithm 778: L-BFGS-B: fortran subroutines for large-scale bound-constrained optimization, ACM Trans. Math Software 23 (4) (1997) 550–560, https://doi.org/10.1145/279232.279236. ISSN 0098-3500.
[55] J. Cohen, A coefficient of agreement for nominal scales, educational and psychological measurement 20, 1960, pp. 37–46, https://doi.org/10.1177/001316446002000104, 1.
[56] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic Differentiation in PyTorch, NIPS Autodiff Workshop, 2017.
[57] P. Cao, F. Ren, C. Wan, J. Yang, O. Zaiane, Efficient multi-kernel multi-instance learning using weakly supervised and imbalanced data for diabetic retinopathy diagnosis, Comput. Med. Imag. Graph. 69 (2018) 112–124, https://doi.org/10.1016/j.compmedimag.2018.08.008. ISSN 0895-6111.
[58] F. J. Martinez-Murcia, A. Ortiz, J. Ramírez, J. M. Górriz, R. Cruz, Deep residual transfer learning for automatic diagnosis and grading of diabetic retinopathy, Neurocomputing ISSN 0925-2312, doi:https://doi.org/10.1016/j.neucom.2020.04.148, URL https://www.sciencedirect.com/science/article/pii/S0925231220316520.