Information Fusion 79 (2022) 146–161

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Retinal microaneurysms detection using adversarial pre-training with unlabeled multimodal images

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
Multimodal imaging
Eye fundus
Generative adversarial networks
Microaneurysms

ABSTRACT

The detection of retinal microaneurysms is crucial for the early detection of important diseases such as diabetic retinopathy. However, the detection of these lesions in retinography, the most widely available retinal imaging modality, remains a very challenging task. This is mainly due to the tiny size and low contrast of the microaneurysms in the images. Consequently, the automated detection of microaneurysms usually relies on extensive ad-hoc processing. In this regard, although microaneurysms can be more easily detected using fluorescein angiography, this alternative imaging modality is invasive and not adequate for regular preventive screening.

In this work, we propose a novel deep learning methodology that takes advantage of unlabeled multimodal image pairs for improving the detection of microaneurysms in retinography. In particular, we propose a novel adversarial multimodal pre-training consisting in the prediction of fluorescein angiography from retinography using generative adversarial networks. This pre-training allows learning about the retina and the microaneurysms without any manually annotated data. Additionally, we also propose to approach the microaneurysms detection as a heatmap regression, which allows an efficient detection and precise localization of multiple microaneurysms. To validate and analyze the proposed methodology, we perform an exhaustive experimentation on different public datasets. Additionally, we provide relevant comparisons against different state-of-the-art approaches. The results show a satisfactory performance of the proposal, achieving an Average Precision of 64.90%, 31.36%, and 33.55% in the E-Ophtha, ROC, and DDR public datasets. Overall, the proposed approach outperforms existing deep learning alternatives while providing a more straightforward detection method that can be effectively applied to raw unprocessed retinal images.

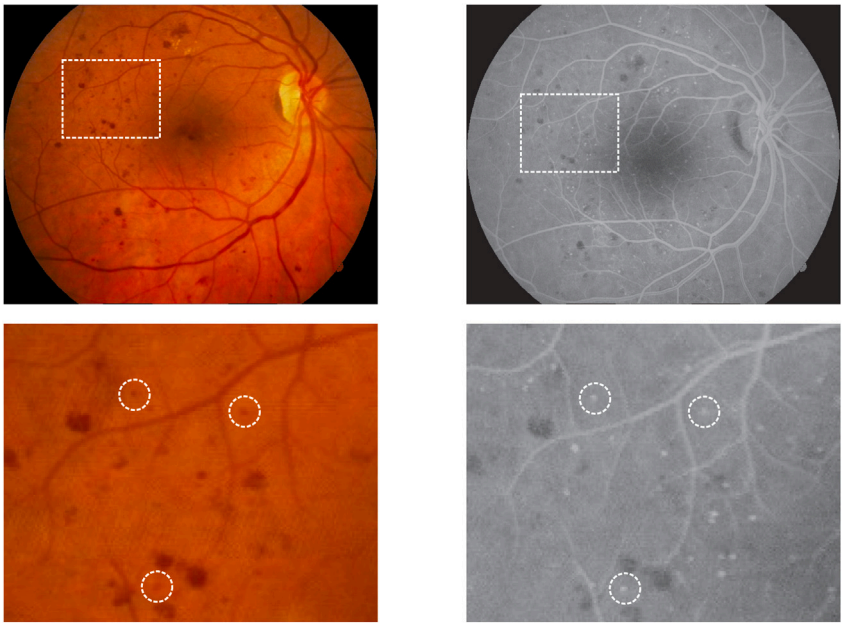
1. Introduction

The detection of microaneurysms (MAs) in eye fundus images is an important and challenging step towards the early diagnosis and prevention of important health conditions. In particular, MAs are small vascular lesions consisting in the swelling of capillaries due to weakened vascular walls [1]. Thus, retinal MAs can be associated with different ophthalmic and cardiovascular conditions [2]. For instance, retinal MAs have been shown to be a risk factor for strokes [3]. Additionally, MAs are the first typical sign of diabetic retinopathy (DR) [4], a retinal disease that represents the leading cause of blindness among the middle-aged population in the world [5]. Therefore, it is crucial to detect the disease at its earliest stages in order to prevent the progression and potential vision loss. However, the tiny dimensions of the MAs make their detection extremely challenging for both clinical experts and automated procedures.

Regarding the analysis of the eye fundus, color retinography is the most widely available retinal imaging technique. These color photographs of the eye can be obtained with affordable equipment and minimal inconvenience for the patients [6,7]. Thus, they are the ideal target imaging modality for screening programs and automated procedures in any healthcare service [7]. In retinography, MAs are shown as small blood-colored dots, typically placed near small vessels. The detection of these reddish lesions is challenging not only due to their small size but also due to the sub-optimal visual conditions. In this sense, MAs may present low contrast with respect to the background or may be affected by uneven illumination across the image. Moreover, MAs can also be confused with other structures in the images [8], such as microhemorrhages, pigmentation changes, or even dust particles in the camera.

* Corresponding author at: Centro de Investigación CITIC, Universidade da Coruña, A Coruña, Spain.
E-mail address: a.suarezh@udc.es (Á.S. Hervella).

https://doi.org/10.1016/j.inffus.2021.10.003
Received 11 February 2021; Received in revised form 8 September 2021; Accepted 9 October 2021
Available online 16 October 2021



(a) Color retinography (b) Fluorescein angiography

Fig. 1. Color retinography and fluorescein angiography for the same eye. Some representative examples of MAs are marked with circles in the zoomed area. In the retinography, different structures are depicted with the same reddish tone, whereas, in the angiography, the MAs are highlighted. In this sense, the angiography (b) can be used as a reference to differentiate the true MAs in the retinography (a). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In clinical practice, in order to overcome the limitations of color retinographies for the detection of MAs, clinicians can opt for injecting a blood contrast dye to the patients. This contrast dye, called fluorescein, produces the fluorescence of the blood and, therefore, can be used to produce a retinal photograph where the vascular structures and lesions are highlighted [9]. The resulting imaging modality, called fluorescein angiography, allows to study in detail the retinal vasculature and to easily visualize the existing MAs. Indeed, the contrast dye specifically produces the fluorescence of MAs avoiding those other structures of similar appearance in color retinography. This makes fluorescein angiography the clinical gold standard for the detection and study of MAs. However, due to the invasive nature of the procedure and potential hazards of the contrast dye, the technique is excluded from common screening programs and reserved for the study of high evidence cases [10]. Fig. 1 depicts some representative examples of MAs in both color retinography and fluorescein angiography.

Given the difficulties for detecting MAs in retinography, the development of automated methods is key towards facilitating successful screening programs of the population. Traditional approaches to automated MAs detection are characterized by the use of several ad-hoc processing stages [11]. In these cases, it is typical to perform the detection of numerous MAs candidates and then proceed to their classification based on a set of hand-engineered features. In contrast, recent works have adopted deep learning-based approaches where the engineering of features is not required and, instead, the required features are automatically learned from the data [12]. However, despite the potential of Deep Neural Networks (DNNs) for learning from the raw images, deep learning-based MAs detection methods still typically require the use of ad-hoc pre-processing steps [12,13]. Moreover, it is also common the necessity of several networks or aggregation mechanisms in order to produce refined predictions [12,14]. Thus, the straightforward and efficient use of DNNs for the detection of MAs remains a significant challenge.

The challenges for applying deep learning to MAs detection come from the very nature of the detection task and the necessity of sufficient annotated data from the training of the networks. Regarding the detection task, it is not always evident how it should be addressed using a DNN. For instance, in the literature, there are examples of different approaches, such as bounding box predictions [15] or the use of segmentation as a surrogate task for the detection [12,13]. The latter being at the cost of requiring significantly more expensive labels in the form of segmentation masks. In that sense, the annotation of MAs is a tedious and difficult task that must be performed by clinicians with expertise in the field. Even then, it is typical to see a high variability among the experts' annotations [8], which shows the difficulty of producing a reproducible analysis. Furthermore, the annotations for a detection task typically consist of a few pixel coordinates per image, which represent a scarce source of information for training DNNs.

In this work, we propose a novel methodology that aims at overcoming the main challenges of applying deep learning to the detection of MAs in retinography. In particular, we aim at providing a straightforward detection method able to process complete raw images in a single step without the necessity of pre-processing or patch-processing the data. For that purpose, we propose to formulate the detection of MAs as a heatmap regression, where a DNN is trained to predict a MAs heatmap (or likelihood map) from the input retinography. Then, the pixel coordinates of the detected MAs are obtained with minimal post-processing, only requiring the extraction of the local maxima (i.e. the points of maximum likelihood) in the predicted heatmaps. Additionally, to facilitate the training of this challenging task, we propose a novel adversarial multimodal pre-training that takes advantage of additional unlabeled multimodal images to learn about the characteristics of the MAs in a self-supervised fashion. Recently, a self-supervised multimodal pre-training consisting in the prediction of angiographies from retinographies has demonstrated to be useful for the later analysis of the retinal anatomy in retinography [16]. Given the condition of the angiography as the gold standard for the analysis of MAs, this kind of multimodal pre-training should be especially advantageous for MAs detection. However, the existing methodology proposed in [17,18] does not successfully learn to distinguish all the small structures in the retina. In fact, this multimodal reconstruction typically fails to recognize the MAs, which are ignored or treated as small microhemorrhages in the generated angiographies. In order to improve the recognition capacity of the multimodal reconstruction, we

propose a novel methodology that is based on the use of an adversarial framework. This novel methodology provides better recognition of the MAs, facilitating their distinction among other similar structures. As a consequence, the proposed adversarial multimodal pre-training allows a successful detection of the MAs in the raw retinographies.

In summary, the methodology proposed in this work for the detection of MAs in retinography presents two important novelties. First, to the best of our knowledge, this is the first work using unlabeled multimodal images as well as generative adversarial networks (GANs) for pre-training the detection of MAs. Second, in contrast to previous works, the detection of MAs is approached as a heatmap regression task using DNNs. In order to demonstrate the advantages of the proposed methodology, we perform an exhaustive experimentation on several public datasets including different scenarios. Additionally, the experimentation includes relevant comparisons against state-of-the-art methods for the detection of MAs and the pre-training of retinal image analysis algorithms.

1.1. Related work

Several approaches have been proposed in the literature for the automated detection of retinal MAs [11]. However, the earliest works addressing this task were applied to angiography instead of retinography [19]. Fluorescein angiography represents the gold standard for the visualization of retinal MAs and, therefore, it was also considered as a natural first step for the automated detection of MAs. These first works applied to angiography followed a classical pipeline consisting of image pre-processing, candidate extraction, and refinement [19,20]. Given the similarities between retinography and angiography, the same pipeline was later adapted for the detection of MAs in retinography [21]. In this regard, the detection of MAs in retinography represents a more challenging task due to the much lower contrast of the MAs in the images. However, the shape characteristics of the MAs remain the same in both imaging modalities.

Regarding the use of the classical pipeline in retinography, the sub-optimal visual conditions of the MAs place great importance on the pre-processing of the images. In this regard, a common approach throughout the literature is to only consider the green channel of the retinography [22–25], where the blood containing structures are more contrasted [26]. Then, different normalization techniques are applied to improve the contrast of the MAs [13,24,27,28], whereas the noise is typically reduced by filtering the image [12,25,28]. With regard to the candidate extraction step, a variety of techniques have been proposed. For instance, Morlet wavelets [22], first-order Gaussian derivative filters [23], sliding band filters [25], or Multi-scale Gaussian Correlation Coefficients [28,29] have been successfully applied in conjunction with different thresholding techniques. However, in these cases, the extraction of the final MAs candidates still requires further processing to remove spurious detections, such as the extraction of the blood vessels [28] or filtering the candidates by shape characteristics [23]. In contrast, Veiga et al. [27] directly extract the MAs candidates via pixel-wise classification using a Support Vector Machine (SVM) with Laws texture features. Given that all these techniques result in a high number of false detections, a classification or refinement stage of the MAs candidates is required. In this final step, the most successful approaches are those based on machine learning techniques. In this regard, the typical approach requires the extraction of handcrafted features and the subsequent training of a classifier. Regarding the feature extraction, shape and intensity features are the most commonly used [24,25,27,28]. Although features based on texture [27] or local convergence index filters [23] have also been successfully applied. For the classification, different algorithms have been explored, such as SVM [27] or RUSBoost [23–25], which is a sampling/boosting algorithm for learning from unbalanced data. Additionally, the use of sparse principal component analysis for improving the classification has also been proposed [28]. In contrast to all these approaches, Javidi et al. [22] avoid the extraction of handcrafted features by using Fisher discrimination dictionary learning.

Recently, several deep learning-based methods have also been proposed. Although the use of DNNs is still not as common as in the segmentation of MAs [7,30]. In this regard, while the training and inference procedures for the segmentation task are well defined, that is not the case for the detection counterpart, which remains an especially challenging issue. Regarding the detection of MAs, despite the ability of DNNs for learning complex patterns from the raw data, the pre-processing of the input images is still present in some of the most successful deep learning-based approaches [12,13]. Regarding the methodological proposals for MAs detection, different alternatives have been explored. For instance, the MAs detection has been successfully approached as a patch-level classification task, for which convolutional neural networks with fully connected layers are required [14]. In this case, Savelli et al. [14] proposed an ensemble of networks that are applied on patches of different dimensions. In contrast, other works [12,13] approached the MAs detection as a segmentation task, which is performed with a fully convolutional network. However, in [12], the images are also processed by patches, given that the final prediction for each pixel is obtained by aggregating the response in all the possible patches containing that pixel. Additionally, the use of patches during training allows to balance the data distribution, correcting the reduced number of MAs in the training set [12,14]. In this regard, another possibility to address the unbalanced training data is to manipulate the loss functions, using e.g. weighted cross-entropy or focal loss [13]. Finally, a third methodological alternative in the literature is to perform a bounding box prediction task, which is a common approach for object detection. In this case, Li et al. [15] explored the use of several state-of-the-art deep learning methodologies for object detection. However, these methodologies resulted in low performance for the detection of MAs, which evidences the difficulty of the problem at hand.

In comparison to previous deep learning methodologies, we approach the MAs detection as a heatmap regression. This allows an easy extraction of the precise MAs locations, which are represented by the local maxima in the predicted heatmaps. In contrast, other comparable approaches such as segmentation [12], despite requiring more expensive labels for training, do not naturally provide the precise location of the MAs. Additionally, the proposed methodology also aims at producing a more straightforward and efficient use of DNNs in comparison to other successful approaches in the literature. This is possible due to the proposed adversarial multimodal pre-training, which successfully leverages the domain-specific knowledge provided by the unlabeled angiographies. In this regard, the proposed methodology adds complexity to the training phase, which is performed only once, to facilitate the detection of MAs at inference time. As a result, the proposed approach avoids both the use of pre-processing and the necessity of patch-processing the images.

The rest of the manuscript is structured as follows. In Section 2, the proposed methodology is described, including (Section 2.1) the adversarial multimodal pre-training, (Section 2.2) the MAs detection using the heatmap regression approach, and (Section 2.3) the network architectures. Section 3 comprises the experiments, results, and their corresponding discussion. First, this section describes (Section 3.1) the datasets and (Section 3.2) the evaluation procedure. Then, several analyses and comparisons are provided, including (Section 3.3) a qualitative analysis and comparison of the pre-training, (Sections 3.4–3.6) complete analyses and comparisons of the MAs detection, and (Section 3.7) an evaluation of the proposed approach for the detection of DR. Limitations and future works are discussed in Section 3.8. Finally, conclusions are drawn in Section 4.

2. Methodology

The automated detection of MAs in retinography is approached as a heatmap regression using DNNs. Following this approach, the

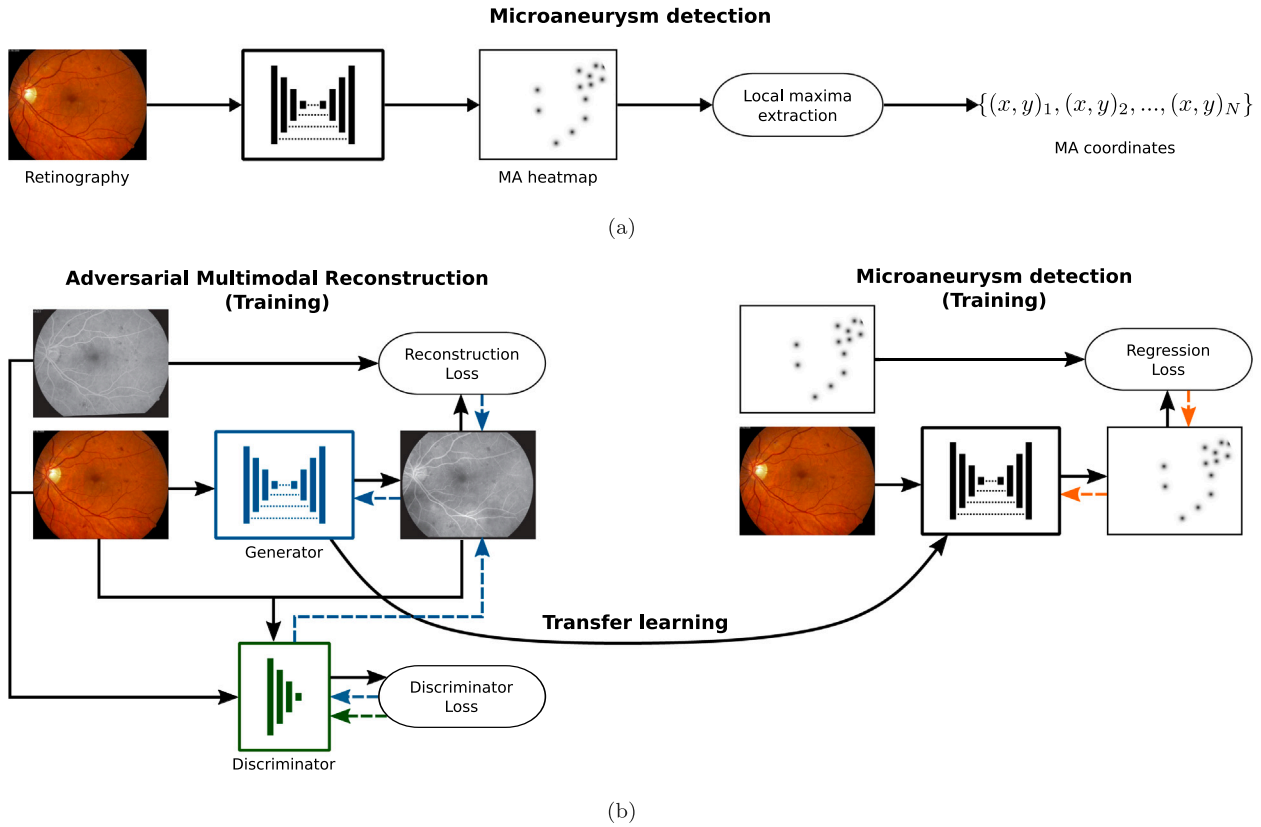


Fig. 2. Methodology for the automated detection of MAs in retinography. (a) MAs detection via heatmap regression. (b) Training procedure for the MAs detection network. Firstly, the network is pre-trained using the adversarial multimodal reconstruction. Then, the same network is fine-tuned for the prediction of the MAs heatmaps.

specific coordinates of the detected MAs will be those corresponding to the local maxima in the predicted heatmaps. In order to facilitate the training of a DNN in this challenging task, the MAs detection network is previously pre-trained in a self-supervised fashion using a novel adversarial multimodal reconstruction approach. The objective of this self-supervised pre-training is to provide the network with relevant knowledge for the detection of MAs before using any annotated data. For that purpose, fluorescein angiography, which is the clinical gold standard for MAs analysis, is used as target image modality in the adversarial multimodal reconstruction. A graphical summary of the proposed methodology for the detection of MAs is depicted in Fig. 2.

2.1. Adversarial multimodal reconstruction

The multimodal reconstruction of angiography from retinography is trained using an adversarial framework consisting of two different networks: the generator G that learns to predict the corresponding angiography from the input retinography and the discriminator D that learns to distinguish between predicted and real angiographies. The training is performed with a paired set of images, where for each retinography there is also available an angiography of the same eye. In order to better take advantage of the paired data, the multimodal image pairs are registered, resulting in a pixel-wise correspondence between the retinography and the angiography of the same eye. The registered image pairs allow the use of pixel-wise distance metrics between the generated and target angiographies, as well as the use of a discriminator network that is pixel-wise conditioned with the corresponding input retinography. The registration of the images is automatically performed by applying the methodology proposed in [31].

In contrast with the approach proposed in [18], in this work we combine both reconstruction and adversarial loss functions for training the generator network. The objective of adding an adversarial loss is to provide additional complementary feedback to the generator. In that sense, while the reconstruction loss has demonstrated to provide high quality feedback for successfully reconstructing most of the retinal structures, the provided feedback is still solely based on low level characteristics of the images. The addition of an adversarial loss function, parameterized by a DNN, results in additional feedback that can be related to both high and low level characteristics of the images. Thus, this combined approach has the potential to improve the recognition of the different retinal structures, especially those that are more challenging such as the MAs. The combined loss function for training the generator is defined as:

$$\mathcal{L}_G = \mathcal{L}_G^{Rec} + \lambda \mathcal{L}_G^{Adv} \tag{1}$$

where $\mathcal{L}_G^{Rec}$ denotes the reconstruction loss, $\mathcal{L}_G^{Adv}$ the adversarial loss, and λ is a hyperparameter that balances the contribution of the two different losses.

For the reconstruction loss, we use the negative Structural Similarity (SSIM) [32], which has been previously proposed for the multimodal reconstruction, demonstrating a superior performance in comparison to other common metrics [18]. In particular, SSIM combines into a single metric the intensity, contrast, and structural similarities between two images. The SSIM value for a pair of pixels (x, y) is obtained using a set of local statistics as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2}$$

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. These local statistics are computed for each pixel by weighting its neighborhood with a Gaussian window of $\sigma = 1.5$ [32]. The constant values C_1 and C_2 are included to avoid instability when the denominator terms are close to zero. Following the definitions given by Wang et al. [32] and considering

that the image values are bounded between 0 and 1, we use $C_1 = 0.0001$ and $C_2 = 0.0009$.

Then, for a certain input retinography $\mathbf{r}$ and the corresponding target angiography $\mathbf{a}$, the reconstruction loss for the generator is obtained as:

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

where N denotes the number of pixels in the images.

For the adversarial loss, we use a least squares approach [33] where the targets of the discriminator are 1 for real samples and 0 for generated samples. Thus, the target of the generator is to produce a value of 1 in the discriminator output. The adversarial loss for the generator is obtained as:

$$\mathcal{L}_G^{Adv} = \frac{1}{M}\sum_{m}^{M}(\mathbf{D}(\mathbf{G}(\mathbf{r}), \mathbf{r})_m - 1)^2 \tag{4}$$

where M denotes the size of the discriminator output.

Simultaneously, the discriminator is trained on both the real angiographies and the fake angiographies that are estimated by the generator. Thus, the loss for the discriminator is obtained as:

$$\mathcal{L}_D = \frac{1}{M}\sum_{m}^{M}\left[(\mathbf{D}(\mathbf{G}(\mathbf{r}), \mathbf{r})_m)^2 + (\mathbf{D}(\mathbf{a}, \mathbf{r})_m - 1)^2\right] \tag{5}$$

As defined in Eqs. (4) and (5), both the real/fake angiography and the corresponding original retinography are inputs to the discriminator. This conditioning of the discriminator on the input retinography is possible due to the use of paired data. The objective is to provide additional information for learning the differences between real and estimated angiographies.

Finally, the generator training is performed by minimizing the combined generator loss, particularly solving:

$$G^* = \arg\min_G \mathcal{L}_G^{Rec} + \lambda \mathcal{L}_G^{Adv} \tag{6}$$

whereas, simultaneously, the discriminator training is performed by minimizing the discriminator loss, solving:

$$D^* = \arg\min_D \mathcal{L}_D \tag{7}$$

The hyperparameter λ in Eqs. (1) and (6) is commonly used in the literature for balancing adversarial and non-adversarial losses, typically reducing the strong feedback that is provided by the adversarial component [34]. The value of this hyperparameter is selected empirically attending to the training stability and the desired effect in the generated images. In particular, in this work, we aim at improving the recognition of challenging retinal structures such as the MAs. On the basis of previous works [34], we considered the candidate values 1, 0.1, and 0.01 in this order of preference. Finally, we empirically selected $\lambda = 0.1$, which provided stable training while producing the desired effect in the generated angiographies.

The optimization for both generator and discriminator is performed using the Adam algorithm [35]. The decay rates of Adam are set to $\beta_1 = 0.5$ and $\beta_2 = 0.999$, which are common values in the literature for the training of generative adversarial networks [36]. The learning rate is set to $\alpha = 5e{-}5$, which is the largest value allowing a stable training. The batch size is one image and the training is stopped when there are signs of overfitting for both the reconstruction and the adversarial loss of the generator. This is monitored over the learning curves using a 25% of the available training data as validation images. The parameters of the generator and the discriminator are initialized with a zero-centered normal distribution following the method proposed by He et al. [37].

To delay the onset of overfitting for both generator and discriminator, data augmentation is applied to the retinography and angiography images. The data augmentation consists of both spatial and intensity/color augmentations. In particular, as spatial augmentation, we use random affine transformations that are commonly applied in biomedical image analysis [38], including scaling, rotation, and shearing. To keep the pixel-wise correspondence between retinography and angiography, the same affine transformation is applied to both images within each multimodal pair. As color augmentation for the retinography, we use random transformations of the different image channels in HSV color space, similar to those proposed in [18]. Similarly, we also apply random transformations to the intensity value of the angiography images.

2.2. Microaneurysms detection

The MAs detection using a DNN is approached as a heatmap regression. In order to train the detection network in this regression task, the ground truth annotations provided by the clinical experts are transformed into the target heatmaps. These heatmaps are able to represent the location of multiple MAs and provide a certain heuristic to guide the training of the network. In particular, the exact location of the MAs will correspond to the maximum values in the heatmap and progressively decreasing values will be assigned to the surrounding pixels. This distribution of the pixel values in the heatmaps automatically increases the feedback that is provided to the network without unnecessarily increasing the annotation effort. With regards to the generation of the targets heatmaps, two steps are required. This procedure is depicted in Fig. 3. First, the provided ground truth pixel coordinates are used to generate a binary map where only the exact locations of the MAs have a non-zero value. Then, the target heatmaps are generated by involving the binary maps with an isotropic kernel of convex and monotonic decreasing kernel profile. In particular, within this family of kernels, we use a Radial Hyperbolic Tangent (Radial Tanh) kernel. This type of kernel, which is depicted in the diagram of Fig. 3, presents a very sharp profile that facilitates the precise localization of the local maxima in the heatmaps. In this regard, in comparison to a Gaussian alternative, the Radial Tanh kernel has demonstrated to provide a more stable performance against changes in the kernel size [39]. This is an important advantage that avoids the necessity of exhaustively searching for the most adequate size. The Radial Tanh kernel is defined as:

$$K(x, y; k) = 1 + tanh\left(-\frac{\pi\sqrt{x^2 + y^2}}{k}\right) \tag{8}$$

where (x, y) are the pixel coordinates with respect to the kernel center and k is the saturation distance for the kernel. This saturation distance is directly related to the kernel size and it allows to control the region of influence for each MA in the heatmap. Then, the target heatmap is obtained as:

$$\mathbf{y}^{heat} = \mathbf{K}(k) * \mathbf{y}^{bin} \tag{9}$$

where $*$ denotes two-dimensional convolution and $\mathbf{y}^{bin}$ is the binary map resulting from the direct mapping of the ground truth pixel coordinates.

For training the network, the Mean Squared Error (MSE) between the network output and the target heatmaps is used as loss function. Thus, the training loss is defined as:

$$\mathcal{L} = \frac{1}{N}\sum_{n}^{N}(\mathbf{F}(\mathbf{r})_n - \mathbf{y}_n^{heat})^2 \tag{10}$$

where $\mathbf{r}$ denotes the input retinography and $\mathbf{F}$ the transformation given by a DNN that generates the predicted heatmaps.

Finally, after the network training, the precise locations of the detected MAs can be easily recovered from the predicted heatmaps by extracting the local maxima. For this purpose, we use a maximum filter and an intensity threshold, which allows the calibration of the method to the desired operating point.

For the experiments in this work, the saturation distance of the kernel $\mathbf{K}$ is fixed to $k = 10$. This value is established according to

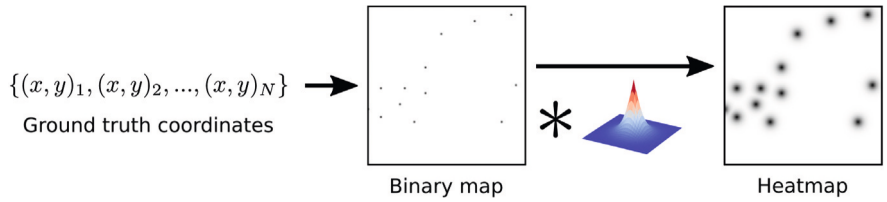


Fig. 3. Procedure for the automated generation of the target heatmaps from the ground truth pixel coordinates. The intermediate binary map is convolved with a Radial Hyperbolic Tangent kernel.

the size of the retinal images and the experimental results reported in [39]. During the training, the optimization is performed using the Adam algorithm [35] with decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which were proposed by the authors in [35]. The batch size is one image. The initial learning rate is set to $\alpha = 1e-4$, being reduced by a factor of 10 when the validation loss does not improve for 10 epochs. The training is early stopped after 20 epochs without any improvement in the validation loss. These values were empirically established according to the evolution of the learning curves. For this purpose, 25% of the available training data is used as validation data. To reduce the overfitting, we use data augmentation in the form of random spatial and color augmentations. In particular, the spatial augmentations are affine transformations and the color augmentations are channel-wise transformations performed in HSV color space, similar to those proposed in [18]. The network parameters are initialized with the optimized values obtained from the previous adversarial multimodal pre-training.

To avoid overfitting, we apply the same data augmentation that was used in the adversarial multimodal reconstruction. In particular, as spatial augmentation, we use random affine transformations that are simultaneously applied to the input retinography and the ground truth coordinates of the MAs. As color augmentation for the retinography, we use random transformations of the different image channels in HSV color space, similar to those proposed in [18].

2.3. Network architectures

Regarding the network architecture for MAs detection, it is important to notice that the proposed approach does not depend on any particular network design. Thus, in order to validate the proposal, we use a standard U-Net architecture [40]. This commonly used network design provides a well-known reference point in terms of performance, especially in medical image analysis. In this regard, this same network has been successfully applied in numerous medical domains [41], also including the analysis of retinal images [16]. U-Net is a fully convolutional DNN with a symmetric encoder–decoder structure. The main characteristic of U-Net is the use of skip connections between the intermediate layers of the encoder and the decoder. The structure of the network and the details of the different layers can be seen in the diagram of Fig. 4(a). The network presents several convolutional blocks consisting of two consecutive convolutions with 3×3 kernels and ReLU activation functions. In the encoder, there is a max pool operation after each convolutional block to downsample the internal feature maps whereas in the decoder there are transposed convolutions before each block to upsample the feature maps. Following the original U-Net design [40], the network performs four downsampling steps in the encoder and four upsampling steps in the decoder. This number of internal scales has been successfully tested with retinal images of a similar resolution in previous works [16]. The last layer presents a linear activation function, which is adequate for both the MAs heatmap regression in the target task and the multimodal reconstruction in the pre-training phase.

In order to apply the proposed pre-training to the MAs detection network (or generator), an additional neural network to act as discriminator in the adversarial setting is required. In general terms, this discriminator should be a fully convolutional DNN with an encoder structure, which allows the use of arbitrarily sized input images. Thus, for convenience, we use a discriminator network with the same structure that the encoder in the generator. The characteristics of the different layers are adapted according to the architecture guidelines proposed in [36] for the training of deep convolutional generative adversarial networks. The structure of the network and the details of the different layers can be seen in the diagram of Fig. 4(b). Specifically, the ReLU activation functions were changed for Leaky ReLU activations and the max pooling operations for strided convolutions. These modifications aim at improving the back-propagation of gradients during the training [36]. An additional empirical adaption is a reduction by half of the number of channels. This is motivated by an unstable training with the original number of channels in the discriminator, which seemed to be due to an excessive capacity of the network.

Due to the use of fully convolutional networks for both the generator and discriminator, the proposed approach can be applied to images of arbitrary size. Nevertheless, for the experiments in this work, the images are resized so that all of them depict the retina at the same scale. In particular, we use a standard retinal width of 720 pixels.

3. Experiments and results

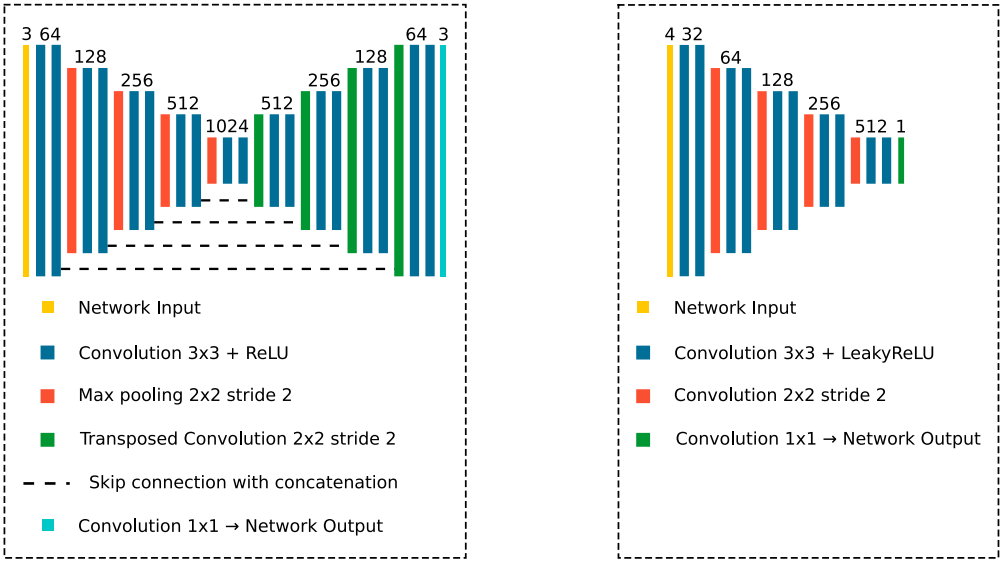
3.1. Datasets

For the adversarial multimodal reconstruction, we use the public multimodal dataset provided by Isfahan MISP [42]. This dataset is composed of 59 image pairs consisting of retinography and angiography of the same eye. Half of the image pairs correspond to patients diagnosed with DR whereas the other half correspond to healthy individuals. The size of the images is 720×576 pixels.

For the MAs detection, we use the public datasets of reference E-Ophtha [43], ROC [8], and DDR [15]. The E-Ophtha dataset consists of 381 images and the provided MAs annotations are in the form of segmentation labels. In total, 148 images have at least one MA, whereas the remaining 233 do not present any MA. The images present different sizes, ranging from $1{,}440 \times 960$ to $2{,}544 \times 1{,}696$ pixels.

The ROC dataset includes 50 images and the provided annotations consist of the pixel coordinates and the estimated radius of the MAs. In total, 37 images have at least one MA, whereas the remaining 13 do not present any MA. The images present different sizes, ranging from 768×576 to $1{,}389 \times 1{,}383$ pixels. Regarding this dataset, there exist 50 additional *test* images for which no ground truth is available. Although some works in the literature evaluated their methods on these *test* images while the online evaluation was possible. Besides consisting of different images, ROC *test* also uses completely different criteria for the reference standard in the evaluation [8]. Thus, the experiments and comparisons in this work are exclusively performed on the publicly available ROC *training* set.

The DDR dataset consists of 757 images and the annotations are provided in two different forms, as segmentation labels and as bounding boxes. The images correspond to different grades of the disease, ranging from mild to proliferative DR. Additionally, the images present varying sizes. The dataset is split by default into 50% training, 20% validation, and 30% test subsets.



(a) Generator (b) Discriminator

Fig. 4. Diagrams of the neural networks. The numbers above the different blocks indicate the number of channels.

In order to have the eye fundus in all the images at the same scale, the images are resized to match the scale of the multimodal Isfahan MISP dataset, i.e. a retinal width of 720 pixels. The ground truth annotations regarding the location and size of the MAs are rescaled accordingly.

For evaluating the detection of DR, we use the public datasets E-Ophtha [43] and Messidor [44]. In both cases, we follow the approaches that are common in the related literature [23,25]. For instance, in the case of E-Ophtha, the images with at least one MA are considered pathological whereas the images with no MAs are considered healthy.

The Messidor dataset consists of 1200 retinal images categorized into four different DR grades, one of them healthy (grade 0) and three pathological (grades 1, 2, and 3). In this case, the evaluation of MAs detection algorithms for the detection of DR is typically performed using only grades 0 and 1, given that grades 2 and 3 contain additional lesions besides the MAs [23,25]. However, in our experiments, we consider both the "0 vs 1" and "0 vs Rest" (i.e. 0 vs 1,2,3) settings. This allows to also study the robustness of the method, including more complex scenarios.

3.2. Evaluation metrics

For the quantitative evaluation, we apply the same gold standard that is defined in [8], which is a common reference in the field. In particular, following this approach, the predicted MAs are matched one-to-one with the ground truth MAs. This means that each ground truth MA can only be detected by a single prediction and, simultaneously, each prediction can only detect a single ground truth MA. In this line, a predicted MA is considered a True Positive (TP) if it is located within a distance d of a still undetected ground truth MA and a False Positive (FP) otherwise. In case of several ground truth MAs within the corresponding range d, the closest one is considered as the detected MA. Finally, the ground truth MAs that remain undetected are considered False Negatives (FN). Then, TP, FP, and FN measures are used to compute Precision and Recall (or Sensitivity), which are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

In order to study the performance independently of the specific operating point, the described metrics are computed for different detection thresholds. The resulting sets of values are used to perform Precision–Recall (PR) and Free-response Receiver Operating Characteristic (FROC) analyses. The PR analysis is the most common approach to evaluate detection algorithms in computer vision. Besides depicting the PR curve, the performance is summarized into the Average Precision (AP), which is computed as the area under the PR curve. In contrast, the FROC analysis is only used for those unbalanced problems where the negative class is much more numerous than the positive class. This is the case of MAs detection and, therefore, FROC analysis is broadly applied in the related literature. The FROC curves represent the sensitivity for different levels of FPs per image (FPI). Additionally, the performance is typically summarized into the Score, which is obtained by averaging the sensitivities for certain representative FPI values. Typically, in the MAs detection, the Score is computed for all or part of the set of values $FPI_{Score} = \{0.125, 0.25, 0.5, 1, 2, 4, 8, 12, 16, 20\}$. We will use all the values in FPI_{Score} unless stated otherwise.

Regarding the gold standard, the distance threshold d that is used to define the TPs depends on the size of the MAs. Particularly, the ROC public dataset directly provides this value for each annotated MA. However, that is not the case for the E-Ophtha and DDR public datasets. In these cases, we define d as the diameter of the circle that has the same area as the specific MA. The aim of this approximation is to facilitate the comparison with those works that use the overlapping area, instead of the distance, as criteria for the TPs. This would be the case of those works that, using segmentation or bounding box prediction as a surrogate for the detection, do not explicitly provide the pixel coordinates of the MAs. In these cases, the same rigorous gold standard that we follow could not be applied.

The evaluation for the detection of DR is performed using Receiver Operating Characteristic (ROC) analysis, which is commonly used for binary classification problems. In particular, we assess the performance by computing the Area Under the ROC curve (AUROC).

3.3. Multimodal reconstruction

The objective of the adversarial multimodal reconstruction (AdvMR) proposed in this work is to improve the recognition ability of the pre-trained networks. This will be evaluated by means of their performance in the downstream target task, i.e., the MAs detection.

However, we also provide a qualitative evaluation regarding the multimodal reconstruction, which may facilitate the comprehension of the results that were obtained for the target MAs detection.

Fig. 5 depicts representative examples of predicted angiographies together with the original retinographies and angiographies. Additionally, the comparison also includes the predicted angiographies obtained with the methodology proposed in [18] (MR), which represents the current state-of-the-art approach for self-supervised multimodal pre-training in medical image analysis [16].

Regarding the predicted angiographies, both methodologies demonstrate the ability to recognize most of the retinal structures and to perform an adequate multimodal transformation. In particular, both approaches show an excellent performance regarding the recognition of the main anatomical structures of the retina, namely the fovea, optic disc, and blood vessels. Additionally, bright lesions in the retinography, which should not be visible in the angiography, are successfully removed in the predicted angiographies using any of the two approaches. However, important differences arise when the treatment given to the red lesions in the retinography is analyzed. The two main types of red lesions in the retinography are the MAs and the microhemorrhages. Although these two types of lesions present a similar color and intensity in the retinography, their appearance is completely different in the angiography. In particular, in the angiography, the microhemorrhages are depicted as dark lesions, keeping a low intensity value, whereas the MAs are depicted as bright lesions, getting a high intensity value like that of the blood vessels. The depicted examples show that the approach proposed in this work, AdvMR, is able to distinguish these two kinds of red lesions and to predict angiographies containing both bright lesions, i.e. MAs, and dark lesions, i.e. microhemorrhages. In contrast, the MR approach only generates dark lesions in the predicted angiography, which means that the MAs are ignored or treated as small microhemorrhages. Thus, AdvMR represents a significative improvement over the current state-of-the-art approach for self-supervised multimodal pre-training, at least by means of the generated images.

Finally, regarding the appearance of the predicted angiographies, although both approaches generate images that resemble real angiographies, the images generated using AdvMR present a more realistic appearance. This is mainly due to two different factors. One is the presence of MAs in the predicted angiographies, which is a distinctive characteristic of the angiographies for those cases with DR. The other factor is the background texture generated by the AdvMR approach, which resembles better the retinal background in real angiographies. This shows that the addition of an adversarial loss in the proposed methodology contributes with both high and low level feedback during the training.

3.4. Microaneurysm detection

To perform a comprehensive analysis of the approach that is proposed in this work, we compare the performances of three different methodologies: (1) the MAs heatmap regression using the adversarial multimodal pre-training (AdvMR), which is the methodology proposed in this work, (2) the MAs heatmap regression training from scratch (FromScratch), and (3) the MAs heatmap regression using the multimodal pre-training that was proposed in [16] for other related tasks in the same domain (MR). These three methodologies are exhaustively evaluated using the E-Ophtha and ROC public datasets. In particular, the E-Ophtha evaluation follows a 5×2 cross-validation approach consisting of 10 experiments with 10 different training–test splits. In the case of the ROC dataset, all the available data is used as test set. However, the variability regarding the training data is still studied by performing 10 evaluations with the 10 different detection networks resulting from the E-Ophtha experiments.

In particular, the results of the quantitative evaluation are depicted in Fig. 6. The depicted charts show the individual curves resulting from the different experiments that are performed. The curves are not aggregated because, for some of the methods, the obtained results are far from being normally distributed. Thus, the result of aggregating the curves by computing their mean and standard deviation would not be representative of the true underlying performance. Regarding the analysis of these charts, it is observed that the best performance is always achieved by AdvMR. In contrast, the worst performance corresponds to the networks that are trained from scratch. In particular, according to the quantitative results, these networks completely fail in the detection task. This shows how challenging is to successfully detect MAs using a straightforward deep learning approach as the one intended in this work. In this regard, the difficulties for learning to detect MAs are mainly due to the tiny size and low contrast of these structures, as well as the heavily unbalanced training data. Nevertheless, the difficulties are overcome by applying the whole methodology consisting in the heatmap regression together with the proposed adversarial multimodal pre-training.

Regarding the comparison between AdvMR and MR approaches, AdvMR not only results in higher values for the considered metrics but it also offers a much lower variability in the performance. In this regard, the MR pre-training demonstrates to be able to produce a performance close to that of AdvMR in some experiments but, in others, the performance is significantly reduced. In these cases, the networks producing low performance are the same for both the E-Ophtha and ROC datasets, which indicates that this effect is due to the different training data among the 10 experiments. These results show that some training samples provide better feedback than others for learning the task and, therefore, the distribution of the training samples is a relevant factor to consider when training the networks. Nevertheless, the proposed approach, AdvMR, does not show to be affected by this variability in the training data. This means that the pre-trained network weights provided by AdvMR are able to overcome a less optimal training data distribution.

Representative examples of predicted MAs heatmaps are depicted in Fig. 7. It is observed that the MAs heatmaps predicted by AdvMR are the ones that best fit the ground truth annotations. In this regard, as previously stated, MR provides satisfactory results in some experiments (see Fig. 7, 1st and 2nd examples), but in others the performance deteriorates drastically (see Fig. 7, 3rd example). With regards to the training from scratch, the networks generate a blank heatmap with no MAs. It must be noticed that this wrong solution still significantly minimizes the training loss given that most of the training data correspond to non-MAs, i.e., blank regions. To ensure that this is the typical outcome of the networks trained from scratch, we performed some additional experiments with extended training times but the networks still converged to the same blank solution.

Finally, it should be noticed that, according to the multimodal reconstruction results (Section 3.3), the MR pre-trained networks do not recognize the MAs. However, despite this fact, the MR pre-training still significantly improves the training from scratch. This indicates that general knowledge about the main structures of the retina, even if the MAs are not included, is helpful for improving the detection of MAs. In any case, as it was expected, the knowledge about MAs provided by AdvMR facilitates even more the MAs detection task, outperforming the results.

3.5. Effect of class balancing the loss function

The proposed methodology, consisting in the MAs heatmap regression together with the adversarial multimodal pre-training, has demonstrated to successfully address the MAs detection. However, the provided experimentation has shown that the MAs heatmap regression alone fails to solve the task and, therefore, the proposed AdvMR is a key element towards achieving a straightforward and efficient detection of MAs using deep learning. In the literature, however, there are some previous works that obtain a satisfactory performance using deep learning without this novel pre-training. A notable difference in

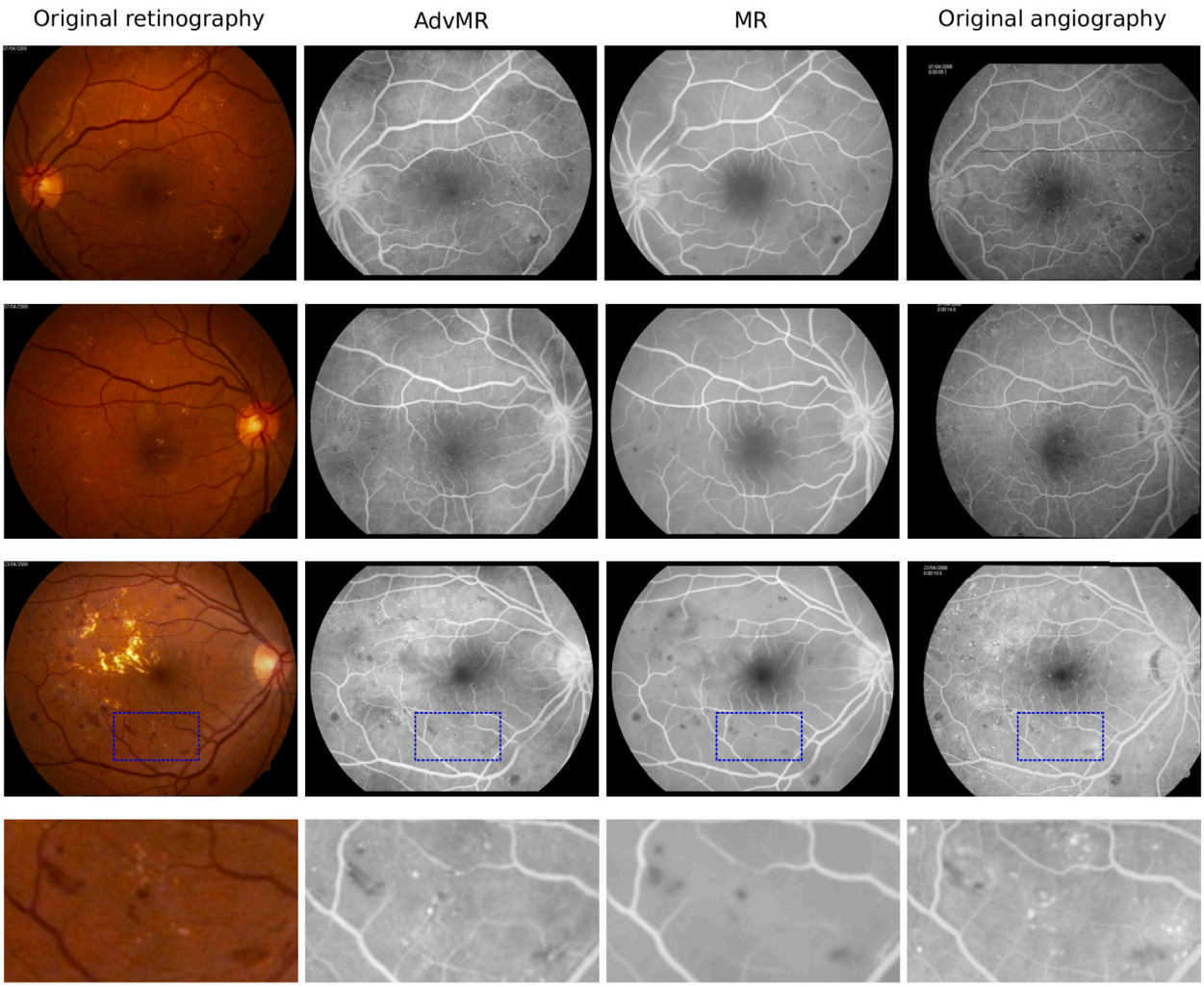


Fig. 5. Representative examples of generated angiographies and comparison between the proposed adversarial multimodal reconstruction (AdvMR) and the multimodal reconstruction previously proposed in [18] (MR). The 4th row depicts zoomed regions from the images in the 3rd row. The images correspond to patients diagnosed with DR.

some of these works is the use of segmentation labels, which provide more training feedback at the cost of increasing the annotation effort. In our proposal, however, the training feedback is also increased by using the MAs heatmaps, which are much more efficient regarding the required annotations. Other differences that could explain the adequate performance of the previous approaches include the use of custom network architectures and processing pipelines, pre-processing of the input images, or class-balancing the task during training. Given the reduced number of MAs in the images, we argue that balancing the task can be a key element towards facilitating the training and avoiding the convergence to the blank solution. In this regard, we performed additional experiments to confirm the hypothesis.

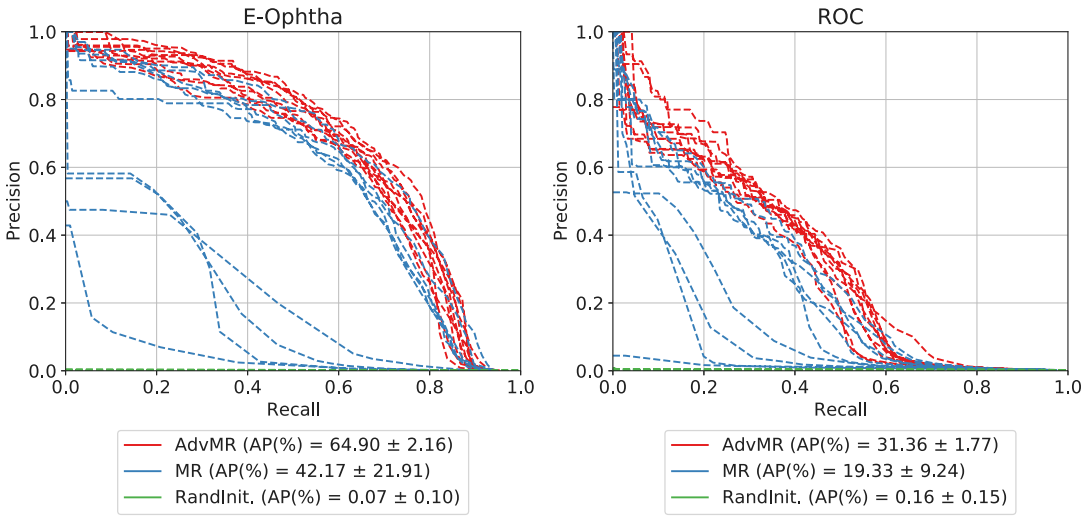
In order to study the effect of class-balancing the detection task during the training, we opt for applying a balanced loss function. In particular, a weight map is computed for each training image containing MAs and it is used to weight the losses at the pixel level. For this weight map, the MA class is defined as each MA location and its neighborhood. The size of the neighborhood is given by the kernel size that was used to create the heatmaps, i.e., a circle with a radius of 10 pixels in this case. The objective of the weight map is to cancel the effect of the unbalanced number of pixels between classes while minimally affecting other factors in the experimentation. In particular, the selected weighting scheme ensures that, if both classes present the same mean error, both classes will have the same importance in the total loss of the image and the importance of the individual image will remain unchanged. For each image $\mathbf{g}$ containing MAs, the weight map $\mathbf{w_g}$ is defined as:

$$w_g(x,y) = \begin{cases} \frac{P+N}{2P} & \text{if } g(x,y) \text{ is MA} \\ \frac{P+N}{2N} & \text{if } g(x,y) \text{ is non-MA} \end{cases} \tag{13}$$

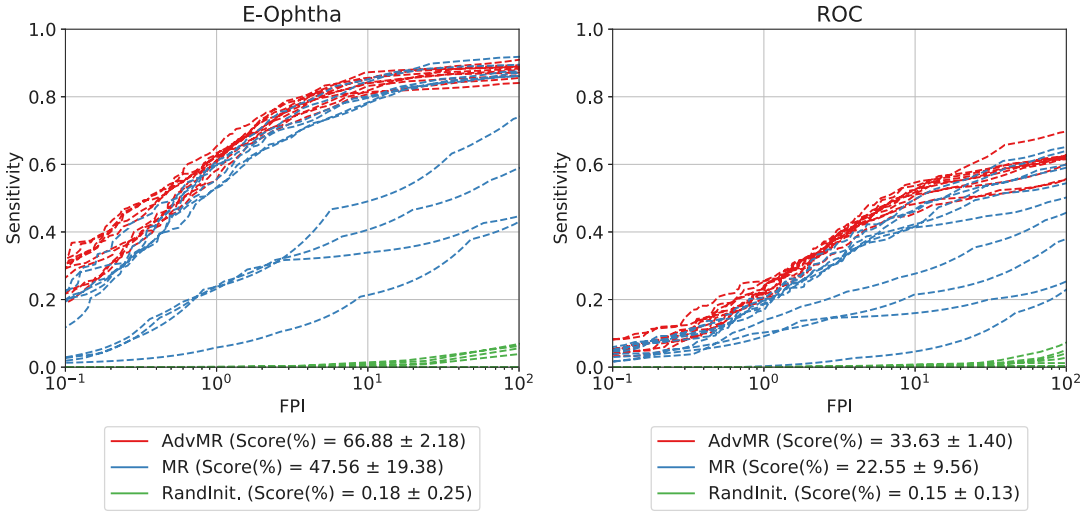
where P denotes the number of pixels within the (Positive) MA class, N the number of pixels within the (Negative) non-MA class, and (x, y) are the pixel coordinates. For those images without MAs, no weight map is used.

We compare the performance of the same three methodologies that were studied in Section 3.4 with the addition of the balanced loss (BL). Moreover, for reference, the results corresponding to the original methodology proposed in this work, AdvMR, are also included in the comparison. The results of the quantitative evaluation are depicted in Fig. 8. The experimental settings are the same that those in Section 3.4, including 10 experiments with the same training–test splits for E-Ophtha. However, in this case, all the methodologies result in a reduced variability, which enables the use of average curves and their standard deviation for the comparison.

Regarding the obtained results, it is observed that using the balanced loss significantly reduces the differences among methodologies. In this case, AdvMR(+BL) and MR(+BL) offer similar performance, and FromScratch(+BL) is much closer to them in the analysis. In this regard, the improvement of the networks trained from scratch is especially notorious, considering that they completely failed to perform



(a) Precision-Recall analysis



(b) FROC analysis

Fig. 6. Quantitative evaluation of the MAs detection using the heatmap regression. Comparison of different alternatives: the proposed adversarial multimodal pre-training (AdvMR), training from scratch (RandInit), and the multimodal pre-training previously proposed in [16] (MR).

the task in the previous experiments without the balanced loss (see Fig. 6). This demonstrates that the small number of MAs present in the images is one of the main challenges when training DNNs for MAs detection. However, in addition to this analysis, Fig. 8 also shows that none of the methodologies using the balanced loss is able to equal the performance of our original proposal, AdvMR. Thus, while class-balancing the task facilitates the training for those approaches with sub-optimal performance, it does not help when the base approach already produces a successful outcome. The reason for this could be that the class-balancing strategy provides a training feedback that is not as representative of the true target problem. To better understand these results, representative examples of predicted heatmaps are depicted in Fig. 9. These examples show that the networks trained with the balanced loss tend to detect more MAs in the images, producing substantially more incorrect detections. This visual analysis fits with the results depicted in Fig. 8, where for the same recall (or sensitivity) the addition of BL increases the false positives and reduces the precision. As a counterpart, the addition of BL allows to achieve higher recall values in the curves, but these are obtained with a very low precision, which is not useful for practical MAs detection. That kind of performance could be useful as a first detection step in a multi-stage pipeline. However, this would unnecessarily complicate the MAs detection. Precisely in this work, we aim at producing a more straightforward and efficient alternative to MAs detection using DNNs.

3.6. Comparison with the state-of-the-art

In this section, we provide a comparison of the proposed approach against relevant works in the literature. In particular, our comparison is mainly focused on recent works applying deep learning approaches to MAs detection. Despite that, other relevant works applying classical methods are also included as a reference.

The comparison for the E-Ophtha, ROC, and DDR public datasets is depicted in Table 1. To produce a fair comparison, we follow the same validation approaches as previous works using deep learning [12,14,15] . Particularly, in the case of E-Ophtha and ROC, the results of our approach correspond to the average of the experiments performed in Section 3.4. This means that, for E-Ophtha, we perform 5 repetitions of 2-fold cross-validation, whereas, for ROC, we perform a cross-dataset evaluation using the networks previously trained on E-Ophtha. For

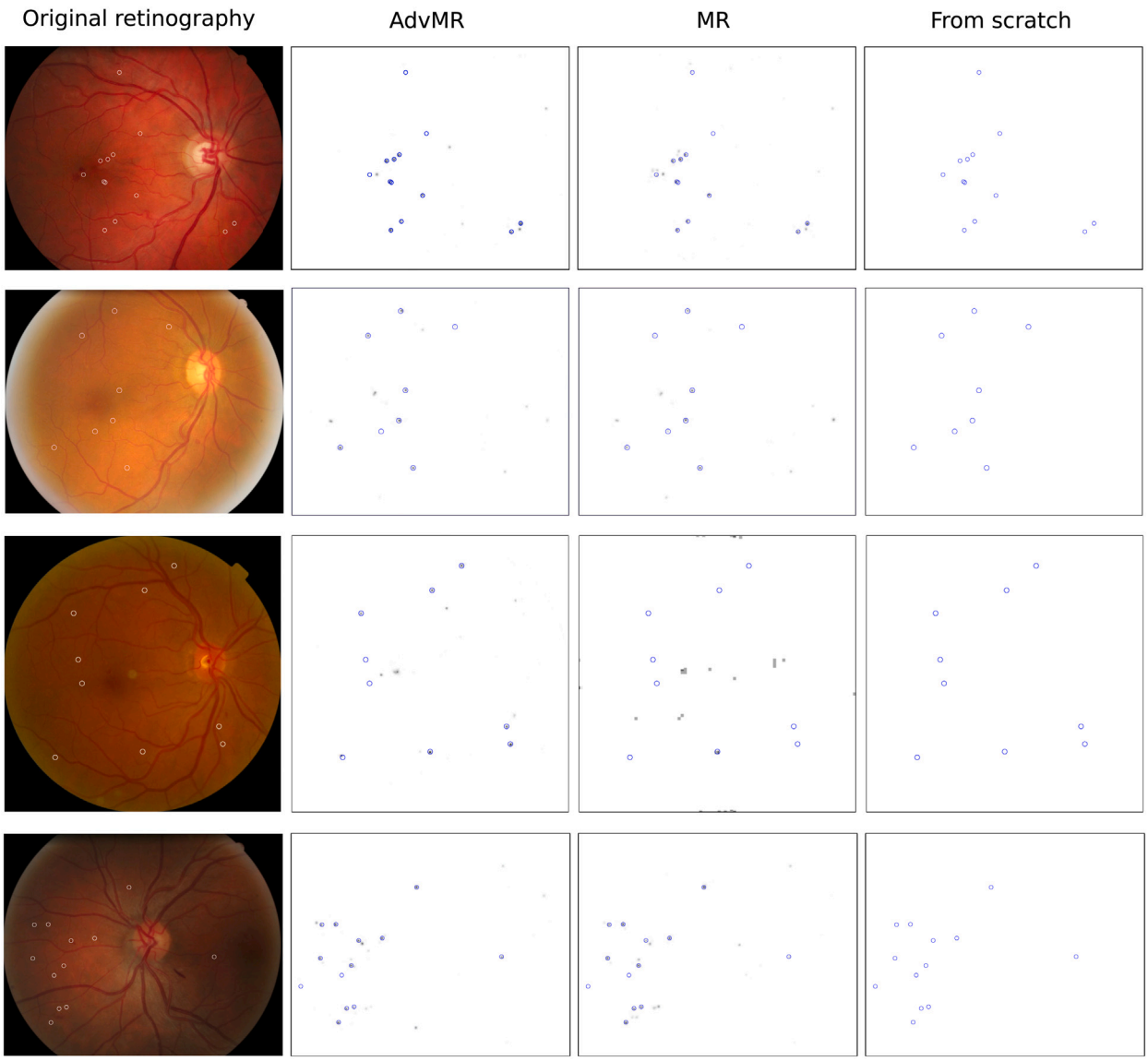


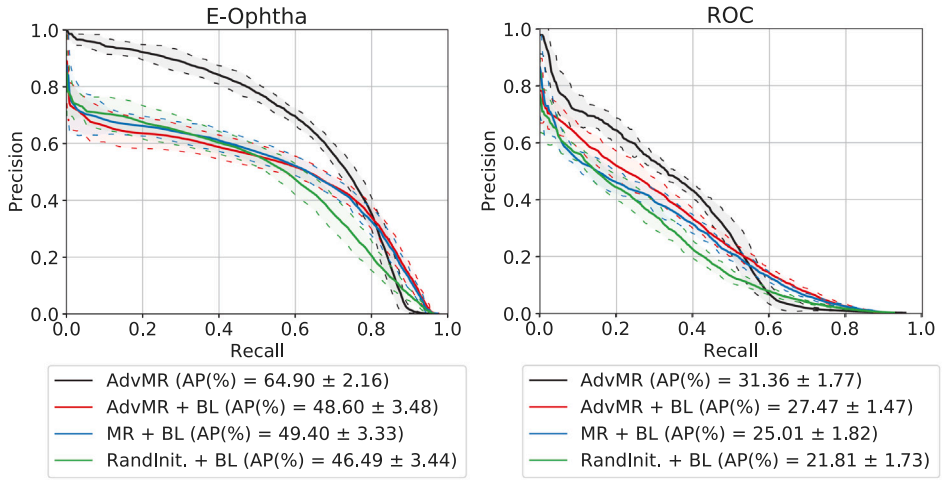
Fig. 7. Representative examples of predicted MAs heatmaps using the heatmap regression. Comparison of different alternatives: the proposed adversarial multimodal pre-training (AdvMR), the multimodal pre-training previously proposed in [16] (MR), and training from scratch. The ground truth MAs are marked with circles. Additionally, each depicted heatmap has been normalized independently to enhance the visualization.

DDR, we apply the default training–test split of the dataset. With regards to the FROC analysis, previous works in the literature are typically evaluated on a smaller subset of FPI values, discarding the lower or higher ends of the range that we use in our experiments. Thus, for the purpose of this comparison, we compute two different Score values, S_{LOW} for the range between 0.125 and 8 FPIs, and S_{HIGH} for the range between 1 and 20 FPIs.

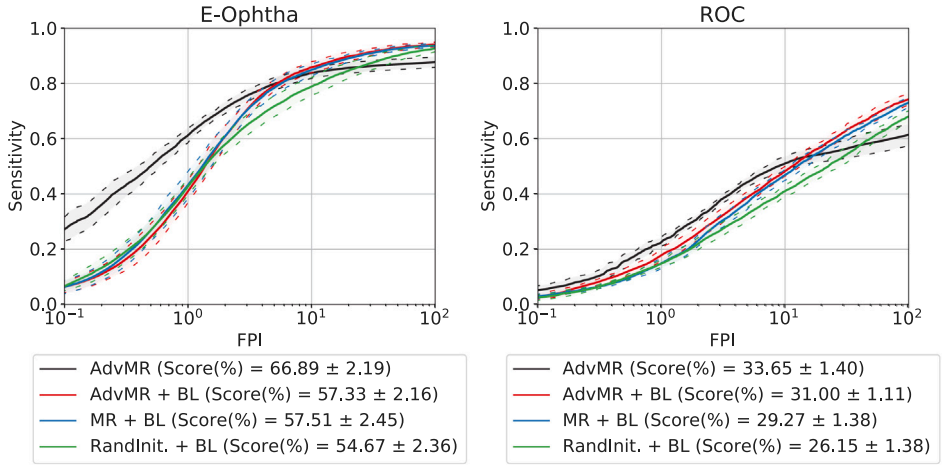
For the E-Ophtha dataset, it is observed that the proposed approach offers the best overall performance. Particularly, the method of Melo et al. [25] produces the best results at the lowest FPIs, whereas the method of Chudzik et al. [12] achieves slightly higher sensitivity values at mid-high FPIs. However, as previously stated, the herein proposed approach produces the best overall results, with a competitive performance across all the range of FPI values. Additionally, it can be observed that the performance of Melo et al. [25] is notoriously reduced when following a validation procedure similar to ours, with fewer available training samples. This evidences how the performance of different machine learning algorithms can be significantly affected by the amount of training data available, even when they rely on hand-engineered features instead of end-to-end learning using DNNs. Precisely in this work, we propose an adversarial multimodal pre-training that compensates for the lack of annotated training data. In this scenario, our proposal successfully facilitates the training of DNNs for the detection of MAs.

For the ROC dataset, as stated in Section 3.1, the provided comparison is performed on the public ROC *training* set. In spite of the cross-dataset evaluation, the proposed approach offers a competitive performance against methods that are specifically tuned for the ROC dataset. In this case, the best results at low and mid FPIs are achieved by the method of Dashtbozorg et al. [23]. However, the proposed approach achieves the best results among the deep learning alternatives following a similar evaluation approach. In particular, our proposal significantly outperforms the method of Chudzik et al. [12] across the full range of FPI values. Additionally, while previous deep learning-based approaches [12] needed to perform an ad-hoc fine-tuning on the ROC dataset, our approach achieves this good performance without any specific fine-tuning. This indicates a greater generalization ability of the proposed methodology.

The DDR public dataset is a recently available dataset that is provided by Li et al. [15]. In their work, several state-of-the-art deep



(a) Precision-Recall analysis



(b) FROC amalysis

Fig. 8. Quantitative evaluation of the MAs detection using the heatmap regression and class-balancing the loss function. Comparison of different alternatives: the proposed adversarial multimodal pre-training (AdvMR + BL), training from scratch (RandInit + BL), and the multimodal pre-training previously proposed in [16] (MR + BL). Additionally, the adversarial multimodal pre-training without the balanced loss is also included (AdvMR) as a reference.

learning approaches for object detection are tested. For the comparison, we select their best results, which are obtained using Faster RCNN [45]. Nevertheless, it is observed that our approach clearly outperforms this method. In fact, considering the PR analysis, Faster RCNN completely fails in the detection task. In this sense, the outcome is similar to the one obtained in our experiments training the networks from scratch (see Fig. 6). This evidences the difficulty of the problem at hand, given that well-proven object detection algorithms cannot be directly successfully applied to MAs detection.

3.7. Diabetic retinopathy detection

In this section, we study the feasibility of using the MAs predictions obtained with the proposed approach for the detection of DR. In order to obtain an estimate of the presence of DR in the eye fundus, we compute the maximum value in the predicted MAs heatmaps. This value can be seen as the likelihood of having at least one MA in the input image, which we use as a risk index for DR.

The evaluation for the detection of DR is performed on the E-Ophtha and Messidor datasets using networks that were previously trained for MAs detection on the E-Ophtha dataset. Thus, the experiments on Messidor correspond to a cross-dataset scenario. In particular, we use the same MAs detection networks that were used for the experiments in Sections 3.4 and 3.6 .

Table 2 depicts the results for DR detection, including a comparison with previous works on the detection of MAs that also provide this kind of study. The results show that the proposed approach provides a satisfactory performance for the detection of DR, using the estimated likelihood for the presence of MAs as a risk index. In this regard, our proposal outperforms previous alternatives on E-Ophtha and provides similar results to Dashtbozorg et al. [23] on Messidor, hence being the best overall alternative. Additionally, the proposed approach also produces satisfactory results on Messidor when the complete dataset is used (i.e. 0 vs Rest), which indicates that the method is robust to the presence of other retinal lesions besides MAs. Regarding the differences between datasets, the results show that the performance is always significantly better on E-Ophtha, which is the dataset used for training the networks. Thus, in part, the difference could be explained by the cross-dataset setting that is used on Messidor. This outcome is also in line with the results obtained for MAs detection.

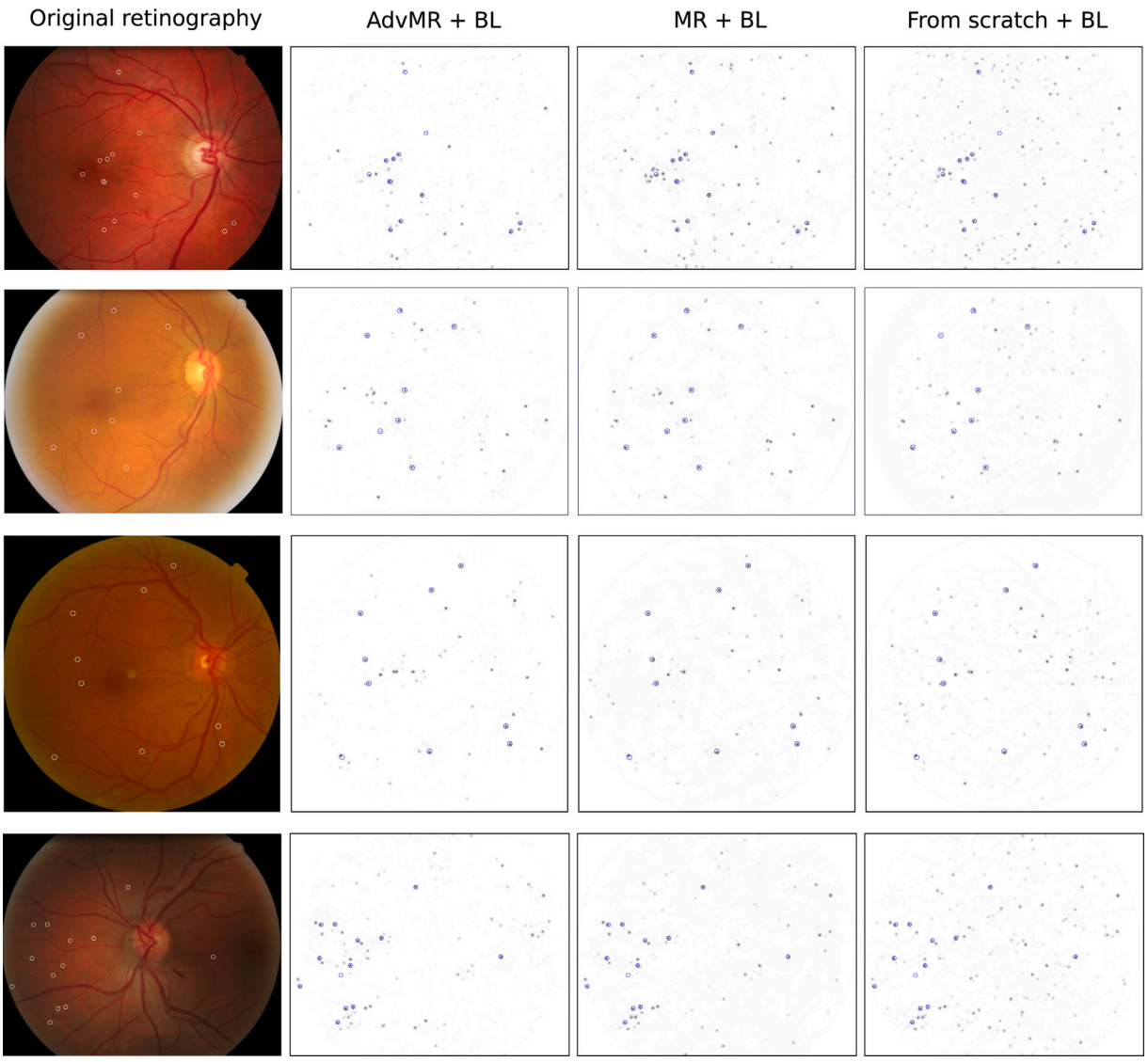


Fig. 9. Representative examples of predicted MAs heatmaps using the heatmap regression and class-balancing the loss function. Comparison of different alternatives: the proposed adversarial multimodal pre-training (AdvMR + BL), the multimodal pre-training previously proposed in [16] (MR + BL), and training from scratch (From scratch + BL). The ground truth MAs are marked with circles. Additionally, each depicted heatmap has been normalized independently to enhance the visualization.

3.8. Limitations and future works

Finally, the comparisons presented in Sections 3.6 and 3.7 demonstrate that the herein proposed methodology offers a satisfactory performance in a variety of different scenarios. The proposed methodology outperforms other deep learning-based methods. Additionally, it is also competitive against the best performing classical approaches, which rely on complex pipelines including specific image filtering techniques for feature extraction. In contrast, our proposal is based on end-to-end learning and it does not even require pre-processing or patch-processing of the images.

However, besides the satisfactory performance in the different scenarios, our experiments and comparisons also show that there is high inter-dataset variability in the detection of MAs. This seems to be due to the different capture devices and procedures as well as the different criteria of the experts when annotating the images [8]. The heterogeneity of the data and its effects on the performance of the algorithms is an important topic of study for future works. In this regard, it would be worth exploring the use of graph neural networks [46] for modeling the complex relations among datasets. Also, domain adaption techniques [47] could be explored for improving the generalization among different data sources. Another future research direction that we consider is the development of novel methods for the diagnosis of DR [48]. In Section 3.7, we already demonstrate that the proposed approach is useful for the detection of DR. However, it is worth exploring different alternatives to take advantage of the knowledge of these networks about the retina and the MAs. In this vein, multi-task learning [49] could be a useful tool for providing a more reliable and explainable diagnosis. Precisely, explainability in deep learning is a research topic that has attracted a lot of recent interest. In this regard, future works should also look for improving both the explainability and the causability (i.e. the quality of the explanation in terms of human understanding) of deep learning algorithms [50]. This will facilitate the application of the algorithms in the medical field [51].

4. Conclusions

The automated detection of MAs in retinography is a very challenging task due to the tiny size and reduced contrast of these vascular lesions. In this paper, we proposed a novel deep learning methodology for the detection of MAs in retinography. In particular, our proposal presents two main novelties. First, we approach the MAs detection

Table 1
Comparison with the state-of-the-art on the E-Ophtha, ROC, and DDR public datasets. In the case of ROC, the comparison is performed on the public ROC *training* set. S_{LOW}(%) is computed for the FPI values between 0.125 and 8 whereas S_{HIGH}(%) is computed between 1 and 20 FPI. DL denotes Deep Learning.

Method	Validation[a]	DL	Sensitivity at given FPI (%)										S_{LOW}(%)	S_{HIGH}(%)	AP(%)
			0.125	0.25	0.5	1	2	4	8	12	16	20			
E-Ophtha															
Veiga et al. (2017) [27]	–	No	11.0	15.2	22.2	30.7	38.8	49.4	62.9	–	–	–	32.8	–	–
Dashtbozorg et al. (2018) [23]	10-fold	No	35.8	41.7	47.1	52.2	55.8	60.5	63.8	–	–	–	51.0	–	–
Chudzik et al. (2018) [12]	2-fold	Yes	18.5	31.3	46.5	60.4	71.6	80.1	84.9	–	–	–	56.2	–	–
Du et al. (2020) [24]	10-fold	No	22.7	30.9	40.7	48.8	62.2	73.9	82.0	–	–	–	51.6	–	–
Melo et al. (2020) [25]	2-fold	No	17.8	28.4	38.3	51.9	58.7	58.7	58.7	–	–	–	44.6	–	–
Melo et al. (2020) [25]	Leave-one-out	No	43.4	43.4	43.4	54.0	57.2	61.8	63.9	–	–	–	52.4	–	–
Andersen et al. (2020) [13]	Fixed split	Yes	24.32	34.90	45.84	55.56	66.23	74.58	79.95	–	–	–	54.48	–	–
Savelli et al. (2020) [14]	2-fold	Yes	18.15	26.02	36.08	47.34	59.25	70.16	78.92	–	–	–	47.99	–	–
Proposed (AdvMR)	(5×)2-fold	Yes	29.97	40.09	50.18	61.28	71.16	78.12	82.67	84.45	85.18	85.69	59.07	78.36	64.90
ROC															
Zhou et al. (2017) [28]	10 Monte-Carlo	No	–	–	–	13.5	15.5	23.2	28.8	32.5	37.0	42.0	–	27.5	–
Javidi et al. (2017) [22]	Fixed split	No	–	–	–	12.98	14.70	20.89	28.70	31.90	35.34	38.33	–	26.12	–
Dashtbozorg et al. (2018) [23]	10-fold	No	43.5	44.3	45.4	47.6	48.1	49.5	50.6	–	–	–	47.1	–	–
Chudzik et al. (2018) [12]	Cross-dataset	Yes	2.8	4.0	6.3	9.0	10.8	12.8	13.9	15.6	16.3	17.7	8.5	13.7	–
Chudzik et al. (2018) [12]	+ Fine-tuning	Yes	3.9	6.7	14.1	17.4	24.3	30.6	38.5	43.1	46.1	48.5	19.3	35.5	–
Du et al. (2020) [24]	10-fold	No	15.5	16.6	19.7	26.0	33.1	43.6	50.4	–	–	–	29.3	–	–
Melo et al. (2020) [25]	10-fold	No	7.7	9.2	11.3	14.9	20.5	28.3	34.8	–	–	–	18.1	–	–
Proposed (AdvMR)	Cross-dataset	Yes	5.85	9.00	15.66	22.61	31.54	41.58	49.27	52.47	53.80	54.54	25.07	43.69	31.36
DDR															
Li et al. (2019) [15]	Fixed split	Yes	–	–	–	–	–	–	–	–	–	–	–	–	0.04
Proposed (AdvMR)	Fixed split	Yes	2.27	3.68	7.19	13.40	21.51	34.79	51.49	57.76	64.15	67.47	19.19	44.37	33.55

[a]"*k*-fold", "leave-one-out", and "*k* Monte-Carlo" indicate different cross-validation approaches. "Cross-dataset" indicates training on E-Ophtha and evaluation on ROC. "+ Fine-tuning" indicates additional training on ROC.

Table 2
Results for DR detection on the E-Ophtha and Messidor datasets.

Method	E-Ophtha	Messidor (0 vs 1)	Messidor (0 vs Rest)
	AUROC (%)	AUROC (%)	AUROC (%)
Melo et al. [25]	94.10	76.20	–
Dashtbozorg et al. [23]	95.78	78.50	–
Proposed (AdvMR)	97.46 ± 0.69	78.11 ± 0.97	86.05 ± 1.32

as a heatmap regression, which allows the simultaneous detection of multiple MAs and facilitates the extraction of precise locations. Second, we propose a novel adversarial multimodal reconstruction pre-training for learning about the retina and the MAs in a self-supervised fashion, i.e., from unlabeled data. The knowledge provided by this multimodal pre-training facilitates the detection of the MAs from the raw retinographies.

In order to analyze the proposed methodology, several experiments were conducted on different public datasets of reference. The obtained results show that our approach offers a satisfactory performance in a variety of different scenarios, outperforming existing deep learning alternatives. Additionally, given the heavily unbalanced scenario in the MAs detection, we also studied the effect of class-balancing the network training. In this regard, in comparison to existing alternatives, the proposed methodology not only avoids the use of pre-processing and the necessity of patch-processing the images but, also, it does not require any ad-hoc manipulation of the training data to mitigate the unbalance issue. In this sense, the proposed methodology is able to both learn and detect the MAs from the raw data as it is. Finally, our experiments also demonstrate that the proposed approach can be successfully used for the detection of diabetic retinopathy.

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, Spain 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 Cultura, Educación e Universidade, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%). Funding for open access charge: Universidade da Coruña/CISUG.

References

[1] E.Y. Chew, F.L. Ferris, Chapter 67 - nonproliferative diabetic retinopathy, in: Retina, fourth ed., Mosby, Edinburgh, 2006, pp. 1271–1284, http://dx.doi.org/10.1016/B978-0-323-02598-0.50073-2.

[2] D.S. Ting, L. Peng, A.V. Varadarajan, P.A. Keane, P.M. Burlina, M.F. Chiang, L. Schmetterer, L.R. Pasquale, N.M. Bressler, D.R. Webster, M. Abramoff, T.Y. Wong, Deep learning in ophthalmology: The technical and clinical considerations, Prog. Ret. Eye Res. 72 (2019) 100759, http://dx.doi.org/10.1016/j.preteyeres.2019.04.003.

[3] T.Y. Wong, R. Klein, D.J. Couper, L.S. Cooper, E. Shahar, L.D. Hubbard, M.R. Wofford, A.R. Sharrett, Retinal microvascular abnormalities and incident stroke: the atherosclerosis risk in communities study, Lancet 6358 (2001) 1134–1140, http://dx.doi.org/10.1016/S0140-6736(01)06253-5.

[4] E. Kohner, I. Stratton, S. Aldington, R. Turner, D. Matthews, Microaneurysms in the development of diabetic retinopathy (UKpds 42), Diabetologia 42 (1999) 1107–1112, http://dx.doi.org/10.1007/s001250051278.

[5] N. Cheung, P. Mitchell, T.Y. Wong, Diabetic retinopathy, Lancet 376 (9735) (2010) 124–136, http://dx.doi.org/10.1016/S0140-6736(09)62124-3.

[6] L. Cao, H. Li, Y. Zhang, L. Zhang, L. Xu, Hierarchical method for cataract grading based on retinal images using improved Haar wavelet, Inf. Fusion 53 (2020) 196–208, http://dx.doi.org/10.1016/j.inffus.2019.06.022.

[7] T. Li, W. Bo, C. Hu, H. Kang, H. Liu, K. Wang, H. Fu, Applications of deep learning in fundus images: A review, Med. Image Anal. 69 (2021) 101971, http://dx.doi.org/10.1016/j.media.2021.101971.

[8] M. Niemeijer, B. van Ginneken, M. Cree, A. Mizutani, G. Quellec, C. Sanchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, X. Wu, G. Cazuguel, J. You, A. Mayo, Q. Li, Y. Hatanaka, B. Cochener, C. Roux, F. Karray, M. Garcia, H. Fujita, M. Abramoff, Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs, IEEE Trans. Med. Imaging 29 (1) (2010) 185–195, http://dx.doi.org/10.1109/TMI.2009.2033909.

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

[10] L.A. Yannuzzi, K.T. Rohrer, L.J. Tindel, R.S. Sobel, M.A. Costanza, W. Shields, E. Zang, Fluorescein angiography complication survey, Ophthalmology 93 (5) (1986) 611–617, http://dx.doi.org/10.1016/S0161-6420(86)33697-2.

[11] R. Biyani, B. Patre, Algorithms for red lesion detection in diabetic retinopathy: A review, Biomed. Pharmacother. 107 (2018) 681–688, http://dx.doi.org/10.1016/j.biopha.2018.07.175.

[12] P. Chudzik, S. Majumdar, F. Calivá, B. Al-Diri, A. Hunter, Microaneurysm detection using fully convolutional neural networks, Comput. Methods Programs Biomed. 158 (2018) 185–192, http://dx.doi.org/10.1016/j.cmpb.2018.02.016.

[13] J.K. Andersen, J. Grauslund, T.R. Savarimuthu, Comparing objective functions for segmentation and detection of microaneurysms in retinal images, in: Proceedings of the Third Conference on Medical Imaging with Deep Learning, in: Proceedings of Machine Learning Research, vol. 121, 2020, pp. 19–32.

[14] B. Savelli, A. Bria, M. Molinara, C. Marrocco, F. Tortorella, A multi-context CNN ensemble for small lesion detection, Artif. Intell. Med. 103 (2020) 101749, http://dx.doi.org/10.1016/j.artmed.2019.101749.

[15] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, H. Kang, Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening, Inform. Sci. 501 (2019) 511–522, http://dx.doi.org/10.1016/j.ins.2019.06.011.

[16] 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. 91 (2020) 106210, http://dx.doi.org/10.1016/j.asoc.2020.106210.

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

[18] A.S. Hervella, J. Rouco, J. Novo, M. Ortega, Self-supervised multimodal reconstruction of retinal images over paired datasets, Expert Syst. Appl. (2020) 113674, http://dx.doi.org/10.1016/j.eswa.2020.113674.

[19] C. Baudoin, B. Lay, J. Klein, Automatic detection of microaneurysms in diabetic fluorescein angiography, Rev. D'Epidemiol. Sante Publ. 32 (3–4) (1984) 254—261.

[20] T. Spencer, J.A. Olson, K.C. McHardy, P.F. Sharp, J.V. Forrester, An image-processing strategy for the segmentation and quantification of microaneurysms in fluorescein angiograms of the ocular fundus, Comput. Biomed. Res. 29 (4) (1996) 284–302, http://dx.doi.org/10.1006/cbmr.1996.0021.

[21] T. Walter, J.-C. Klein, Automatic detection of microaneurysms in color fundus images of the human retina by means of the bounding box closing, in: Proceedings of the Third International Symposium on Medical Data Analysis, in: ISMDA '02, Springer-Verlag, Berlin, Heidelberg, 2002, pp. 210–220.

[22] M. Javidi, H.-R. Pourreza, A. Harati, Vessel segmentation and microaneurysm detection using discriminative dictionary learning and sparse representation, Comput. Methods Programs Biomed. 139 (2017) 93–108, http://dx.doi.org/10.1016/j.cmpb.2016.10.015.

[23] B. Dashtbozorg, J. Zhang, F. Huang, B. ter Haar Romeny, Retinal microaneurysms detection using local convergence index features, IEEE Trans. Image Process. 27 (7) (2018) 3300–3315, http://dx.doi.org/10.1109/TIP.2018.2815345.

[24] J. Du, B. Zou, C. Chen, Z. Xu, Q. Liu, Automatic microaneurysm detection in fundus image based on local cross-section transformation and multi-feature fusion, Comput. Methods Programs Biomed. 196 (2020) 105687, http://dx.doi.org/10.1016/j.cmpb.2020.105687.

[25] T. Melo, A.M.M. ca, A. Campilho, Microaneurysm detection in color eye fundus images for diabetic retinopathy screening, Comput. Biol. Med. 126 (2020) 103995, http://dx.doi.org/10.1016/j.compbiomed.2020.103995.

[26] T. Hellstedt, E. Vesti, I. Immonen, Identification of individual microaneurysms: a comparison between fluorescein angiograms and red-free and colour photographs, Graefes Arch. Clin. Exp. Ophthalmol. 234 (1996) 13–17, http://dx.doi.org/10.1007/BF02343042.

[27] D. Veiga, N. Martins, M. Ferreira, J. ao Monteiro, Automatic microaneurysm detection using laws texture masks and support vector machines, Comput. Methods Biomech. Biomed. Eng.: Imag. Visualiz. 6 (4) (2018) 405–416, http://dx.doi.org/10.1080/21681163.2017.1296379.

[28] W. Zhou, C. Wu, D. Chen, Y. Yi, W. Du, Automatic microaneurysm detection using the sparse principal component analysis-based unsupervised classification method, IEEE Access 5 (2017) 2563–2572, http://dx.doi.org/10.1109/ACCESS.2017.2671918.

[29] B. Zhang, X. Wu, J. You, Q. Li, F. Karray, Detection of microaneurysms using multi-scale correlation coefficients, Pattern Recognit. 43 (6) (2010) 2237–2248, http://dx.doi.org/10.1016/j.patcog.2009.12.017.

[30] C. Kou, W. Li, Z. Yu, L. Yuan, An enhanced residual U-net for microaneurysms and exudates segmentation in fundus images, IEEE Access 8 (2020) 185514–185525, http://dx.doi.org/10.1109/ACCESS.2020.3029117.

[31] A.S. Hervella, J. Rouco, J. Novo, M. Ortega, Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES), 2018, http://dx.doi.org/10.1016/j.procs.2018.07.213.

[32] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.

[33] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: The IEEE International Conference on Computer Vision (ICCV), 2017, http://dx.doi.org/10.1109/ICCV.2017.304.

[34] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.

[35] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, ICLR.

[36] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: 4th International Conference on Learning Representations, ICLR.

[37] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: International Conference on Computer Vision (ICCV), 2015, http://dx.doi.org/10.1109/ICCV.2015.123.

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

[39] A.S. Hervella, J. Rouco, J. Novo, M.G. Penedo, M. Ortega, Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images, Comput. Methods Programs Biomed. 186 (2020) 105201, http://dx.doi.org/10.1016/j.cmpb.2019.105201.

[40] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[41] F. Piccialli, V.D. Somma, F. Giampaolo, S. Cuomo, G. Fortino, A survey on deep learning in medicine: Why, how and when? Inf. Fusion 66 (2021) 111–137, http://dx.doi.org/10.1016/j.inffus.2020.09.006.

[42] S.H.M. Alipour, H. Rabbani, M.R. Akhlaghi, Diabetic retinopathy grading by digital curvelet transform, Comput. Math. Methods Med. 2012 (2012) http://dx.doi.org/10.1155/2012/761901.

[43] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, A. Chabouis, TeleOphta: Machine learning and image processing methods for teleophthalmology, IRBM 34 (2) (2013) 196–203, http://dx.doi.org/10.1016/j.irbm.2013.01.010, Special issue : ANR TECSAN : Technologies for Health and Autonomy.

[44] 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, http://dx.doi.org/10.5566/ias.1155.

[45] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, http://dx.doi.org/10.1109/TPAMI.2016.2577031.

[46] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (1) (2021) 4–24, http://dx.doi.org/10.1109/TNNLS.2020.2978386.

[47] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153, http://dx.doi.org/10.1016/j.neucom.2018.05.083.

[48] S. Stolte, R. Fang, A survey on medical image analysis in diabetic retinopathy, Med. Image Anal. 64 (2020) 101742, http://dx.doi.org/10.1016/j.media.2020.101742.

[49] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. (2021) http://dx.doi.org/10.1109/TKDE.2021.3070203, 1–1.
[50] D. Schneeberger, K. Stöger, A. Holzinger, The European legal framework for medical AI, in: A. Holzinger, P. Kieseberg, A.M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, 2020, pp. 209–226.
[51] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI, Inf. Fusion 71 (2021) 28–37, http://dx.doi.org/10.1016/j.inffus.2021.01.008.