16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

EyeWeS: Weakly Supervised Pre-Trained Convolutional Neural Networks for Diabetic Retinopathy Detection

Pedro Costa[1], **Teresa Araújo**[1,2], **Guilherme Aresta**[1,2], **Adrian Galdran**[1], **Ana Maria Mendonça**[2], **Asim Smailagic**[3], and **Aurélio Campilho**[2]

[1]INESC TEC
[2]Faculty of Engineering, University of Porto
[3]Department of Electrical and Computer Engineering, Carnegie Mellon University
[1]{pvcosta,tfaraujo,gmaresta,adrian.galdran}@inesctec.pt
[2]{amendon,campilho}@fe.up.pt
[3]asim@cs.cmu.edu

Abstract

Diabetic Retinopathy (DR) is one of the leading causes of preventable blindness in the developed world. With the increasing number of diabetic patients there is a growing need of an automated system for DR detection. We propose EyeWeS, a method that not only detects DR in eye fundus images but also pinpoints the regions of the image that contain lesions, while being trained with image labels only. We show that it is possible to convert any pre-trained convolutional neural network into a weakly-supervised model while increasing their performance and efficiency. EyeWeS improved the results of Inception V3 from 94.9% *Area Under the Receiver Operating Curve (AUC) to* 95.8% *AUC while maintaining only approximately* 5% *of the Inception V3's number of parameters. The same model is able to achieve* 97.1% *AUC in a cross-dataset experiment.*

1 Introduction

Diabetic Retinopathy (DR) is a worldwide leading cause of preventable blindness, affecting more than 25% [1] of the estimated 425 million diabetic patients in the world. The prevalence of diabetes is expected to grow to 629 million by 2045, and the number of patients requiring treatment will increase significantly in the following years [2]. In this context, early DR detection is important for successful treatment and thus large-scale screening programs are regularly implemented by hospitals and local authorities in both developed and developing countries. In these programs, diabetic patients are called to a clinic to obtain eye fundus images, which are then sent to ophthalmologists for diagnosis. Moreover, 34% of diabetic patients live in rural areas [2], where the access to medical specialists is difficult and screening programs are scarce. Also, the large number of images to analyze and other factors such as stress due to high clinical work-loads hinder the diagnosis procedure. For these reasons, systems that are capable of automatically detecting DR are becoming increasingly important for screening the growing number of diabetic patients and reaching a larger percentage of this population.

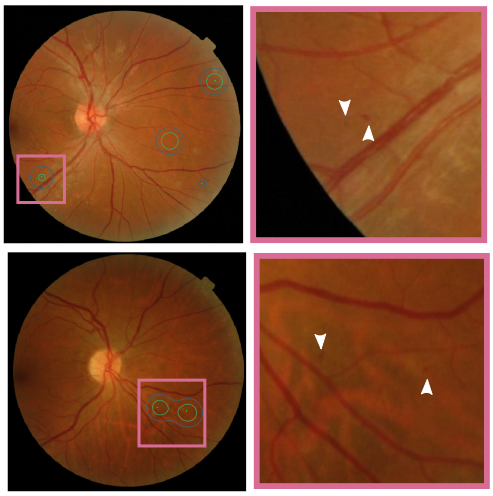


Figure 1: **EyeWeS is able to detect DR lesions being trained with image labels only.** Given a new image, our method not only outputs whether an image displays signs of DR or not but also highlights the regions that contain lesions.

Over the last years, Deep Learning has become the standard approach for the development of CAD systems. However, the development and application of these systems in real practice is hindered by the lack of annotated data, which is expensive to obtain, as well as the lack of explainability of the prediction. To deal with these obstacles, we introduce a novel general approach to train modern Convolutional Neural Network (CNN) architectures for the task of DR detection, illustrating how to easily convert such architectures into weakly-supervised models. This conversion removes the need of data annotated lesion-wise at the pixel-level for train-

ing, while maintaining the ability to pin-point regions of the image that contain lesions relevant to diagnosis, as illustrated in Figure 1. The main contributions of this work are hence in terms of: *Accuracy* - our method is shown to be accurate even in cross-dataset experiments; *Explainability* - the designed approach finds regions with signs of DR; *Efficiency* - the approach introduced here represents a straightforward technique to convert any current CNN architecture designed for the task of classification into a weakly-supervised model with a much reduced set of parameters; and *Speed* - our method is fast to both classify and explain the results.

2 Deep Learning for Diabetic Retinopathy

In recent years, Deep Neural Networks (DNNs) have started to have a strong presence in the field of medical data analysis. Gulshan *et al.* [3] fine tuned an ensemble of 10 inception v3 networks on a large private dataset and reported results comparable to a panel of seven certified ophthalmologists. However, end-to-end Deep Learning approaches, such as the one proposed by Gulshan *et al.* [3], suffer from the common criticism of being unable to explain their decision.

Concurrently, Abràmoff *et al.* [4] trained several Deep Learning models to detect DR lesions and anatomical landmarks. Then, the features extracted by these CNN models are provided to a fusion algorithm that outputs the final grade of the exam. This way, the decision of the method is slightly easier to explain since it is able to identify which lesions are present in a given image.

The next step would be to locate the lesions in the image to further improve the model's explainability. The problem with this approach is that most of the existing methods require additional time consuming annotations, such as pixel-level annotations, in order to output such information. Weakly-Supervised methods are a promising solution for this problem. As opposed to Fully-Supervised methods, a Weakly-Supervised model is trained to learn low-level information from data, even if the corresponding low-level ground-truth is not available but rather a higher-level source of information is present.

A form of Weak Supervision that has been particularly successful in biomedical applications is the Multiple Instance Learning (MIL) framework [5, 6, 7]. In MIL, instead of supplying a learning algorithm with pairs of inputs and labels, these labels are associated with sets of inputs, also known as *bags*. In the case of binary labels, the fundamental MIL assumption states that a positive bag (*e.g.* an image with signs of disease) contains at least one positive instance (*e.g.* a lesion).

Training MIL-based models requires only weak annotations. This is translated in practice to using only image-level labels, while still classifying images based on instance-level information, *i.e.* on the presence of lesions. Requiring weak annotations simplifies enormously data collection, which is a major issue in medical image analysis. However, these weaker annotations need to be compensated by larger datasets [5]. Accordingly, the availability of such large datasets is a typical prerequisite for training MIL-based algorithms. This need is stressed if the models to be trained are deep CNNs, which contain a great amount of parameters to be learned. For this reason, in this paper we propose to embed standard Transfer Learning strategies into a MIL framework. The technical details on how to achieve this goal are explained in the next section.

3 EyeWeS for explainable Diabetic Retinopathy detection

The approach proposed in this paper, referred to as EyeWeS, consists of a combination of MIL and Transfer Learning for deep CNNs. As such, the method is capable of formulating a decision regarding the presence of DR on a retinal image, while detecting the image's regions that better explain such decision. We start by formalizing the problem through the MIL framework and then, following these insights, we outline a strategy to modify current state-of-the-art CNN architectures to better suit the *Standard MIL Assumption*.

3.1 Intuition

In order to train a MIL-based model on retinal images to detect DR based on lesion presence, our goal is to train an *instance* classifier from *bag* labels. In this context, an image is regarded as a *bag* composed of several rectangular spatial neighborhoods, *i.e.* image patches. These patches are considered as instances. Hence, the goal becomes to train a patch classifier from image labels only.

Since patch-level labels are not available, we consider these instance's labels, y_i, as latent variables, because they are unknown during training. The latent labels can be combined by means of a pooling function f into the corresponding bag label $Y = f(y_1, ..., y_N)$. The pooling function is responsible for encoding the relation between the instances' and bag's labels.

The objective is to learn an instance classifier $P(y_i|x_i, \theta)$, where x_i denotes the i-th instance in an image, and θ are the classifier's parameters. As the instances' labels are unknown, we can only maximize the likelihood $P(Y|y_1, ..., y_N, x_1, ..., x_N, \theta)$. Furthermore, we assume that $\{X, \theta\}$ and Y are conditional independent given y. This means that, given the labels of the instances, both the instances and the model's parameters do not provide information on the likelihood of Y. Therefore, the goal is to find the parameters θ that maximize the likelihood:

$$\theta = \arg\max_{\theta} P(Y|y_1, ..., y_N; \theta) \quad (1)$$

There are some design choices that need to be made before implementing this idea, namely: 1) what is the

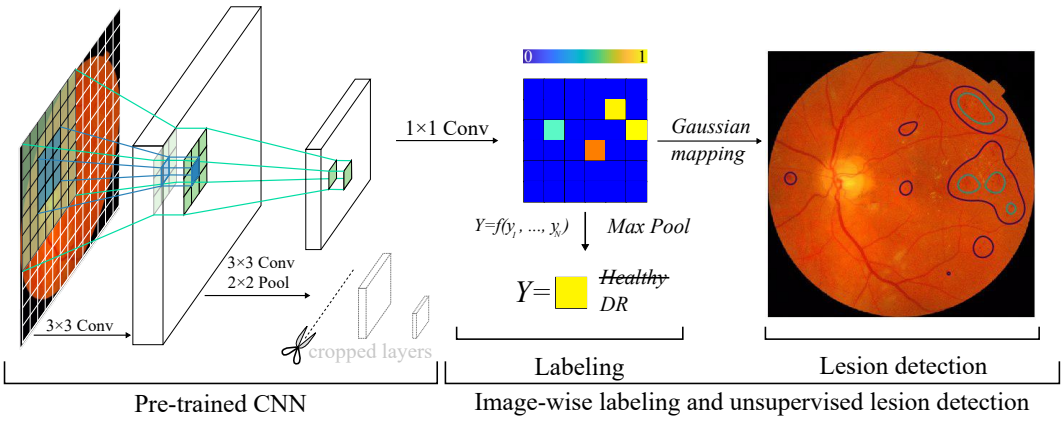


Figure 2: Early layers of pre-trained networks are kept to constrain the receptive field of their output units. For instance, the receptive field of the first 3×3 Convolution is a 3×3 region of the input image, but the receptive field of two 3×3 Convolutions is a 5×5 region of the input. Then, 1×1 Convolution layers are used to classify the input patches. Finally, the image prediction is obtained by max pooling all the patch predictions.

learning algorithm to use for the patch classifier and 2) what pooling function f to use.

3.2 Fully-Convolutional Patch Classifier

We chose to use a CNN for the patch classifier algorithm. However, instead of training a CNN on individual image patches and their corresponding labels, we use a Fully-Convolutional Network (FCN) architecture, which is capable of classifying image patches trained only with the full images and the respective labels.

For that, we build on receptive field properties of Convolutional Layers. The receptive field can be defined as the local region in the input space that conditions the value of a particular activation unit in a given layer. Therefore, we can design a Fully-Convolutional Network architecture such that the receptive field of the last layer's units corresponds to the desired patch size, as depicted in Figure 2.

3.3 Combining Patch Predictions

The computed image patch predictions are then combined into the image label using the max-pooling function $Y = max(y_1, ..., y_N)$, since a single positive instance is enough to predict the full image as positive. As the *max* function is almost everywhere differentiable, the misclassification loss function can be directly applied to Y and the network trained with the backpropagation algorithm.

It is worth noting that, in practice, most binary disease detection problems follow the *Standard MIL Assumption* since the presence of a disease can usually be inferred by the existence of a lesion in the image. However, since the prediction is based on local information, EyeWeS can not be used for classifying the severity of a disease, as this usually requires the detection, classification and counting of multiple lesions that can be spread over the entire image.

3.4 Transfer Learning in EyeWeS

When few training data is available, training a deep CNN from scratch can be unfeasible. In these situations, it is common practice to initialize the network from a set of pre-trained weights and fine tune it for the given task.

We propose to find the layer of the pre-trained network whose receptive field is closer to a desired patch size and discard all subsequent layers. This patch size should be large enough to contain the lesions to be detected within the image. The output of this intermediate layer is then a $H \times W \times K$ tensor with a K-dimensional feature vector for each of the $N = H \times W$ image patches. Then, 1×1 Convolutional Layers are added in order to perform the patch classification without increasing the receptive field, as shown in Figure 2.

4 Application and Results

We train and test EyeWeS on the Messidor [8] dataset for solving the task of DR detection. Messidor images were annotated by specialists with the corresponding DR grade, *i.e.*, DR level of severity, ranging from 0 (no pathology) to 3 (most severe stage). In this work we are only interested in detecting DR and, therefore, we pose the problem as binary one-*vs*-all classification task, distinguishing between healthy images (*i.e.* grade 0) and DR images (*i.e.* grades 1, 2 and 3). We randomly divided Messidor into three sets: a training set with 768 images (64%), a validation set with 192 images (16%), and a test set with the remaining 240 images (20%).

We also perform a cross-dataset experiment to both evaluate the generalization capabilities of EyeWeS and

Table 1: **EyeWeS' receptive field size.** The optimal receptive field (RF) size in pixels (px) for each network is much smaller than the original image resolution.

	VGG	ResNet50	InceptionV3
Crop L.	block4_conv1	add 3	mixed 2
RF	52×52px	30×30px	114×114px
Overlap	44×44px	26×26px	106×106px
N	64×64	127×127	61×61

perform a qualitative assessment of its lesion detection capability. For this, we used the E-ophtha MA dataset [9] that contains 148 images with microaneurysms or small hemorrhages (together with segmentations) and 233 images with no lesions.

4.1 Implementation Details

The training process for EyeWeS proceeds in 3 steps: 1) select the layer of the given pre-trained model to use; 2) train the newly added layers, while keeping the parameters of the pre-trained layers constant and 3) train the full model. All our experiments were implemented using Keras [11] and all networks were pre-trained on ImageNet.

In this work, we experimented with three different CNN architectures: VGG16 [12], Inception V3 [13] and Resnet50 [14]. The number of layers of the given pre-trained model to reuse was treated as a hyper-parameter and it was chosen independently for each architecture using grid-search over all intermediate layer blocks. Following Keras' naming conventions, the layers that achieved optimal performance for each architecture were: block4_conv1 (VGG16), mixed 2 (Inception V3) and add 3 (Resnet50). Details on the receptive field size and overlap of each architecture are displayed in Table 1 along with the number of patches that are classified prior to the max pooling operation.

Two 1×1 Convolution Layers are used after the output of the selected intermediate layer, the first one with 1024 units, followed by a LeakyReLU activation function with slope of 0.02 and the second one with a single unit, followed by a sigmoid activation function in order to perform patch-level classification. These last two Convolution Layers are trained for the first 30 epochs with a learning rate of 10^{-3}. Then, the full model is trained with a learning rate of 2×10^{-4} using early stopping with a patience of 15. Adam [15] optimizer was used in both steps with default parameters.

The input images were cropped to the Field-of-View and resized to 512×512 pixels. In order to improve the generalization of the model, dataset augmentation was used. On top of the standard horizontal and vertical flip, translation, rotation and scaling of the input images, a color balance method [16] was also used as a dataset augmentation operator as it was shown to improve results in segmenting vessels in eye fundus images.

Table 2: **EyeWeS' DR detection results on Messidor**. EyeWeS achieves state-of-the-art Area Under the Receiver Operating Curve (AUC) compared with other weakly-supervised methods.

Architecture	AUC	Weakly-Sup.
Costa *et al.* [7]	90.00%	✓
Zoom-In-Net [10]	92.10%	✓
VGG16	83.44%	
ResNet50	93.77%	
Inception V3	94.97%	
EyeWeS VGG16	90.00%	✓
EyeWeS ResNet50	94.53%	✓
EyeWeS Inception V3	**95.85%**	✓

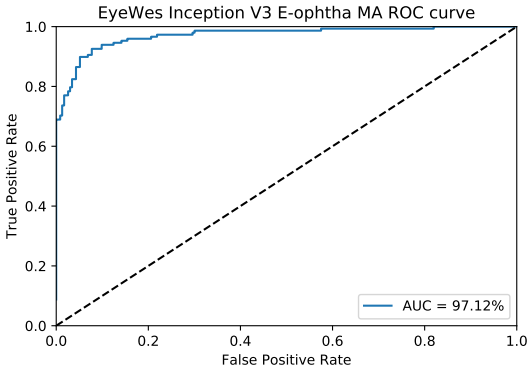
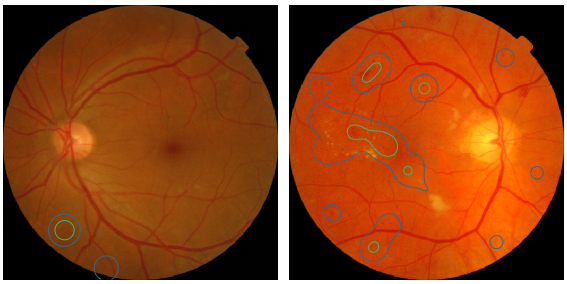


Figure 3: **EyeWeS generalizes to other datasets**. ROC of EyeWeS on the E-ophtha MA dataset.

4.2 Quantitative Results

In order to properly evaluate our method, we performed two experiments: 1. we tested on Messidor and 2. we performed a cross-dataset experiment, testing EyeWeS on e-ophta MA. Results from experiment 1) are shown in Table 2. The Inception V3 network obtains the best results (95.85% AUC) closely followed by ResNet50 (94.53%). On the other hand, VGG16 is not able to achieve comparable results, attaining only 90% AUC. Nonetheless, is is worth noting that this result is already at the same level as other recent works [7, 10].

We also compared EyeWeS with the standard pre-trained architectures. For that, we removed the last fully-connected layers of the given network and applied a Global Average Pooling layer to accommodate for the increase in image resolution. Then, similarly to the EyeWeS, we added two 1×1 Convolution layers, the first one with 1024 units followed by a LeakyReLU activation function and the last one with a single unit followed by a sigmoid. All networks were trained in the same conditions as EyeWeS. As seen in Table 2, EyeWeS always obtains better AUC results than their



Grade 1 ($Y = 1$) Grade 3 ($Y = 1$)

Figure 4: **Examples of EyeWeS's explainable results on Messidor's test images.** Grades are displayed for visualization purposes only, we do not use grade information nor lesion segmentations when training our method. Attention map: 0 1

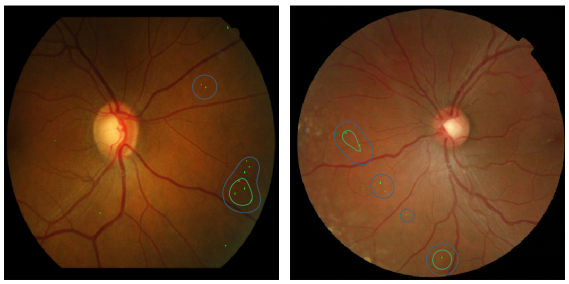


Figure 5: Microaneurysm true locations from E-ophtha MA are displayed in light green. It is possible to see that the method's detection mostly coincides with the dataset's true locations. Attention map: 0 1

standard counterparts.

Finally, in order to fairly evaluate if EyeWeS generalizes to other datasets, we selected the model that performed best in Messidor, in this case, the EyeWeS Inception V3 model, and tested it on the full E-ophtha MA dataset. We obtained 97.12% AUC in this cross-dataset experiment as shown in Figure 3, indicating that our method is effectively finding microaneurysms or small hemorrhages.

4.3 Explainability

EyeWeS produces patch-level predictions even though it is only trained with image labels, which can help in visually interpreting the reasons of the models' decisions, enhancing its explainability. As expected, the attention maps produced by EyeWeS focus on eye lesions. In Figure 4 it is possible to visualize some sample results on images from Messidor's test set with different severity labels. As can be observed, regions of the image that the method considers pathological lie on top of red lesions.

Another interesting result of our method is that it focuses on small lesions while ignoring the larger more visible ones. It is possible to see in the grade 3 image of Figure 4 that EyeWeS does not focus on large hemorrhages and bright yellow lesions. These results potentially mean that the method was able to correctly identify microaneurysms as the earliest and most subtle indication of DR.

In order to test this last hypothesis, we visualize the attention maps produced by our method on E-ophtha MA in Figure 5. It is possible to see that the model's detections mostly coincide with microaneurysm locations.

4.4 Efficiency and Inference Time

In this section we explore the impact of our method in the reduction of the number of parameters and inference time. For the performance measurements, experiments were run on a laptop equipped with a mobile Nvidia GTX 1060 GPU. Inference times were averaged over 100 experiments. The resulting performances, within a 95% confidence interval (CI), are reported in Table 3. We also compare EyeWeS with each of the three corresponding CNN baseline architectures. In this case, EyeWeS ResNet50 was 2.43× faster than its original counterpart at test time, while EyeWeS Inception V3 was 1.68× faster, and EyeWeS VGG16 was 1.32× faster.

Implementing EyeWeS allows to reduce the number of parameters in more than a quarter with respect to the corresponding reference architectures, as can also be observed from Table 3. In the extreme case of Resnet50, the number of parameters was reduced to less than 2% of the original number. EyeWeS also reduces the number of parameters of Inception V3 to approximately 5% of its original number, while VGG16 parameter numbers are reduced to approximately 23%.

5 Conclusions

We have introduced EyeWeS, which addresses the problem of explainable detection of Diabetic Retinopathy from eye fundus images. Out of global image labels, our method trains a patch classifier that can be used at test time to detect the regions of the image that contain lesions of interest for DR diagnosis. The proposed method has been shown to be capable of accurately detecting DR on retinal images, basing its decisions on local information. EyeWeS's architecture allows to pinpoint the spatial locations triggering the model's decisions, which enhances its explainability.

EyeWeS has been comprehensively validated through experimental tests on the Messidor dataset for DR detection, achieving state-of-the-art results. In addition, the same model has been tested on E-ophtha MA without further re-training. Comparison with the pixel-wise lesion ground-truth available for E-opththa MA showed that EyeWeS's patch classifier detects microaneurysms as small as 3 pixels in diameter. It has also been shown that it is possible to decrease the number of parameters

Table 3: **EyeWeS is faster and has fewer parameters.** Mean inference time in milliseconds with 95% CI for a single image and number of parameters of each network.

	EyeWeS Time (ms)	EyeWeS Parameters	Full Time (ms)	Full Parameters
Inception V3	38.92 ± 2.67	$1,240,673$	65.50 ± 4.04	$23,715,265$
Resnet50	33.74 ± 1.30	$498,049$	81.93 ± 3.74	$25,691,009$
VGG16	57.08 ± 1.25	$3,446,081$	75.39 ± 1.78	$15,245,121$

by more than 98% with respect to the reference CNN architecture from which EyeWeS is derived and still obtain more accurate results with the added ability to explain the results.

Acknowledgments

This work is financed by the ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, and by National Funds through the FCT - Fundao para a Cincia e a Tecnologia within project CMUPERI/TIC/0028/2014 and by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement within the project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145- FEDER-000016". Teresa Arajo is funded by the FCT grant contract SFRH/BD/122365/2016. Guilherme Aresta is funded by the FCT grant contract SFRH/BD/120435/2016.

References

[1] LM Ruta, DJ Magliano, R Lemesurier, HR Taylor, PZ Zimmet, and JE Shaw. Prevalence of diabetic retinopathy in type 2 diabetes in developing and developed countries. *Diabetic Medicine*, 30(4):387–398, 2013.

[2] K Ogurtsova, JD da Rocha Fernandes, Y Huang, U Linnenkamp, L Guariguata, NH Cho, D Cavan, JE Shaw, and LE Makaroff. Idf diabetes atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, 128:40–50, 2017.

[3] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316 (22):2402–2410, 2016.

[4] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology & Visual Science*, 57(13):5200–5206, 2016.

[5] Gwenolé Quellec, Katia Charrière, Yassine Boudi, Béatrice Cochener, and Mathieu Lamard. Deep image mining for diabetic retinopathy screening. *Medical Image Analysis*, 39:178–193, 2017.

[6] Pedro Costa and Aurélio Campilho. Convolutional bag of words for diabetic retinopathy detection from eye fundus images. *IPSJ Transactions on Computer Vision and Applications*, 9(1):10, 2017.

[7] P Costa, A Galdran, A Smailagic, and A Campilho. A weakly-supervised framework for interpretable diabetic retinopathy detection on retinal images. *IEEE Access*, 2018.

[8] E Decenciere, X Zhang, G Cazuguel, B Lay, B Cochener, C Trone, P Gain, JR Ordonez-Varela, P Massin, A Erginay, B Charton, and JC Klein. Feedback on a publicly distributed image database: The Messidor database. *Image Analysis and Stereology*, 33(3):231–234, 2014.

[9] 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, and A. Chabouis. TeleOphta: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2):196–203, 2013.

[10] Z Wang, Y Yin, J Shi, W Fang, H Li, and X Wang. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 267–275. Springer, 2017.

[11] F Chollet et al. Keras. `https://keras.io`, 2015.

[12] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[13] C Szegedy, V Vanhoucke, S Ioffe, J Shlens, and Z Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE, 6 2016.

[14] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] DP Kingma and J Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] B Savelli, A Bria, A Galdran, C Marrocco, M Molinara, A Campilho, and F Tortorella. Illumination correction by dehazing for retinal vessel segmentation. In *Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on*, pages 219–224. IEEE, 2017.