EyeQual: Accurate, Explainable, Retinal Image Quality Assessment

Pedro Costa[†*], Aurelio Campilho[†**], Bryan Hooi[‡], Asim Smailagic[*], Kris Kitani[¶], Shenghua Liu[§], Christos Faloutsos[‖] and Adrian Galdran[†]
[†]INESC TEC, Porto
[*]Department of Electrical and Computer Engineering, Carnegie Mellon University
[**]Faculty of Engineering, University of Porto
[‡]School of Computer Science, Department of Statistics and Data Science, Carnegie Mellon University
[¶]Robotics Institute, Carnegie Mellon University
[§]CAS Key Lab of Network Data Science and Technology Institute of Computing Technology, Chinese Academy of Sciences, China
[‖]School of Computer Science, Carnegie Mellon University

***Abstract*—Given a retinal image, can we automatically determine whether it is of high quality (suitable for medical diagnosis)? Can we also explain our decision, pinpointing the region or regions that led to our decision? Images from human retinas are vital for the diagnosis of multiple health issues, like hypertension, diabetes, and Alzheimer's; low quality images may force the patient to come back again for a second scanning, wasting time and possibly delaying treatment. However, existing retinal image quality assessment methods are either black boxes without explanations of the results or depend heavily on feature engineering or on complex and error-prone anatomical structures' segmentation.**

Therefore, we propose EyeQual, that solves exactly this problem. EyeQual is novel, fast for inference, accurate and *explainable*, pinpointing low-quality regions on the image. We evaluated EyeQual on two real datasets where it achieved 100% accuracy taking just 36 milliseconds for each image.

I. Introduction

Can we automatically detect a low quality retinal image? Can we pinpoint the regions on the image, that led to our conclusion? The human retina is an excellent source of biomarkers that can help identifying early signs of several disorders, such as heart diseases, hypertension, Alzheimer's or Diabetic Retinopathy (DR) [1], [2]. About 98% of people with type I diabetes have at least background retinopathy after 25 years of the disease [3]. At the time of diagnosis of type II diabetes about a third of the people already have diabetic retinopathy to certain extent. For this reason, diabetic patients need to be routinely examined by experienced ophthalmologists.

In short, technicians acquire retinal images, and send them to doctors who perform the diagnosis, usually in the context of screening programs. The problem is that often images have low quality (too dark, or too bright, etc), and doctors cannot do reliable diagnosis. The research problems we focus in this paper are: (a) detect retinal images that are of low quality, as early as possible (so that the technician can acquire another image) and (b) *explain* why the image is low-quality, so that the technician can solve the appropriate problem (e.g., increase/decrease the illumination, improve the focus, etc).

The implementation of large-scale screening programs has led to a great increase in the amount of retinal images that need to be reviewed by specialists. Unfortunately, in these programs images are acquired within different sites, by different cameras that are operated by technicians with a varying level of experience. As a consequence, the proportion of retinal images acquired in clinical settings depicting insufficient quality is variable and substantial. Niemeijer *et al.* [4] reported that 12% of the images obtained in a web-based screening program involving $1,676$ patients was considered as unreadable by the ophthalmologists, while Fleming *et al.* [5] indicated that between 5.6% and 20.5% of $33,535$ patients undergoing screening had an ungradable image in at least one eye. This is a relevant issue, since the missing step of immediate low-quality control can force patients to come back to the medical center to re-acquire extra images, or give up with the risk of delayed diagnosis. Low-quality images also complicate diagnosis for doctors. Therefore, there is a great interest among the retinal image analysis community in designing reliable automatic image quality assessment algorithms.

In short, we want to solve the problem of *Explainable Image Quality Assessment:*

- **Given** several retinal images, with label (yes/no for quality),
- **Classify** new, unlabeled images, and **explain** the decision.

In this paper we propose to solve the the retinal image quality assessment problem by learning a patch classifier given a set of eye fundus images and corresponding quality labels. Therefore, our method not only classifies the input image, but also returns a heatmap pinpointing the location of the high/low quality patches, as shown in Figure 1. The first two images are high quality, and EyeQual does label them as such; moreover, it finds very few low-quality patches (in yellow). The rightmost two images are of low quality, and EyeQual also correctly declares them as such; for the first one, it pinpoints two low-quality regions (the bottom-left, because it is too bright; the top-right, because it is too dark). For the second low-quality image, EyeQual classifies all the patches as low-quality, which, as we see, are all very bright and prevent the differentiation

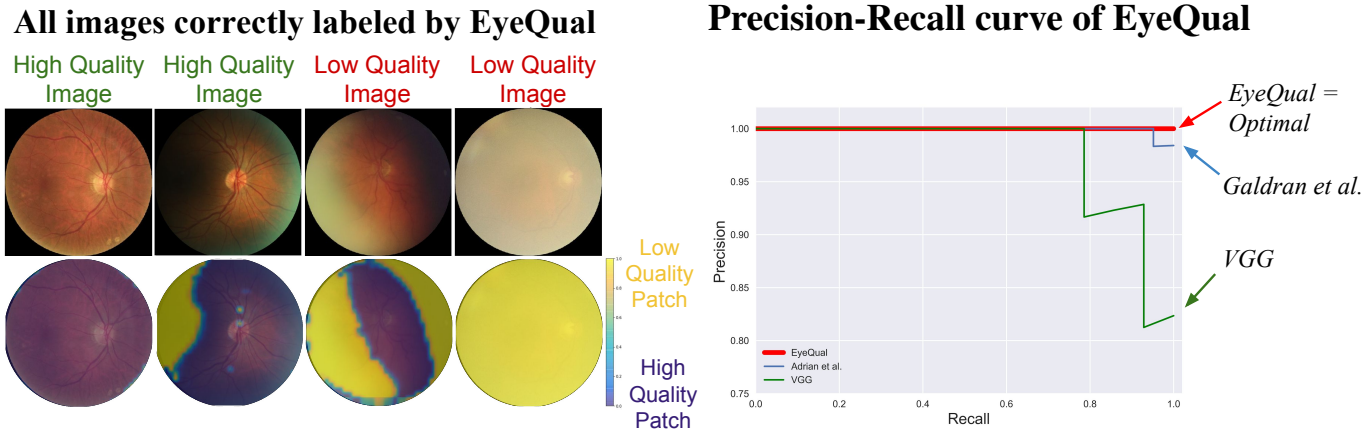

Fig. 1: **EyeQual can explain its decision.** EyeQual gives the appropriate label, as well as marks the culprit low quality regions (in yellow). We evaluated EyeQual in two datasets, obtaining 100% accuracy in both.

of any anatomical structure. From all images, we manually clipped the four cornerns, for visual clarity (our EyeQual learns to ignore them, as we describe later in section III).

In summary, our contributions are as follows:

- *Novelty* New method for image quality assessment:
  - Learns to classify patches using only image labels;
  - The proposed Shifted Weighted Average Pooling (S-WAP) enables the automatic specification of the importance of every region of the retina;
  - Careful pooling of patch scores into the image score.
- *Explainability*: Interpretable results (visualization and attention routing). Capable of pinpointing low quality regions in the image.
- *Accuracy*: Our proposed approach achieves top performance (100% accuracy) on a recent medium-sized and a smaller publicly available datasets.
- *Speed*: Inference is fast (36 ms).

Reproducibility: our work is reproducible as we evaluate EyeQual on a publicly available dataset and we open-source our code[1].

II. Background and Related Work

A. Image quality assessment

Existing image quality assessment methods can be broadly divided into four different categories: *Structural* approaches, *Generic* techniques, *Hybrid* methods, and more recently, methods based on *Deep Learning*. *Structural* approaches are techniques based on building and analyzing image representations specifically designed for retinal images. Thus, they take into account, for instance, if the amount of visible vessels is within a reasonable proportion [6], if the main anatomical structures are present in the image [4], or if the image intensity histogram follows a distribution similar to that of good-quality retinal images [7]. The major drawback of this approach is that most of these methods rely on the segmentation of anatomical landmarks, which is a complex and error-prone process, specially in the case of low-quality images. On the contrary, *Generic* approaches employ or adapt visual features coming from general image quality assessment methods designed to deal with natural images, such as geometric or texture features [8], attempting to measure defocus, lack of sharpness, wrong illumination, etc. [9]. A particularly interesting subset of this kind of techniques are methods supported by features inspired on the Human Visual System, such as [10] or [11]. *Generic* techniques can achieve good results in discriminating low and high-quality retinal images, but they often lack specificity. *Hybrid* approaches attempt to combine both structural and generic principles [12], [13], [14]. Lastly, the *Deep Learning* category comprises the most recent techniques proposed for retinal image quality assessment [15], [16]. In this case, large databases labeled with quality information are supplied to a deep neural network, which automatically learns optimal image representations for the task at hand. The main disadvantage of *Deep Learning* techniques is the need for a large set of labeled data, and the typical lack of interpretability of the system's output.

TABLE I: Comparison between EyeQual and other image quality assessment approaches.

	Low Data Req.	*Speed (Test)*	*No Feature Eng.*	*Scalability*	*Explainability*
Structural [7], [4], [6]	✓				✓
Generic [9], [8]	✓	✓		✓	
HVS [10], [11]	✓	✓		✓	✓
Hybrid [12], [13], [14]	✓			✓	✓
DL [15], [16]		✓	✓	✓	
EyeQual	✔	✔	✔	✔	✔

[1]https://github.com/costapt/EyeQual

B. Multiple Instance Learning

The Multiple Instance Learning (MIL) framework is a generalization of standard supervised learning in which the assumption of having one label for each sample (or *instance*) is relaxed. In this case, we consider that a collection of instances is grouped into a single set to which a label has been assigned. This set is called a *bag of instances*, and its label influences the behavior of the classifier regarding instances contained on it. The goal of the learning process is then to infer instance level predictions out of bag-level labels. Note that a given bag may contain an arbitrary number of elements. Hence, a key difference between standard supervised learning and MIL is that while supervised learning methods map a fixed-size feature vector describing a given instance into a prediction, MIL methods map sets containing a variable number of feature vectors, corresponding to all the instances within a bag, into a single label.

Fundamentals of MIL: Most MIL methods follow a common hypothesis known as the *Standard MIL Assumption* [17]. According to this principle, all the instances inside a negative bag are negative, while a positive bag contains at least one positive instance and an arbitrary number of negative instances. To optimize a model based on the Standard MIL Assumption, a metric called *Diverse Density* was proposed in [18] based on computing the difference between the intersection of positive bags and the union of negative bags. In [19] two SVM-based formulations were proposed for the MIL problem: one that maximizes the margin between instances, named *mi-SVM*, and another that maximizes the margin between bags, named *MI-SVM*. The former trains a model to classify instances as follows: it assigns a negative label to all the instances within negative bags, while at the same time it imputes the labels of the instances inside positive bags, imposing that there is always at least one positive label on them. The model is trained iteratively in an Expectation-Maximization fashion until convergence. The *MI-SVM* also trains an instance classifier but it aggregates the classification of all the bag's instances by selecting the maximum among every prediction.

MIL for medical images: In the field of Computer-Aided Diagnosis for eye fundus images, MIL-based methods have been successfully proposed for different tasks [20], [21], [22]. In particular, techniques based on the *Bag-of-Visual-Words* [23] have been widely used for Diabetic Retinopathy detection [24], [25], [21]. In these approaches, a relaxed version of the *Standard MIL Assumption* is applied, capable of learning richer relationships between input instances. To achieve this, instances are first encoded into a latent representation. Representations extracted from all the instances are then pooled into a single feature vector corresponding to a bag, and finally a classifier is trained to distinguish between negative and positive bags. In this way, these methods can learn a decision function based on the presence/absence of different instances and, as such, are more general than the *Standard MIL Assumption*. Unfortunately, a relevant disadvantage of these approaches is that they are not as interpretable as methods that learn an instance classifier. Although some efforts have been made to make the BoVW more interpretable [21], current methods are unable to establish which positive instances correspond to actual lesions.

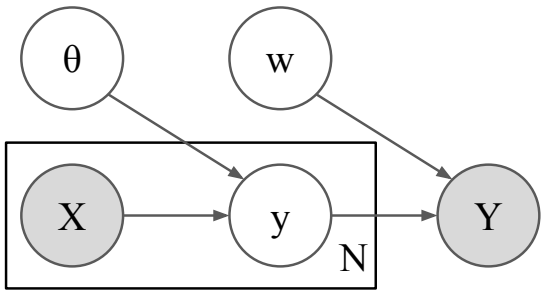

Fig. 2: **Proposed graphical model.** Only the N input instances X and the bag label Y are observed. Each instance has an associated label y that depends on parameters θ. The instance's labels y are combined by means of a pooling function parameterized by w to produce Y.

EyeQual on the other hand, is a MIL method that is able to learn with small datasets (*low data requirements*), is fast at inference time, does not require any feature engineering, scales to large datasets and provides an explanation of its decision. A comparison of EyeQual with the existing methods is shown in Table I.

III. Proposed EyeQual Method

The approach proposed in this paper belongs to the category of *Deep Learning* techniques. However, our method is carefully designed to avoid the dependence on a large database of labeled images while also providing interpretable results, by means of the MIL framework.

We start by formalizing our method by a graphical model view and then we show how to apply it to the image quality assessment problem. Finally we propose a pooling function that is better suited for the task of retinal image quality assessment than the standard Max or Average Pooling.

A. Graphical Model

In this work, we propose a method to train an instance classifier with bag labels only (*i.e.* training a patch classifier using only image labels). For that, the instances' labels y_i are treated as latent variables which are then combined by a pooling function f parameterized by w to infer the bag's label $Y = f(y_1, ..., y_N; w)$. Therefore, it is important to carefully design the pooling function in order to properly encode the relationship between the instances and the bag labels.

The graphical model depiction of our approach is shown in Figure 2. We define X as a random variable representing the set of input instances, θ as the parameters of the instance classifier and $P(y_i|X_i, \theta)$ as a Bernoulli distribution although, in principle, it could follow any other distribution such as a categorical distribution. The choice of this distribution is tied with the problem to solve and influences the design of the subsequent pooling function f. We focus on problems

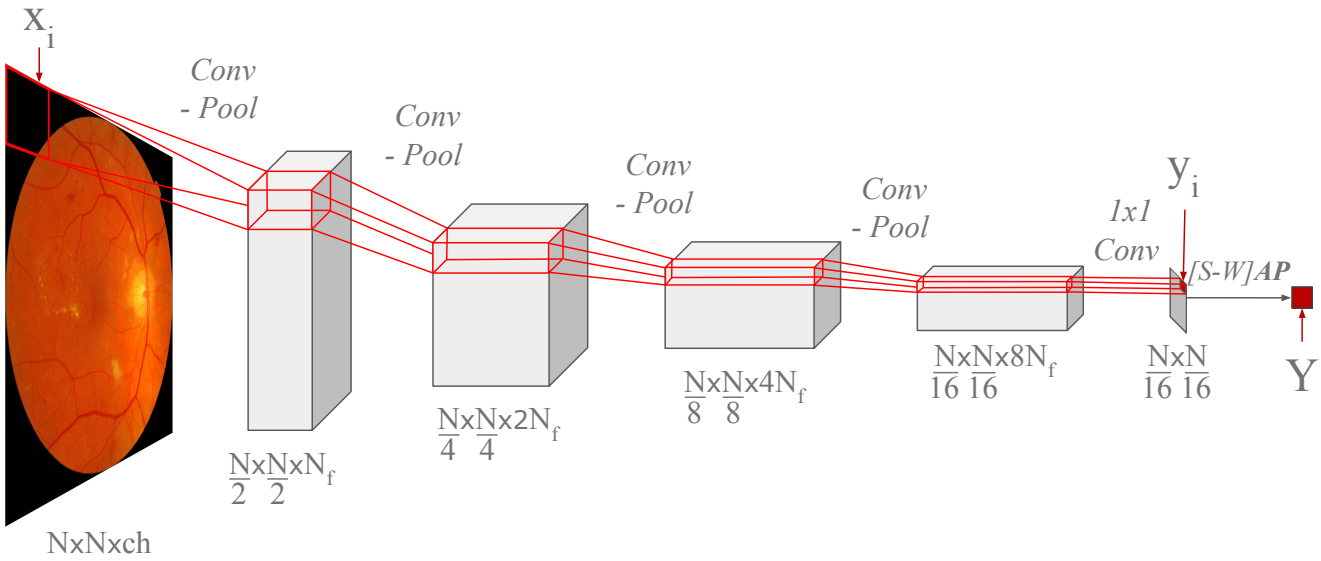

Fig. 3: **Proposed EyeQual method.** Convolutional Layers with 3×3 kernels are followed by max-pooling layers until the receptive field of the last layer achieves the desired patch size. Note that y_i only depends on the patch x_i. Then, the instance's labels are combined using a pooling function (*i.e.* Shifted Weighted Average Pooling) into the image label Y. In this work we used 512×512 RGB input images ($N = 512$ and $ch = 3$) and $N_f = 64$.

where each input instance has a binary label. Moreover, Y and (X, θ) are conditionally independent given y, meaning that the label of the bag is completely determined by the labels of the instances and w. This allows us to write the likelihood of Y as $P(Y|y_1, ..., y_N; w)$. The goal is, then, to find the parameters (θ, w) that maximize the likelihood:

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

There are some design choices that need to be considered:

Choice #1 How to define the input instances X? We could use wavelets, patches or any other feature extraction method.

Choice #2 What learning algorithm should be used to model $P(y_i|x_i, \theta)$? Some choices include SVM and logistic regression.

Choice #3 What pooling function f should be used? For instance, Sum Pooling, Max Pooling or Average Pooling could be used.

Since these choices depend on the problem to solve, we decided to test this model on the problem of image quality assessment from eye fundus images. For that we chose to *#1* use patches of the input eye fundus image as instances X and *#2* use a Convolutional Neural Network (CNN) to perform the patch classification. Finally, we devised a novel pooling function, derived from the Average Pooling (AP), for choice *#3*.

B. Instance Feature Learning and Classification

We answer choices *#1* and *#2* are related. CNNs have been used with great success for image classification problems. These models are able to extract features from raw data often achieving superior results compared to feature engineering approaches. However, as CNNs can easily learn irrelevant features, they usually require large amounts of data to avoid overfiting. To minimize this issue, some works use CNNs to classify patches of the images making it impossible for the network to correlate two pixels that are far away from each other in the image.

Instead of extracting patches from the image and using the same CNN model on different patches, we use a Fully Convolutional Network (FCN) [26] to perform patch classification given the full input image. We achieve this by realizing that the receptive field of each layer grows as the network gets deeper. For instance, the receptive field of a 3×3 Convolution Layer is, indeed, 3×3, while two 3×3 Convolution Layers have a 5×5 receptive field. Therefore, we apply Convolutions followed by pooling layers until the receptive field of the last layer reaches the desired patch size. However, unlike FCNs, we do not upsample the activation maps back to the size of the input image but apply a 1×1 convolution followed by a sigmoid function to obtain the label of each patch. When there is overlap between patches, it is more efficient to apply the model to the entire image than to each patch. The architecture of the model is shown in Figure 3.

This architecture has the disadvantage that the input patches are not independent from each other which may result in a decreased convergence rate while training the network with backpropagation [27]. However, it has been shown that using the full image to train a FCN has the same convergence rate than sampling patches to train the same network [26] and, due to increase in computational efficiency, training with the full image ends up being faster.

After computing the patch labels, we need to combine them

to produce the image label. When the problems follow the *Standard MIL Assumption* the max-pooling function can be used [19], [21]. However, the image quality assessment does not follow this assumption as some good quality images contain bad quality patches. To overcome this issue we developed a new MIL assumption for the image quality problem. We assume that a high quality image contains more high quality patches than low quality ones, and that a low quality image contains more low quality patches than high quality ones. This can be modeled by the average function $Y = \frac{1}{N}\sum_{i=1}^{N} y_i$.

We can see that $Y > 0.5$ when there are more low quality patches ($y_i = 1$) than high quality ones and that $Y < 0.5$ when there are more high quality patches than low quality ones.

C. Proposed Shifted Weighted Average Pooling

All eye fundus images are centered, having a circular region called Field-of-View (FoV) where all the anatomical structures are visible. Patches outside the FoV are always dark and, therefore, are not relevant to discriminate between high/low quality images. However, with average pooling, all patches contribute equally to the decision. Moreover, since a black region inside the FoV should be considered as low quality, patches outside the FoV end up being classified as low quality, artificially raising the final score of the image. This introduces competition between the two classes that hinders the optimization process and reduces the quality of the patch classifier.

As an intermediate step, we introduce a Weighted Average Pooling (WAP) function to solve this issue. By assigning a non-negative weight $w_i \geq 0$ to all the input patches it is possible to model the importance of each region of the image $Y = \frac{\sum_i^N w_i y_i}{\sum_i^N w_i}$.

WAP improves on Average Pooling by giving a larger weight to more discriminative regions of the retina. However, it still makes the assumption that the (weighted) number of high quality discriminative patches needs to be larger than the number of low quality discriminative patches in order to classify the image as high quality. A small number of low quality patches inside the FoV might be sufficient to classify the image as low quality. Therefore, we introduce a shifted version of WAP that not only learns what are the discriminative regions of the image, but also learns the ratio of low/high quality patches needed inside this region to reach a decision.

We do this by realizing that the WAP is a linear classifier. By setting $s = \sum_i^N |w_i|$, the model follows the following decision:

$$Y = \begin{cases} 1 & \text{if } \sum_i^N \frac{|w_i|}{s} y_i \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The differences between the decision function of WAP and a logistic regression classifier are threefold: *1)* WAP has a constant bias $b = -0.5$ while the logistic regression is able to learn it; *2)* the weights of the WAP are scaled by s prior to the multiplication by y while the logistic regression uses the sigmoid function to normalize its output and; *3)* the weights in the logistic regression are allowed to be negative.

TABLE II: **EyeQual outperforms competing methods.** Results on the ARSN dataset.

Method	Accuracy
Galdran *et al.* [11]	83.35%
VGG [28]	96.88%
Remeseiro *et al.* [8]	99.09%
EyeQual - Average Pooling	98.51%
EyeQual - S-WAP	**100.00%**

Both models have desirable properties for the image quality assessment problem. It is important to constrain the weights to be nonnegative, otherwise the model could potentially learn that a low quality patch in a certain region of the retina increases the likelihood of the image being high quality. On the other hand, the bias term allows the model to learn the ratio between low and high quality patches inside the discriminative region of the retina needed to reach a decision. By combining the advantages of both models, we get to the Shifted Weighted Average Pooling (S-WAP):

$$\tilde{Y} = \sum_i^N |w_i| y_i + b, \quad (3)$$

$$Y = \frac{1}{1 + e^{-\tilde{Y}}}. \quad (4)$$

S-WAP models the image quality assessment problem better than Average Pooling and, as we will show in the evaluation section, improves on both the results and the quality of the produced heatmaps.

IV. Results and Evaluation

In this section we answer the following questions:

Q1. **Accuracy**: How accurate is EyeQual compared with competing methods?

Q2. **Explainability**: How well do the EyeQual's results explain its image level decision?

Q3. **Efficiency**: What is the inference speed of EyeQual?

To answer these questions, we evaluated our method on two datasets: *1)* ARSN and *2)* DRIMDB [13].

The *ARSN* dataset was collected by the Portuguese *Health Authority of the Norther Region* (ARSN) that screens diabetic patients from the northern region of Portugal for signs of DR. Eye fundus images that were successfully used by ophthalmologists to diagnose patients were labeled as high quality, while images from undiagnosable exams were labeled as low quality. This dataset is proprietary and contains 330 images (183 high and 147 low quality). The dataset was randomly divided into a train (211 images), validation (52) and test (67) sets.

The *DRIMDB* is a public dataset that contains eye fundus images labeled in 3 classes: high quality, low quality and outlier. We discarded the outlier images and only used the remaining 194 high (125) and low (69) quality images. Again, we further randomly divided this dataset into a train (124 images), validation (31) and test (39) sets.

TABLE III: **EyeQual outperforms competing methods.** Results on the DRIMDB dataset.

Method	Accuracy
VGG [28]	90.47%
Sevik *et al.* [13]	98.08%
Galdran *et al.* [11]	98.40%
EyeQual - Average Pooling	**100.00%**
EyeQual - S-WAP	**100.00%**

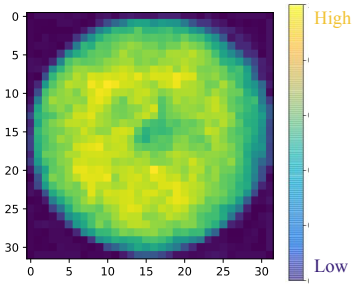

Fig. 4: **S-WAP discards patches outside the FoV.** The figure shows the weights learned by EyeQual with S-WAP when trained on the ARSN dataset. Patches outside the FoV are given a weight close to 0 while patches inside the FoV are given a similar weight.

A. Q1 - Accuracy

We trained EyeQual with *Early Stopping*, by ending the training when the validation loss stopped improving. We then report our results on the test set using the parameters that achieved the lowest loss on the validation set. To properly evaluate the method, we trained EyeQual with several pooling functions and compared with other published methods. Results on the *ARSN* dataset are shown in Table II and on the *DRIMBD* in Table III.

We evaluated EyeQual with two different pooling functions f: Average Pooling and S-WAP. We also tried using the Max Pooling function but the accuracy was consistently below 90% on the *ARSN* test set, confirming that the image quality assessment problem does not follow the *Standard MIL Assumption*. On the other hand, the Average Pooling function obtains comparable results to other state-of-the-art methods in the *ARSN* dataset while S-WAP is able to improve the results achieving 100% accuracy. For DRIMDB, Average Pooling also achieves 100% accuracy. This is due to DRIMDB images being cropped to the FoV (Figure 6 and without black borders, which turns every image patch into a discriminative instance.

EyeQual also outperforms VGG [28], which is a widely used Deep Learning architecture for image classification. The gap in performance between EyeQual and VGG increases as the amount of training data decreases. While VGG performs reasonably in the *ARSN* dataset, it gets low accuracy on the smaller *DRIMDB* dataset. On the other hand, EyeQual is robust to the amount of training data achieving the same results on both datasets: 100% accuracy. This supports our claim that, by focusing on image patches, EyeQual avoids overfiting.

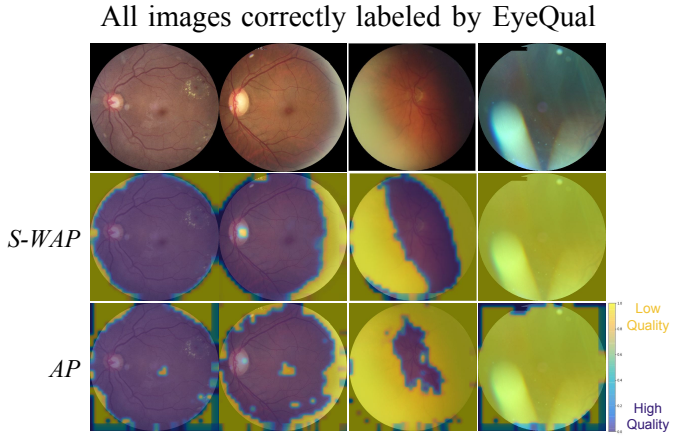

Fig. 5: **EyeQual explains.** Results on the *ARSN* dataset. Yellow regions are considered by the method as low quality. S-WAP produces more meaningful heatmaps than Average Pooling.

B. Q2 - Explainability

We can explain the results at two levels: at a dataset level and at an image level. In the former case, S-WAP can explain what are the regions of the dataset images that are more discriminative and, more importantly, in the latter EyeQual can pinpoint the regions of a given image that are of low quality.

To get the important regions of the retina one can simply plot the weights that S-WAP learnt. As shown in Figure 4, S-WAP learns to give a higher weight to patches inside the FoV. Patches outside the FoV are always dark and, therefore, are not informative. Moreover, patches inside the FoV are assigned with similar weights, with the exception of patches near the border that get a slightly lower weight. The explanation for this is that, due to the spherical shape of the retina, all images acquired with a fundus camera show certain darkness within the borders of the FoV, independently of their degree of quality.

In Figure 5 we show the results of the EyeQual's patch classifier when trained with S-WAP and Average Pooling. It is possible to see that both pooling methods output meaningful heatmaps. The dark regions outside the FoV tend to be classified as low quality patches since a dark patch inside the FoV is a low quality one. This is why the Average Pooling is not capable of producing results as good as the S-WAP that weights each patch contribution taking into account its location. S-WAP heatmaps tend to be smoother and more meaningfull.

C. Q3 - Efficiency

Inference with EyeQual is also fast. Deep Learning methods, such as EyeQual and VGG, are able to exploit GPU's ability to perform parallel computations which allows for huge speedups in inference times. We ran the Deep Learning experiments (EyeQual and VGG) on a laptop with a mobile Nvidia GTX 1060 GPU. Both models were implemented with

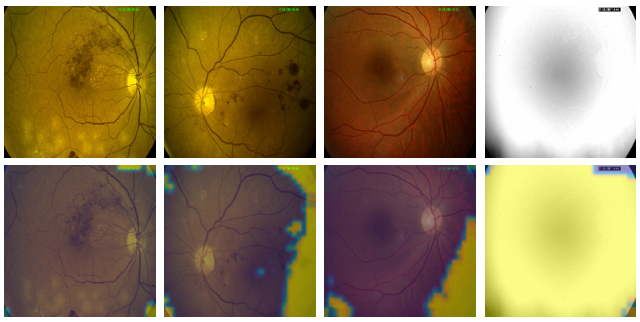

Fig. 6: **EyeQual is general.** Results on a different dataset (DRIMDB) dataset using S-WAP.

TABLE IV: **EyeQual is fast.** Mean inference time in milliseconds with 95% CI for a single image.

	EyeQual	VGG	Remeseiro *et al.* [8]
Time (ms)	36 ± 9	79 ± 9	470

Keras using the Theano backend. Remeseiro *et al.*'s [8] method was run on an Intel(R) Core(TM) i5-2400 CPU @ 3.10GHz, however, the comparison is still fair as the method is not able to exploit GPUs. Table IV shows that EyeQual is effectively instantaneous on a laptop with a stock GPU, being faster than both VGG and Remeseiro *et al.* [8]. Moreover, EyeQual still takes less than one second when running on a stock CPU (611 ± 16) while VGG takes more than two seconds (2272 ± 136).

V. Conclusions

We presented EyeQual, which addresses the problem of detecting low quality retinal images. The main idea is to learn a patch classifier using only image labels.

The main advantages of the method are:

- *Novelty*: we propose a new method that is carefully designed to learn a patch classifier despite the fact that it is only trained with image labels.
- *Explainability*: EyeQual pinpoints the region(s) of low image quality as shown in Figure 5.
- *Accuracy*: it is the only method to achieve 100% accuracy on two datasets (Table II and III).
- *Speed*: inference is faster than competing state-of-the-art methods (Table IV).

We tested EyeQual on two datasets and we show that not only do we achieve better results than all competing methods, but we can also explain the decision of the method by pinpointing the low quality patches. Our method is faster and more robust to the size of the dataset than standard deep CNN architectures like VGG.

In the future we want to evaluate our method on a larger dataset before deploying it in real screenings. As EyeQual is effectively instantaneous running on a stock GPU, it could be implemented directly on a fundus camera in order to provide imediate feedback to the technician on whether she should take another picture. Also, the heatmaps produced by EyeQual should be visually validated by ophthalmologists.

Acknowledgments

This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project CMUP-ERI/TIC/0028/2014.

References

[1] R. Kawasaki, N. Cheung, J. J. Wang, R. Klein, B. E. Klein, M. F. Cotch, A. R. Sharrett, S. Shea, F. A. Islam, and T. Y. Wong, "Retinal vessel diameters and risk of hypertension: the Multiethnic Study of Atherosclerosis," *Journal of hypertension*, vol. 27, no. 12, pp. 2386–2393, Dec. 2009.

[2] J. K. H. Lim, Q.-X. Li, Z. He, A. J. Vingrys, V. H. Y. Wong, N. Currier, J. Mullen, B. V. Bui, and C. T. O. Nguyen, "The Eye As a Biomarker for Alzheimer's Disease," *Frontiers in Neuroscience*, vol. 10, Nov. 2016.

[3] R. Taylor and D. Batey, *Handbook of retinal screening in diabetes: diagnosis and management*. John Wiley & Sons, 2012.

[4] M. Niemeijer, M. D. Abràmoff, and B. van Ginneken, "Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening," *Medical Image Analysis*, vol. 10, no. 6, pp. 888–898, Dec. 2006.

[5] A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, and P. F. Sharp, "Automated Assessment of Diabetic Retinal Image Quality Based on Clarity and Field Definition," *Investigative Ophthalmology & Visual Science*, vol. 47, no. 3, pp. 1120–1125, Mar. 2006.

[6] T. Köhler, A. Budai, M. F. Kraus, J. Odstrčilik, G. Michelson, and J. Hornegger, "Automatic no-reference quality assessment for retinal fundus images using vessel segmentation," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, Jun. 2013, pp. 95–100.

[7] S. C. Lee and Y. Wang, "Automatic retinal image quality assessment and enhancement," vol. 3661, 1999, pp. 1581–1590.

[8] B. Remeseiro, A. Mendonça, and A. Campilho, "Objective Quality Assessment of Retinal Images Based on Texture Analysis," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 4520–4527.

[9] J. M. Pires Dias, C. M. Oliveira, and L. A. da Silva Cruz, "Retinal image quality assessment using generic image quality indicators," *Information Fusion*, vol. 19, pp. 73–90, Sep. 2014.

[10] S. Wang, K. Jin, H. Lu, C. Cheng, J. Ye, and D. Qian, "Human Visual System-Based Fundus Image Quality Assessment of Portable Fundus Camera Photographs," *IEEE Transactions on Medical Imaging*, vol. 35, no. 4, pp. 1046–1055, Apr. 2016.

[11] A. Galdran, T. Araújo, A. Mendonça, and A. Campilho, "Retinal Image Quality Assessment by Mean-Subtracted Contrast-Normalized Coefficients," in *Computational Vision and Medical Image Processing VI, 6th ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing - VipIMAGE 2017*. CRC Press, Oct. 2017.

[12] J. Paulus, J. Meier, R. Bock, J. Hornegger, and G. Michelson, "Automated quality assessment of retinal fundus photos," *International Journal of Computer Assisted Radiology and Surgery*, vol. 5, no. 6, pp. 557–564, Nov. 2010.

[13] U. Şevik, C. Köse, T. Berber, and H. Erdöl, "Identification of suitable fundus images using automated quality assessment methods," *Journal of Biomedical Optics*, vol. 19, no. 4, p. 046006, Apr. 2014.

[14] L. Abdel-Hamid, A. El-Rafei, S. El-Ramly, G. Michelson, and J. Hornegger, "Retinal image quality assessment based on image clarity and content," *Journal of Biomedical Optics*, vol. 21, no. 9, 2016.

[15] D. Mahapatra, P. K. Roy, S. Sedai, and R. Garnavi, "Retinal Image Quality Classification Using Saliency Maps and CNNs," in *Machine Learning in Medical Imaging*. Springer, Oct. 2016, pp. 172–179.

[16] S. K. Saha, B. Fernando, J. Cuadros, D. Xiao, and Y. Kanagasingam, "Deep Learning for Automated Quality Assessment of Color Fundus Images in Diabetic Retinopathy Screening," *arXiv:1703.02511 [cs]*, Mar. 2017.

[17] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[18] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, pp. 570–576, 1998.

[19] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," *Advances in neural information processing systems*, pp. 577–584, 2003.

[20] F. Khalvati, J. Zhang, A. Wong, and M. A. Haider, "Bag of bags: Nested multi instance classification for prostate cancer detection," in *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE, 2016, pp. 146–151.

[21] P. Costa and A. Campilho, "Convolutional bag of words for diabetic retinopathy detection from eye fundus images," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 10, 2017.

[22] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE reviews in biomedical engineering*, 2017.

[23] J. Sivic, A. Zisserman *et al.*, "Video google: A text retrieval approach to object matching in videos." in *iccv*, vol. 2, no. 1470, 2003, pp. 1470–1477.

[24] G. Quellec, M. Lamard, A. Erginay, A. Chabouis, P. Massin, B. Cochener, and G. Cazuguel, "Automatic detection of referral patients due to retinal pathologies through data mining," *Medical image analysis*, vol. 29, pp. 47–64, 2016.

[25] R. Pires, S. Avila, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Beyond lesion-based diabetic retinopathy: a direct approach for referral," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 193–200, 2017.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[27] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 1998, pp. 9–48.

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