# FunPiQ: A New Benchmark for Pixel-Level Quality Assessment in Fundus Images

Pengwei Wang[1](✉)[0009−0000−5036−3377], José Morano[1,2][0000−0003−3785−8185], Virginia Mares[1][0000−0001−6127−6770], and Hrvoje Bogunović[1,2][0000−0002−9168−0894]

[1] Institute of Artificial Intelligence, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria
[2] Christian Doppler Lab for Artificial Intelligence in Retina, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria
{pengwei.wang,jose.moranosanchez,hrvoje.bogunovic}@meduniwien.ac.at

**Abstract.** Color fundus photography (CFP) is the most common ophthalmic imaging modality for large-scale screening. However, it is highly susceptible to degradations, making robust fundus image quality assessment (FIQA) crucial. The criteria for what constitutes high-quality at the image level vary across clinical tasks, making FIQA dependent on expert knowledge. This motivated the development of automated methods and datasets. While existing datasets aim to standardize image-level quality, their criteria often differ. Furthermore, image-level labels preclude the quantitative evaluation of localized degradations, which is essential for trustworthy FIQA. We argue that pixel-level FIQA based on anatomical visibility represents a more task-agnostic, explainable approach. In this work, we introduce FunPiQ, the first FIQA benchmark to provide pixel-level quality annotations. In addition, we propose EFIQA-CP, an explainable-by-design (EBD) method that uses quality pseudo-labels based on anatomical visibility to train a CNN via Non-Negative Positive-Unlabeled learning. Extensive evaluations of classification methods with post-hoc explanations, anomaly detection methods, and EBD methods demonstrate the superior performance of the last and, particularly, of EFIQA-CP. Our code, weights, and dataset are publicly available at https://github.com/penway/FunPiQ

**Keywords:** Image quality assessment · Retina · Benchmark

## 1 Introduction

Color fundus photography (CFP) is one of the most widely used imaging modalities in ophthalmology [7], especially for screening. CFP enables the discovery not only of retinal diseases but also of systemic health conditions [32,11]. In addition, it is arguably the most accessible ophthalmic modality, with devices ranging from professional desktop systems to portable, mobile-phone-based cameras, making large-scale screening feasible in remote areas [23].

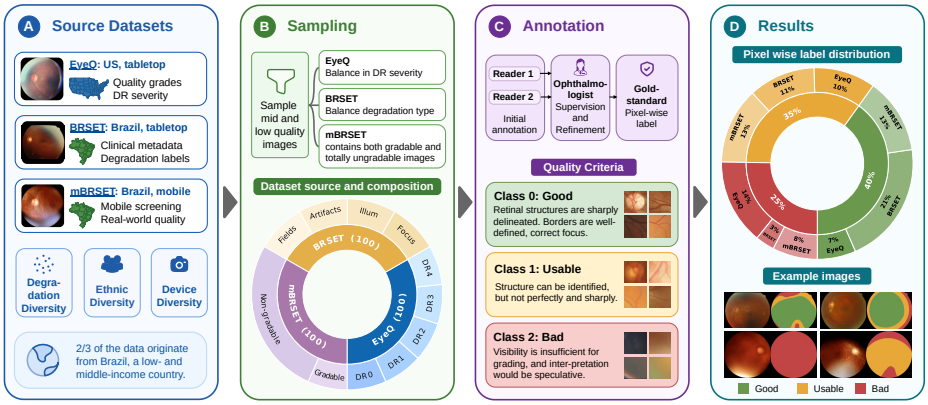


**Fig. 1.** Overview of our FunPiQ dataset.

However, CFP is highly susceptible to degradations, such as inconsistent illumination and poor contrast [8]. These issues often affect *local* regions rather than the entire image, meaning certain areas remain of high quality while others are degraded (see Fig. 1). Such degradations are frequent even in professional settings. For example, in the EyePACS benchmark [6], 42% of images had quality issues, with 19% deemed unsuitable for diagnosis [8]. These challenges are further exacerbated when using portable fundus cameras in uncontrolled environments. Thus, Fundus Image Quality Assessment (FIQA) represents a crucial step.

Nonetheless, FIQA faces important challenges, primarily due to the lack of standardized criteria. The definition of *image-level* quality is inherently tied to specific downstream tasks, as the visibility of certain anatomical structures may be critical for diagnosing some diseases while remaining less relevant, or irrelevant, for others. This task-dependency makes manual, image-level FIQA partly subjective, time-consuming, and dependent on expert knowledge, rendering it unfeasible for large-scale screening. Consequently, there is a growing need for automated methods and datasets.

Table 1 summarizes the most relevant datasets for FIQA. While these benchmarks provide various quality-related labels, ranging from global classes to specific degradation types, they implement distinct criteria for what constitutes "quality". As previously discussed, a universal, image-level standard remains elusive because diagnostic requirements vary by disease; consequently, some images labeled as *bad* quality in these datasets are actually usable for specific clinical tasks. For example, in MSHF [14], images flagged as *bad* quality were still deemed suitable for glaucoma screening. Furthermore, the separate standardization efforts have led to fairly dissimilar classifications across benchmarks, evidenced by the suboptimal generalization of models on external tests [29], and by differences in quality labels assigned to the same images by independent teams—for instance, an 11% difference in a subset of EyePACS [8,31]. Given the inherently local nature of degradations, we argue that defining quality locally, through the

**Table 1.** Summary of current CFP datasets with quality labels, including classes (C) and degradation types (DT). 2C includes *good* and *bad*, while 3C adds *usable*. FunPiQ is the first evaluation dataset providing pixel-level quality annotations.

| Granularity | Dataset | Year | Size | Country | Quality Labels |
|---|---|---|---|---|---|
| **Image-level** | DRIMDB [26] | 2014 | 216 | Turkey | 2C and *outlier* class |
| | EyePACS-Q [31] | 2018 | 88,702 | US | 2C |
| | EyeQ [8] | 2019 | 30,000 | US | 3C |
| | DeepDRiD [17] | 2022 | 2,256 | China | 2C and 0–10 score with 3DT |
| | MSHF [14] | 2023 | 1,302 | China | 2C and 3DT |
| | BRSET [21] | 2024 | 16,266 | Brazil | 2C and 4DT |
| | mBRSET [30] | 2025 | 5,164 | Brazil | 2C |
| | FQS [10] | 2025 | 2,246 | China | 3C and mean opinion score |
| **Pixel-level** | ***FunPiQ*** | 2026 | 300 | US, Brazil | 3C |

visibility of underlying anatomy, offers a more task-agnostic, explainable alternative. This approach can be coupled with specific anatomical requirements (e.g., optic disc visibility) to implement dynamic image-level quality criteria.

Most FIQA methods rely on supervised deep learning on the aforementioned datasets, incorporating various architectural innovations. For instance, MCF-Net utilizes multi-color space fusion [8]; FGR-Net, reconstruction-based representations [15], and Shen et al. use semi-tied adversarial domain adaptation for domain-invariant features [27]. However, these approaches depend on image-level labels, inheriting the biases of their training datasets. Furthermore, while post-hoc visualization techniques like Grad-CAM [25] are frequently used [27,15], the explainability provided by these tools is limited, and the localization of degradations is generally imprecise. To address the latter, EyeQual [4] proposes an explainable-by-design architecture that outputs local-level quality maps via patch-level scoring. Aiming to address both issues, EFIQA [29] proposed using vessel unsupervised anomaly detection (VUAD) as a prior for producing local-level quality maps without the need for quality-related labels. While it showed superior performance on external tests, EFIQA still presents some limitations. First, it relies on a single linear adapter for patch-level quality scores from foundation model features, with no global context. This leads to incorrect predictions in areas with naturally low vessel density, such as the fovea. Second, it is trained on the noisy pseudo-labels from the VUAD model via a simple Binary Cross-Entropy (BCE) loss, leading to pseudo-label overfitting. Finally, despite the recognized importance of explainability and localization [15,27], no benchmark currently exists to quantitatively evaluate the precision of these outputs.

In this context, the contributions of this work are threefold:

1. We introduce the Fundus Pixel Quality (FunPiQ) dataset, the first public benchmark dedicated to pixel-level quality assessment for CFP. The benchmark comprises 300 images sourced from three public datasets and annotated by expert readers supervised by a board-certified ophthalmologist.

2. We propose EFIQA-CP, a method that mitigates EFIQA's limitations by using a shallow CNN with a large receptive field as feature adapter and Non-Negative Positive-Unlabeled (nnPU) learning [16] to handle pseudo-labels.
3. We conduct a thorough evaluation of existing and proposed methods on our benchmark, demonstrating the superior performance of EFIQA-CP and highlighting the potential of explainable-by-design approaches for implementing robust, local-level quality assessment systems.

## 2 Materials and Methods

### 2.1 Dataset

*Data Selection.* Fig. 1 shows an overview of the dataset. To maximize data diversity in terms of degradation types, ethnicity, and acquisition devices, we sampled from three existing benchmarks from two different countries: EyeQ [8] (US), BRSET [21,20,9], and mBRSET [30,22,9] (both Brazil). Thus, 2/3 of the data come from a low- and middle-income country (LMIC), typically underrepresented in benchmarks. Moreover, these datasets include images captured using a wide range of hardware, from high-end tabletop fundus cameras to mobile phone-based portable devices. Since FunPiQ is designed to evaluate the localization of degradations by IQA methods (i.e. their explainability) task-agnostically (not mirroring specific prevalences), for all three datasets, we selected the images labeled as *bad*. This produces a mix of good, usable, and bad regions within each image (Fig. 1 D), so that every sample tests a model's ability to differentiate quality locally. For EyeQ, we balanced the dataset by selecting all types of DR level. For BRSET, we balanced among degradation types (illumination, focus, field, and artifact). For mBRSET, some images have diagnosis, while others are not gradable. We sampled all the gradable ones, and randomly sampled the rest.

*Labeling.* Images were annotated by expert readers and supervised by a board-certified ophthalmologist. Quality was assessed at the pixel level across the entire fundus image, based on the visibility and interpretability of anatomical or pathological features. Three quality categories were defined: *good* (regions of optimal visibility with sharply delineated retinal features and no relevant artifacts), *usable* (regions with mild reduction in sharpness or contrast without critical impact on structural identification), and *bad* (non-gradable regions due to severe artifacts, signal dropout, shadowing, media opacity, or marked defocus).

### 2.2 Method

Our method follows a similar overall pipeline to EFIQA [29], and utilizes its VUAD model to generate local quality pseudo-labels. These labels are then used to train an adapter that maps features from a pretrained foundation model to final quality maps. However, these pseudo-labels are inherently noisy, as they rely solely on the absence of visible vessels. In EFIQA, the adapter is a linear layer

that takes the features of a single patch, with no global context, and outputs its quality score. This setup, together with the noisy pseudo-labels, causes the network to flag as bad quality regions with naturally low vessel density (e.g., the macula). To mitigate these issues, we propose a wide-yet-shallow network that integrates spatial information through a broader receptive field. Furthermore, we employ nnPU learning to address the label distribution problem and improve robustness against the noise of the pseudo-labels.

*Adapter Architecture.* We extract features from a frozen DINOv3 backbone [28]. The network processes these features through five ConvNeXt-style blocks utilizing $7 \times 7$ depthwise convolutions, GELU activations [13], and channel-wise Layer Normalization [1]. Different from standard setting, we use the MLP ratio as 1 instead of 4 to limit the capacity of the model and avoid overfitting.

*nnPU.* To handle noisy pseudo-labels, we apply a strict threshold to isolate reliable low-quality pixels, treating the rest as unlabeled. This motivates a Positive-Unlabeled (PU) learning approach, which effectively learns an unbiased classifier without requiring guaranteed high-quality (negative) labels. Furthermore, because highly flexible neural networks can overfit in this setup and push the estimated risk of the hidden negatives below zero, we utilize the non-negative PU (nnPU) framework [16] to clamp this risk at zero, stabilizing training. Using a softplus surrogate loss and a positive class prior $\pi = 0.05$, the objective is:

$$\mathcal{L} = \pi \widehat{R}_p^+ + \max\left(0, \widehat{R}_u^- - \pi \widehat{R}_p^-\right) \tag{1}$$

where $\widehat{R}_p^+$ and $\widehat{R}_p^-$ are empirical losses on positive data treated as positive and negative, respectively, and $\widehat{R}_u^-$ is the loss on unlabeled data treated as negative.

## 3 Experiments

*Benchmarked Methods.* We selected methods from three categories. First, end-to-end FIQA methods with Grad-CAM [25]: MCF-Net [8] and FGR-Net [15]. Second, explainable FIQA methods: EyeQual [4] and EFIQA [29]. Lastly, inspired by the use of Anomaly Detection (AD) techniques in EFIQA, we also trained and evaluated two AD methods: PatchCore [24] and UniVAD [12]. These models are specifically designed to identify deviations from a "normal" appearance without explicit labels, thus offering a potential solution for label-free FIQA.

*Training Details.* To keep FunPiQ solely for evaluation, all the methods that do not provide pretrained weights were trained on other datasets. End-to-end methods were trained on EyeQ, as proposed in the original papers [8,15], and the saliency map of the class *bad* was used as quality map. For EyeQual, we reimplemented the method using a stronger backbone (ConvNeXt-tiny [18]) and training dataset (MSHF) to ensure a fair comparison with more recent methods. Both modifications yielded significant performance improvements. AD methods were

trained on a subset of the 1000 images dataset [2] composed of pristine images. For MCF-Net and EFIQA official weights were used. To train the EFIQA-CP, we used an internal dataset consisting of 4064 CFP images from a single center, captured using a tabletop device (Topcon Maestro 2) with mixed quality. For training, we use AdamW [19] with a learning rate of $1 \times 10^{-3}$, 20 epochs, batch size of 16, and input resolution of $1024 \times 1024$, with no data augmentation.

*Evaluation Metrics.* Given the ordinality of labels and that all models output continuous pixel-level scores, we evaluate performance using Quadratic Weighted Kappa (QWK) [3] and the mean Dice coefficient (mDice) [5]. QWK measures the agreement between predicted and ground truth (GT) masks while accounting for the severity of disagreements in ordinal tasks. On the other hand, mDice (mean of the Dice scores for each class) measures the overlap between the predictions and the GT. In clinical applications, identifying bad regions is of primary interest for quality assurance. Consequently, we also evaluate performance on a binary "Reject" task (*good*+*usable* vs. *bad*). We report Dice, Area Under the Receiver Operating Characteristic (AUROC) and Precision-Recall Curves (AUPRC), and sensitivity at 95% specificity (S95), using *bad* as the positive class. To ensure a fair comparison, we report the optimal results achieved by each method across a range of thresholds. Statistical significance was assessed using bootstrapping with $10,000$ iterations and a sample size of 300 per replicate with replacement.

## 4 Results

### 4.1 Quantitative Results

The quantitative benchmark results are summarized in Table 2. Overall, EFIQA-CP achieves the best performance, followed by EFIQA and EyeQual, suggesting that explainable-by-design approaches are particularly effective for this task. EFIQA-CP ranks first on all metrics in BRSET and EyeQ. On mBRSET, it remains the top-performing method in terms of mDice and Reject Dice. On the merged dataset, compared with EFIQA, EFIQA-CP yields statistically significant improvements in QWK ($p < 0.0001$), mDice ($p < 0.01$), and Reject Dice ($p < 0.05$). Furthermore, EFIQA-CP significantly outperforms all other methods across these three metrics (all $p < 0.0001$). Interestingly, although EFIQA is the overall second-best method, EyeQual consistently surpasses EFIQA in QWK but not in mDice. EyeQual tends to confuse *usable* and *bad* more, with lower performance for Reject, but has better precision in the *good* class, making the overall QWK better. Among the three strongest methods, their performance becomes closer in mBRSET, the most challenging dataset; notably, EFIQA attains the best AUPRC and S95. Since mBRSET is collected with mobile devices, this observation suggests that EFIQA may be more robust to domain shift.

Among the remaining methods, MCF-Net and PatchCore perform relatively well overall. MCF-Net is stronger on BRSET, whereas PatchCore performs better on EyeQ and mBRSET. The performance of general AD methods falls short of the expectations set by their success on industrial and medical benchmarks.

**Table 2.** Comparison of quality assessment models in FunPiQ. Cell backgrounds indicate relative performance using min-max normalization within each dataset and metric (darker is better). Best results are in **bold**; second best, underlined.

| Method | BRSET | | | | | | EyeQ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3-Class | | Reject | | | | 3-Class | | Reject | | | |
| | QWK | mDice | Dice | AUROC | AUPRC | S95 | QWK | mDice | Dice | AUROC | AUPRC | S95 |
| MCF-Net | 39.54 | 48.13 | 28.61 | 80.31 | 26.49 | 22.98 | 9.73 | 35.58 | 62.35 | 54.66 | 50.57 | 8.28 |
| FGR-Net | 19.16 | 39.06 | 20.60 | 62.19 | 12.28 | 5.48 | 14.48 | 37.16 | 60.86 | 51.90 | 45.47 | 3.69 |
| PatchCore | 13.95 | 38.86 | 19.95 | 58.18 | 15.35 | 12.30 | 23.40 | 39.91 | 58.83 | 62.92 | 60.49 | 15.42 |
| UniVAD | 17.71 | 39.29 | 17.17 | 63.10 | 16.93 | 14.38 | 11.81 | 36.12 | 62.92 | 56.58 | 51.49 | 7.43 |
| EyeQual | 51.32 | 53.94 | 40.58 | 85.42 | 37.33 | 33.18 | 64.89 | 61.58 | 76.39 | 83.17 | 76.18 | 19.63 |
| EFIQA | 43.50 | 54.07 | 46.10 | 85.45 | 55.04 | 52.23 | 65.27 | 62.09 | 79.67 | 89.27 | 89.43 | 61.70 |
| *EFIQA-CP* | **60.17** | **60.68** | **56.12** | **92.10** | **55.95** | **56.98** | **72.94** | **66.04** | **83.21** | **92.68** | **91.00** | **69.23** |

| Method | mBRSET | | | | | | *Merged* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3-Class | | Reject | | | | 3-Class | | Reject | | | |
| | QWK | mDice | Dice | AUROC | AUPRC | S95 | QWK | mDice | Dice | AUROC | AUPRC | S95 |
| MCF-Net | 21.97 | 39.40 | 40.09 | 62.06 | 29.95 | 8.25 | 31.18 | 44.14 | 47.05 | 65.47 | 35.83 | 10.31 |
| FGR-Net | 13.73 | 38.53 | 37.24 | 55.49 | 26.80 | 8.22 | 20.79 | 41.01 | 42.63 | 58.87 | 30.48 | 6.06 |
| PatchCore | 23.22 | 41.69 | 40.72 | 62.99 | 36.69 | 15.66 | 26.55 | 42.96 | 44.64 | 64.48 | 40.50 | 15.68 |
| UniVAD | 0.78 | 33.90 | 37.47 | 51.73 | 27.76 | 11.15 | 13.59 | 38.33 | 40.83 | 57.63 | 33.42 | 11.58 |
| EyeQual | **62.43** | 59.03 | 57.80 | 81.65 | 47.92 | 17.34 | 66.54 | 62.42 | 65.92 | 86.43 | 64.67 | 31.73 |
| EFIQA | 55.69 | 57.68 | 65.27 | 87.96 | **73.64** | **49.24** | 63.26 | 62.67 | 71.36 | 89.45 | **80.01** | **59.01** |
| *EFIQA-CP* | 59.12 | **60.06** | **66.13** | **89.32** | 64.40 | 40.32 | **68.41** | **65.10** | **73.45** | **91.30** | 75.83 | 55.61 |

This indicates that modern AD methods may be inherently biased toward localizing compact anomalies, struggling to capture larger quality degradations. We believe that domain-specific design is still required to fully exploit AD for FIQA. Overall, EFIQA-CP is the most stable method across datasets and metrics.

### 4.2 Qualitative Results

Qualitative results are shown in Fig. 2. We observe that EFIQA-CP produces the most precise maps, with smooth regions and a better-utilized score range. In line with the quantitative results, the three explainable-by-design methods (EFIQA-CP, EFIQA, and EyeQual) consistently yield meaningful maps across all examples. EyeQual localizes well *usable* and *bad* areas, but its outputs appear strongly binarized, making *usable* vs. *bad* separation less distinct. EFIQA shows better separability, while the *good*–*usable* transition appears less accurate and the maps are noisier than EFIQA-CP. For the baselines, we observe that MCF-Net often produces plausible saliency maps but does not fully cover globally degraded images (see Fig. 2, 3$^{rd}$ row). FGR-Net appears less reliable and triggers frequent false positives. PatchCore and UniVAD highlight degraded regions, but their scores remain high across patches and the maps appear noisy. Overall, explainable-by-design methods provide better quality maps, and EFIQA-CP shows a clear advantage in spatial precision and thus practical usability.

### 4.3 Ablation Study

We conducted an ablation study to assess the impact of our architectural changes and nnPU learning on final performance. All models were trained on the same

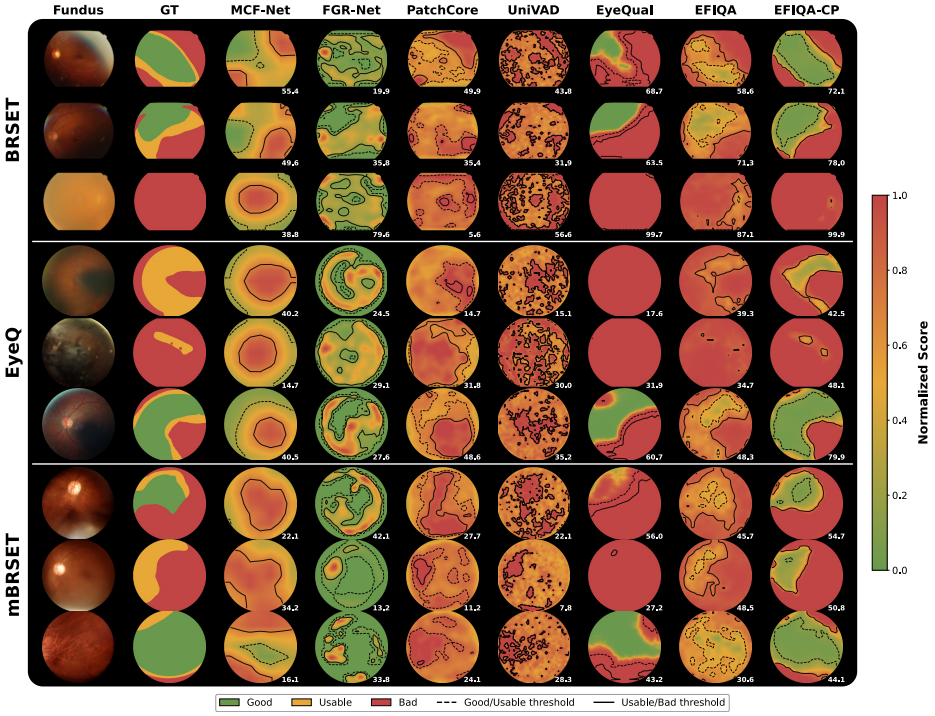


**Fig. 2.** Qualitative comparison of quality maps from each method. Lower-right numbers report per-image mDice (%), averaged over present ground-truth classes.

in-house dataset. The model using our new architecture but trained with BCE achieved 63.36 mDice, 67.00 QWK and 90.76 AUROC, outperforming EFIQA (62.67 mDice, 63.26 QWK, 89.45 AUROC) but not the full EFIQA-CP (65.10 mDice, 68.41 QWK, 91.30 AUROC). Notably, while the BCE-trained model suffered from overfitting and performance degradation during extended training, nnPU learning ensured stable convergence.

## 5 Conclusions

In this work, we introduced FunPiQ, the first dataset with pixel-level quality annotations for evaluating FIQA models and their explanations in a task-agnostic way. In addition, we proposed EFIQA-CP, a method for precise identification of degradations that leverages a wide-yet-shallow feature adapter and nnPU learning to mitigate the inherent noise of quality pseudo-labels. Our extensive evaluation in FunPiQ revealed that explainable-by-design methods hold a distinct advantage for local quality prediction, with EFIQA-CP achieving state-of-the-art performance. In the future, we plan to extend FunPiQ with more samples and multi-reader annotations to quantify inter-annotator agreement. In addition,

the combination of quality maps with anatomical localization to enable dynamic image-level quality criteria represents an interesting research direction. Beyond benchmarking explainability, pixel-level maps also have direct application value: by showing *where* quality is insufficient rather than returning an opaque global score, they support quality assurance for non-expert operators and can guide targeted image reacquisition. In addition, they could provide a strong prior for CFP enhancement, by distinguishing anatomy from artifacts, helping restoration models enhance recoverable regions while avoiding hallucinations. We believe FunPiQ benchmark will encourage research on pixel-level CFP quality detection and EBD methods, and facilitate the development of more robust and flexible FIQA methods with stronger clinical applicability.

**Acknowledgments.** Funded by the European Union (ERC, HealthAEye, 101171183). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was also supported in part by the Christian Doppler Research Association, Austrian Federal Ministry of Economy, Energy and Tourism, and the National Foundation for Research, Technology.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016)
2. Cen, L.P., Ji, J., Lin, J.W., et al.: Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nature communications **12**(1), 4828 (2021)
3. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin **70**(4), 213 (1968)
4. Costa, P., Campilho, A., Hooi, B., et al.: EyeQual: Accurate, explainable, retinal image quality assessment. In: 2017 16th IEEE International Conference on machine learning and applications (ICMLA). pp. 323–330. IEEE (2017)
5. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
6. Dugas, E., Jared, Jorge, Cukierski, W.: Diabetic Retinopathy Detection. `https://kaggle.com/competitions/diabetic-retinopathy-detection` (2015), kaggle
7. Fogel-Levin, M., Sadda, S.R., Rosenfeld, P.J., et al.: Advanced retinal imaging and applications for clinical practice: A consensus review. Survey of Ophthalmology **67**(5), 1373–1390 (2022)
8. Fu, H., Wang, B., Shen, J., et al.: Evaluation of retinal image quality assessment networks in different color-spaces. In: International conference on medical image computing and computer-assisted intervention. pp. 48–56. Springer (2019)
9. Goldberger, A.L., Amaral, L.A.N., Glass, L., et al.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000), rRID:SCR_007345

10. Gong, Z., Deng, Z., Gan, R., et al.: Acquire continuous and precise score for fundus image quality assessment: FTHNet and FQS dataset. Scientific Reports **15**(1), 40524 (2025)
11. Grzybowski, A., Jin, K., Zhou, J., et al.: Retina fundus photograph-based artificial intelligence algorithms in medicine: a systematic review. Ophthalmology and therapy **13**(8), 2125–2149 (2024)
12. Gu, Z., Zhu, B., Zhu, G., et al.: UniVAD: A training-free unified model for few-shot visual anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15194–15203 (2025)
13. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv:1606.08415 (2016)
14. Jin, K., Gao, Z., Jiang, X., et al.: MSHF: a multi-source heterogeneous fundus (MSHF) dataset for image quality assessment. Scientific data **10**(1), 286 (2023)
15. Khalid, S., Rashwan, H.A., Abdulwahab, S., et al.: FGR-Net: Interpretable fundus image gradeability classification based on deep reconstruction learning. Expert Systems With Applications **238**, 121644 (2024)
16. Kiryo, R., Niu, G., Du Plessis, M.C., Sugiyama, M.: Positive-unlabeled learning with non-negative risk estimator. Advances in neural information processing systems **30** (2017)
17. Liu, R., Wang, X., Wu, Q., et al.: DeepDRiD: Diabetic retinopathy—grading and image quality estimation challenge. Patterns **3**(6) (2022)
18. Liu, Z., Mao, H., Wu, C.Y., et al.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
20. Nakayama, L.F., Goncalves, M., Zago Ribeiro, L., et al.: A Brazilian Multilabel Ophthalmological Dataset (BRSET). PhysioNet (Mar 2023), version 1.0.0
21. Nakayama, L.F., Restrepo, D., Matos, J., et al.: BRSET: a Brazilian multilabel ophthalmological dataset of retina fundus photos. PLOS Digital Health **3**(7), e0000454 (2024)
22. Nakayama, L.F., Zago Ribeiro, L., Restrepo, D., et al.: mBRSET, a Mobile Brazilian Retinal Dataset. PhysioNet (Jun 2024), version 1.0
23. Panwar, N., Huang, P., Lee, J., et al.: Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare. Telemedicine and e-Health **22**(3), 198–208 (2016)
24. Roth, K., Pemula, L., Zepeda, J., et al.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2022)
25. Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
26. Şevik, U., Köse, C., Berber, T., Erdöl, H.: Identification of suitable fundus images using automated quality assessment methods. Journal of biomedical optics **19**(4), 046006–046006 (2014)
27. Shen, Y., Sheng, B., Fang, R., et al.: Domain-invariant interpretable fundus image quality assessment. Medical image analysis **61**, 101654 (2020)
28. Siméoni, O., Vo, H.V., Seitzer, M., et al.: DINOv3. Transactions on Machine Learning Research (2026)

29. Wang, P., Morano, J., Wan, Q., Bogunović, H.: EFIQA: Explainable Fundus Image Quality Assessment via Anatomical Priors. In: Medical Imaging with Deep Learning (2026)
30. Wu, C., Restrepo, D., Nakayama, L.F., et al.: A portable retina fundus photos dataset for clinical, demographic, and diabetic retinopathy prediction. Scientific Data **12**(1), 323 (2025)
31. Zhou, K., Gu, Z., Li, A., et al.: Fundus image quality-guided diabetic retinopathy grading. In: International Workshop on Ophthalmic Medical Image Analysis. pp. 245–252. Springer (2018)
32. Zhu, Z., Wang, Y., Qi, Z., et al.: Oculomics: Current concepts and evidence. Progress in Retinal and Eye Research **106**, 101350 (2025)