Generalized Portrait Quality Assessment

Nicolas Chahine, Sira Ferradans, Javier Vazquez-Corral, Jean Ponce

***Abstract*—Automated and robust portrait quality assessment (PQA) is of paramount importance in high-impact applications such as smartphone photography. This paper presents FHIQA, a learning-based approach to PQA that introduces a simple but effective quality score rescaling method based on image semantics, to enhance the precision of fine-grained image quality metrics while ensuring robust generalization to various scene settings beyond the training dataset. The proposed approach is validated by extensive experiments on the PIQ23 benchmark and comparisons with the current state of the art. The source code of FHIQA will be made publicly available on the PIQ23 GitHub repository at https://github.com/DXOMARK-Research/PIQ2023.**

***Index Terms*—Blind image quality assessment, Portrait quality assessment, Deep learning.**

I. Introduction

SMARTPHONES have significantly altered the landscape of photographic practices, with a notable emphasis on portrait photography. As the consumer base becomes increasingly discerning, there is a clear escalation in expectations regarding the quality of portrait images. This trend pushes a corresponding advancement in camera technology driving manufacturers to focus on improving image quality, which usually involves developing costly image quality assessment (IQA) protocols for optimizing camera performance. Consequently, automated IQA methods are employed to cut the cost of smartphone camera tuning.

Traditional IQA [11, 16, 17, 19] offers relevant insights about image quality but often doesn't capture the complexities of modern camera systems and weakly correlates with human perception [2]. Deep learning-based IQA has emerged as a promising alternative leveraging the widespread availability of modern smartphone images and image quality datasets. Popular datasets like LIVE [19], CSIQ [11], and TID [16, 17] present a large amount of synthetically distorted images, but they don't encapsulate the multifaceted nature of modern smartphone camera systems. "In-the-wild" collections, including LIVE Challenge [4], KonIQ-10k [6], and PaQ-2-PiQ [23], offer a better representation of real-world conditions, but their uncontrolled data collection and subjective labeling make them less suited for rigorous digital camera evaluations. Recently, datasets such as PIQ23 [1] have emerged, offering a setting tailored to portrait photography including diverse scenes, each independently annotated by image quality experts. The PIQ23 dataset is crafted to achieve high precision in annotations by employing two key strategies: a) utilizing pairwise comparisons for image annotation, a method known for enhancing consistency in IQA experiments [12, 15]; and b) categorizing images by content into distinct scenes to harness the human visual system's precision when evaluating images with shared content. Notably, the PIQ23 dataset distinguishes itself with a comprehensive array of portrait images sourced from over 100 smartphone models, encompassing 50 varied scenarios that represent a wide range of photographic conditions.

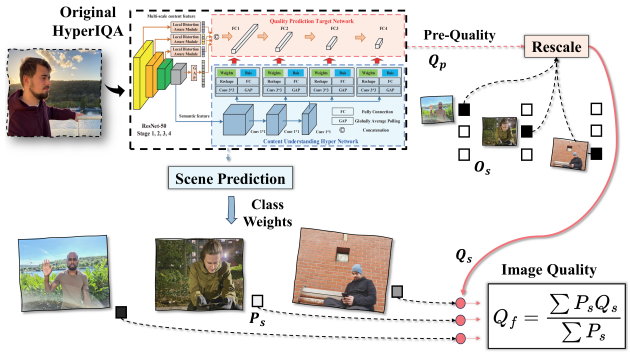

Fig. 1. Diagram of FULL-HyperIQA (FHIQA). The figure illustrates how FHIQA processes input images, extracts semantic information, and adapts the quality prediction based on scene-specific evaluations.

Blind IQA (BIQA) has gained significant interest in recent years, as it offers universal metrics for quality assessment without the need for pristine reference images, often scarce in practical photography scenarios. BIQA methods, based on convolutional neural networks (CNN) [8, 10, 25], have shown marked improvements over their classical counterparts, due to their ability to extract perceptual quality information [5, 13, 14, 18, 22]. However, many existing BIQA methods exhibit limitations in capturing scene-specific semantics, often treating diverse scenes with a one-size-fits-all approach. Given that image quality is inherently subjective and varies under different conditions, numerous studies underscore the significance of integrating semantic information into the assessment of image quality, tackling a challenge known as domain shift [21, 24, 26, 27]. This is often achieved through a semantics-aware multitasking framework [1, 2, 3, 7, 20]. Moreover, the variability of quality scales across IQA datasets introduces ambiguities, complicating the aggregation of IQA insights. As a result, most BIQA methods struggle to generalize to new conditions, underscoring a critical challenge in cross-domain

This work was funded in part by the French government under the management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), the Louis Vuitton/ENS chair in artificial intelligence and the Inria/NYU collaboration. This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011013850). NC was supported in part by a DXOMARK/PRAIRIE CIFRE Fellowship. JVC was supported by Grant PID2021-128178OB-I00 funded by MCIN/AEI/10.13039/501100011033, ERDF "A way of making Europe", the Departament de Recerca i Universitats from Generalitat de Catalunya with reference 2021SGR01499.

N. Chahine is with DXOMARK, Paris, France & the Departement d'informatique de l'Ecole normale suprerieure (ENS-PSL, CNRS, Inria) (e-mail: nchahine@dxomark.com).

Ferradans is with DXOMARK, Paris, France (e-mail: sferradans@dxomark.com).

J. Vazquez Corral, is with Computer Vision Center, Barcelona, Spain & Computer Sciences Departament Universitat Autònoma de Barcelona, Barcelona, Spain (e-mail: javier.vazquez@cvc.uab.cat).

J. Ponce is with the Departement d'informatique de l'Ecole normale suprerieure (ENS-PSL, CNRS, Inria) & the Institute of Mathematical Sciences and Center for Data Science, New York University (e-mail: jean.ponce@ens.fr).

image quality evaluations.

In tackling the limitations of current BIQA methodologies, such as combining IQA knowledge and limitations on generalization, this work presents FULL-HyperIQA (FHIQA), an advancement over prior HyperNetwork models ([20, 1]). FHIQA not only considers scene-specific semantics in predicting image quality but also merges knowledge across various scenes for improved generalization. This approach offers a context-aware IQA method, better suited to adapting to new, unfamiliar conditions.

Our contribution is a new BIQA model, FHIQA, focusing on, a) precision in quality metrics, meaning, the capacity to generate highly granular results thus enabling the differentiation of closely matched cameras; and b) generality by extending the scope of previous work to tackle image content (scenes) not encountered during training. Our comprehensive tests on a newly introduced PIQ23 test split demonstrate FHIQA's enhanced generalization capabilities over existing BIQA benchmarks.

II. Method

We present FULL-HyperIQA (Fig. 1), an enhanced version of SEM-HyperIQA [1], with a specific focus on adapting to scenes not encountered during training.

a) Scene-specific rescaling: SEM-HyperIQA combines the HyperIQA architecture [20], which integrates semantic information, with multitasking, which allows scene-specific quality score rescaling. Semantic features from multiple random crops are concatenated and fed to a multi-layer perceptron (MLP) that predicts the scene category for the image (s). Then, the predicted category is passed to a smaller MLP that predicts a multiplier a_i^s and offset b_i^s to adapt the predicted quality score of each patch, q_i, to its respective scene quality scale, such as $\hat{q}_i^s = a_i^s q_i + b_i^s$, where $\hat{q}_i^s$ is the rescaled quality score of patch i. The final image quality score is computed by averaging the individual patch scores. This approach provides a good basis for solving the problem of domain shift and scene-specific rescaling. However, SEM-HyperIQA does not explicitly allow generalization to new scenes, since it relies on predicting a single category for each image, supposing we only use images from known scenes.

b) Understanding new content: The novelty of FHIQA lies in its approach to quality score rescaling. Instead of adjusting the final score based solely on the scene predicted for the image, FHIQA utilizes the entire scene prediction vector. This vector can be interpreted as the set of coordinates in the vectorial space of scenes defined in the training set. The coordinate (or weight) of each scene indicates how similar the input is to that scene. For the scenes of the training set, the input image is assigned a one-hot vector corresponding to that scene (O_s in Fig. 1). During training, the model is trained to predict the scene in the image and to align the predicted pre-quality score ($Q_p \in \mathbb{R}$ in Fig. 1) with the scene scale, by passing the predicted classification vector as input to the rescaling layer. The only difference is that we split the predicted classification vector into separate one-hot encodings and then rescale the pre-quality score on each of the training scenes. The final quality score is obtained by weighting the rescaled scores with the predicted classification vector. Finally, our main concern is to generalize the quality predictions to scene categories that were not included in the training set. We hypothesize that the information needed for this is naturally encoded in the class prediction weights. In short, if the scene classification vector considers an input image to be closer to one scene category, its quality score should also be closer to that category. Mathematically, we can define our hypothesis as follows:

$$Q_f = \frac{\sum_{i=1}^{k} P_{s_i}(a_i^s Q_p + b_i^s)}{\sum_{j=1}^{k} P_{s_j}}, \text{ where} \tag{1}$$

- Q_f is the final quality score for the input image.
- P_{s_i} denotes the weight that the input image belongs to scene s_i.
- a_i^s and b_i^s are the multiplier and offset, respectively, for the scene s_i.
- Q_p is the pre-quality score predicted by the fully connected layer.

To optimize this approach, instead of considering all scenes, only the top k scenes are considered. The weights are then normalized, and a weighted average is taken based on these top k scenes to produce the final quality score for the input image.

III. Dataset

Since our work focuses on portrait quality assessment, we have chosen to evaluate our model on PIQ23 [1]. Portrait photography captures the essence of human emotions, expressions, and individual characteristics, presenting unique challenges in the domain of image quality assessment. Recognizing the need for a specialized benchmarking tool, PIQ23 [1] has been introduced as a dataset aimed at a more subtle evaluation of IQA models within the context of portrait photography. Benchmarking against PIQ23 ensures a comprehensive and representative evaluation of real-world scenarios for smartphone portrait photography. A significant challenge in IQA is the model's ability to generalize to multiple scenes, especially those not encountered during training. To rigorously evaluate this aspect, we have carefully chosen 15 out of 50 scenes featured in PIQ23 for testing, and the rest for training. This selection accounts for approximately 30% of the total images, uniformly distributed across the different lighting conditions, encompassing around 1486 out of the 5116 images of PIQ23. During the scene selection process, we ensured that both sets captured a rich blend of conditions, i.e., framing, lighting, and skin tones (Fig. 2).

IV. Experiments

A. Baselines methods

We have compared FHIQA with several well-known BIQA models: DB-CNN [25], HyperIQA [20], MUSIQ [9], and SEM-HyperIQA [1]. Using their official implementations, we have fine-tuned these models on the PIQ23 dataset. DB-CNN and two MUSIQ models were initially trained on the LIVE

Fig. 2. Examples from the new scene split for PIQ23. The test set incorporates various framing settings, backgrounds, subject characteristics, and weather conditions that are significantly distinct from the training set.

Challenge, KonIQ-10k, and PaQ-2-PiQ datasets, respectively. For all HyperIQA variants, only the Resnet50 backbone was pre-trained on ImageNet without any subsequent IQA pre-training. Due to HyperIQA's input size constraint of 224x224, we had to modify its architecture to handle resolutions that are multiples of 224 to be able to train on larger images.

B. Training strategy

We test different training configurations for all the proposed methods and report the best results. Specifically, we randomly crop the images to square patches of one of the three following sizes: 224 (5 patches per image), 672 (3 patches), and 1344 (1 patch). We use Adam stochastic optimization with different learning rates between 10^{-6} and 10^{-4}. For HyperIQA, SEM-HyperIQA, and FHIQA, we adopt different learning rates per module. For instance, we apply a smaller learning rate for the Resnet50 backbone compared to the hypernetwork, the rescaling, and the classification blocks. We fix the training for 300 epochs and adopt a learning rate decay factor of 0.05 for every 10 epochs. For FHIQA, we experimented with different values of k (the number of scenes utilized), including 3, 5, 10, and 25. We use early stopping with a patience of 40 epochs. Finally, we use Huber loss for the quality output and cross-entropy loss for the classification output of multitasking models. For the latter models, we apply a weighted sum of losses with a weight of 0.5 for the classification loss and 1.0 for the quality loss.

C. Metrics

To evaluate the performance, we compute Pearson's linear correlation coefficient (PLCC), Spearman's rank correlation coefficient (SRCC), Kendal's rank correlation (KRCC), the averaged correlations, and the mean absolute error (MAE) between the model outputs and the ground-truth scores. In the PIQ23 dataset, each scene is annotated individually, thus quality scores cannot be merged. Therefore, we calculate metrics for each scene separately. The aggregate performance across all scenes for a given metric is determined by the median, $M_{\text{Med}} = M_{(\frac{s}{2})}$ where s denotes the total number of scenes, and $M_{(i)}$ represents the i-th smallest scene metric value among the sorted scenes. For the early stopping, we evaluate our models on their SRCC performance. It is noteworthy that achieving a high SRCC doesn't necessarily correlate with other metrics. For future works, we might indeed benefit from considering a mix of metrics for early stopping, as small variations in correlation might arise due to a variety of factors, such as the number of images for each scene, score distribution, etc.

D. Results

The results presented in Tab. I and Fig. 3 offer several significant insights. A deep look into the performance metrics reveals that FHIQA is consistently competitive across the different attributes. It stands out and outperforms other models for "Overall", demonstrating its comprehensive assessment capabilities and semantic understanding. Specifically, for the "Overall" attribute, we utilize the entire image, in contrast to other attributes where the evaluation is confined solely to the face region. Therefore, where semantic understanding matters the most, FHIQA proves that it can adapt the knowledge from the training set to generalize to new conditions. Furthermore, when looking at the three attributes altogether, our model is best in 7 out of 15 cases and second best in 6 others. On the one hand, where other HyperIQA variants perform consistently worse, FHIQA stands out as a better solution for generalization, even when compared to MUSIQ, which is pre-trained on IQA datasets. On the other hand, MUSIQ variants also display compelling performances. The PaQ-2-PiQ pre-trained variant dominates in "Exposure". In contrast, DB-CNN doesn't fare as well as the other models. In summary, while different models exhibit strengths in specific areas, FHIQA, with no prior IQA pre-training, stands out, emphasizing its generalization and contextual assessment capabilities across a diverse range of attributes and conditions.

E. Understanding new content

Fig. 4 illustrates the distribution of FHIQA's predicted classes for unfamiliar conditions. These distributions highlight

TABLE I
PERFORMANCE METRICS OF VARIOUS IQA MODELS ON PIQ23. THE RESULTS ARE PRESENTED AS THE MEDIAN ACROSS SCENES FOR EACH METRIC.

Model\Attribute	Overall				Exposure				Details			
	SRCC	PLCC	KRCC	MAE	SRCC	PLCC	KRCC	MAE	SRCC	PLCC	KRCC	MAE
DB-CNN (LIVE C)	0.59	0.64	0.43	1.04	0.69	0.69	0.51	0.91	0.59	0.51	0.45	0.99
MUSIQ (KonIQ-10k)	0.76	0.75	0.57	**0.95**	0.74	0.70	0.55	0.93	0.71	0.67	0.52	0.88
MUSIQ (PaQ-2-PiQ)	0.74	0.74	0.54	1.09	**0.79**	**0.78**	**0.59**	0.87	0.72	**0.77**	0.53	0.90
HyperIQA*	0.74	0.74	0.55	0.99	0.69	0.68	0.50	0.86	0.70	0.67	0.50	0.94
SEM-HyperIQA*	0.75	0.75	0.56	1.03	0.72	0.70	0.53	0.97	0.73	0.65	**0.55**	0.88
SEM-HyperIQA-CO*	0.74	0.74	0.55	1.04	0.70	0.70	0.52	0.94	**0.75**	0.71	**0.55**	0.85
FULL-HyperIQA*	**0.78**	**0.78**	**0.59**	1.12	0.76	0.71	0.57	**0.85**	0.74	0.72	**0.55**	**0.80**

*ImageNet, backbone only; **best**; second best.

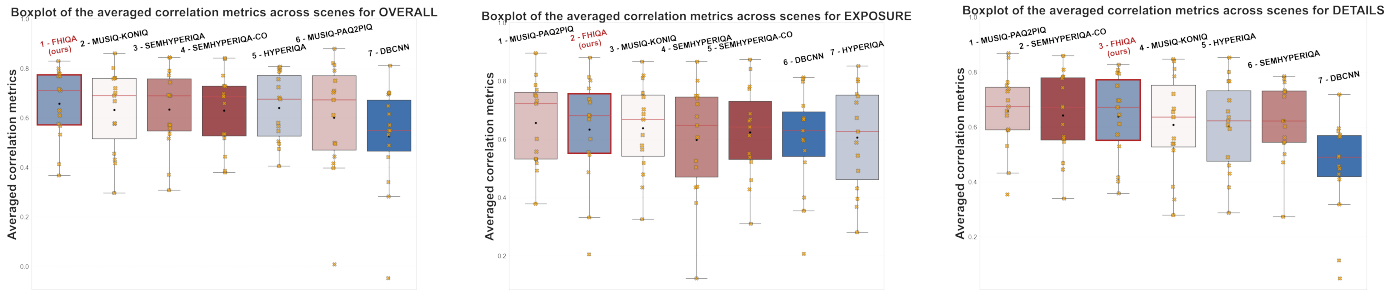

Fig. 3. Comparative analysis of IQA models based on the averaged correlation metrics distribution across all scenes and for the three attributes of PIQ23.

Fig. 4. Histograms showing the classification distribution across training scenes for various unseen testing conditions. The same testing scene can be projected to multiple training scenes with similar features, showcasing the necessity to consider multiple scenes for inference on new conditions.

the significance of semantic discovery when generalizing an IQA measure for unseen content. For instance, scene 17 depicts a man at his computer desk in a lowlight setting. The model projects this scene onto related contexts like desk scenes, library settings, and other lowlight scenarios. FHIQA leverages these diverse classification distributions to derive a quality assessment that integrates various content perspectives, rather than restricting itself to a singular scene or content. This approach recognizes the pivotal role of scene semantics in determining image quality. The model's robust performance across all attributes —particularly its standout results on "Overall"— points to the need for future IQA models to offer content-specific evaluations that are both precise and nuanced.

V. CONCLUSION

In this paper, we have introduced Full-HyperIQA (FHIQA), a novel BIQA method focused on scene generalization. We hypothesize that the information needed for generalization is naturally encoded in the class prediction weights and that the quality of unfamiliar conditions can be extracted based on similar conditions in the training set. Our model obtains competitive or top performance on all attributes of PIQ23, demonstrating the effectiveness of our approach. This paper underscores the pivotal role of semantic understanding in achieving effective generalization for image quality assessment, particularly in portrait scenarios. We advocate for the IQA community to move towards content-specific evaluations, especially in the field of smartphone photography, with the emergence of AI-driven image enhancement, which challenges the traditional IQA methodologies. FHIQA represents a step in this direction, setting the stage for future IQA models that can adapt to varied scenarios and quality criteria.

REFERENCES

[1] Nicolas Chahine, Stefania Calarasanu, Davide Garcia-Civiero, Theo Cayla, Sira Ferradans, and Jean Ponce. "An Image Quality Assessment Dataset for Portraits". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 9968–9978.

[2] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. "Perceptual quality assessment of smartphone photography". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3677–3686.

[3] Zhaopeng Feng, Keyang Zhang, Shuyue Jia, Baoliang Chen, and Shiqi Wang. "Learning from mixed datasets: A monotonic image quality assessment model". In: *Electronics Letters* 59.3 (2023), e12698.

[4] Deepti Ghadiyaram and Alan C Bovik. "Massive online crowdsourced study of subjective and objective picture quality". In: *IEEE Transactions on Image Processing* 25.1 (2015), pp. 372–387.

[5] Deepti Ghadiyaram and Alan C Bovik. "Perceptual quality prediction on authentically distorted images using a bag of features approach". In: *Journal of vision* 17.1 (2017), pp. 32–32.

[6] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4041–4056.

[7] Chen-Hsiu Huang and Ja-Ling Wu. "Multi-task deep CNN model for no-reference image quality assessment on smartphone camera photos". In: *arXiv preprint arXiv:2008.11961* (2020).

[8] Le Kang, Peng Ye, Yi Li, and David Doermann. "Convolutional neural networks for no-reference image quality assessment". In: *IEEE conference on computer vision and pattern recognition*. 2014, pp. 1733–1740.

[9] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. "Musiq: Multi-scale image quality transformer". In: *IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5148–5157.

[10] Jongyoo Kim and Sanghoon Lee. "Fully deep blind image quality predictor". In: *IEEE Journal of selected topics in signal processing* 11.1 (2016), pp. 206–220.

[11] Eric Cooper Larson and Damon Michael Chandler. "Most apparent distortion: full-reference image quality assessment and the role of strategy". In: *Journal of electronic imaging* 19.1 (2010), p. 011006.

[12] Rafał K Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. "Comparison of four subjective methods for image quality assessment". In: *Computer graphics forum*. Vol. 31. 8. Wiley Online Library. 2012, pp. 2478–2491.

[13] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. "No-reference image quality assessment in the spatial domain". In: *IEEE Transactions on image processing* 21.12 (2012), pp. 4695–4708.

[14] Anush Krishna Moorthy and Alan Conrad Bovik. "Blind image quality assessment: From natural scene statistics to perceptual quality". In: *IEEE transactions on Image Processing* 20.12 (2011), pp. 3350–3364.

[15] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk. "From pairwise comparisons and rating to a unified quality scale". In: *IEEE Transactions on Image Processing* 29 (2019), pp. 1139–1151.

[16] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. "Image database TID2013: Peculiarities, results and perspectives". In: *Signal processing: Image communication* 30 (2015), pp. 57–77.

[17] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. "TID2008-a database for evaluation of full-reference visual quality assessment metrics". In: *Advances of Modern Radioelectronics* 10.4 (2009), pp. 30–45.

[18] Michele A Saad, Alan C Bovik, and Christophe Charrier. "Blind image quality assessment: A natural scene statistics approach in the DCT domain". In: *IEEE transactions on Image Processing* 21.8 (2012), pp. 3339–3352.

[19] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. "A statistical evaluation of recent full reference image quality assessment algorithms". In: *IEEE Transactions on image processing* 15.11 (2006), pp. 3440–3451.

[20] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. "Blindly assess image quality in the wild guided by a self-adaptive hyper network". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3667–3676.

[21] Wei Sun, Xiongkuo Min, Guangtao Zhai, and Siwei Ma. "Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training". In: *arXiv preprint arXiv:2105.14550* (2021).

[22] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. "Unsupervised feature learning framework for no-reference image quality assessment". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 1098–1105.

[23] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3575–3585.

[24] Emin Zerman, Giuseppe Valenzise, and Frederic Dufaux. "An extensive performance evaluation of full-reference HDR image quality metrics". In: *Quality and User Experience* 2.1 (2017), pp. 1–16.

[25] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. "Blind image quality assessment using a deep bilinear convolutional neural network". In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.1 (2018), pp. 36–47.

[26] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. "Learning to blindly assess image quality in the laboratory and wild". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 111–115.

[27] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. "Uncertainty-aware blind image quality assessment in the laboratory and wild". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 3474–3486.