Self Patch Labeling Using Quality Distribution Estimation for CNN-Based 360-IQA Training

Abderrezzaq Sendjasni, Mohamed-Chaker Larabi

SELF PATCH LABELING USING QUALITY DISTRIBUTION ESTIMATION FOR CNN-BASED 360-IQA TRAINING

Abderrezzaq Sendjasni and Mohamed-Chaker Larabi

CNRS, Univ. Poitiers, XLIM, UMR 7252, France

ABSTRACT

In this study, we propose a methodology for estimating quality score distribution (QSD) for 360-IQA patch labeling. A collection of 2D-IQA models is used to generate a QSD for patches, inspired by how subjective quality ratings are gathered and handled. The proposed framework is first benchmarked on a subjectively annotated dataset, namely KonPatch-32k, in terms of patch quality classification. The best composition of QSD is then used to derive quality labels for patches sampled from 360-degree images. Furthermore, the quality labels are used in a multi-regression training strategy of CNN models. The ResNet-50 and EfficientNet-B5 are used to test the effectiveness of the proposed labeling framework on two publicly available 360-IQA datasets, namely OIQA and MVAQD. The experimental results demonstrated the efficacy of jointly using local and global qualities. The multi-regression proved to be a bit challenging on OIQA compared to MVAQD, reflecting the necessity to accurately regulate the training process.

***Index Terms*—** Convolutional neural networks, Adaptive Patch labeling, Image quality assessment, 360-degree.

1. INTRODUCTION

Quality of experience (QoE) evaluation is undoubtedly a difficult undertaking, especially for immersive media, such as 360-degree images. The nature and distinctive features of such a content, as well as the used display devices, head-mounted displays (HMDs), necessitate a deeper understanding in order to correctly evaluate QoE [1]. The latter is very often approached from the image quality assessment (IQA) perspective, which considers the quality of an image as the property of its entire content. The quality evaluation is primarily performed by human observers, where the severity of visual degradation is quantified according to a rating scale. Subjective studies allow for gathering such information and deriving the mean opinion scores (MOS) from collected distributions of observer responses. This MOS is considered as the ground truth about quality and is the target to reach when developing an approach to predict it.

With the introduction of ConvNets (CNNs) for 360-degree IQA, models are trained to associate the learned feature representations to the MOS. Because of the unavailability of MOS per individual portions, multichannel CNNs are designed. The latter comprise several CNNs in parallel to process a portion of the image because of its high resolution, but also to mimic the visualization process of the users. Such models are known to be significantly complex due to the multi CNNs [2]. In contrast, patch-based CNNs can be adopted. These models require a labeling step to label the input patches prior to training the model. In existing methods [3,4], patch quality is directly inherited from the corresponding image, inducing redundancy among quality labels. This may be problematic since local (patch content) perceptual quality is not always consistent with global (whole image) quality due to the non-stationarity of image content across spatial locations and the intricate interactions between content and distortions [5]. This is more emphasized for 360-degree images, where the images may include a high diversity of content within the 360 degrees. Moreover, some processing such as compression may create impairments that are inequally distributed over the sphere. Consequently, the use of MOS as labels for sampled patches is somewhat questionable. Finding a way to label the patches that copes with the above drawbacks is important to reach a reliable results.

As an attempt to answer the aforementioned problem, this paper presents a quality score distribution methodology in response to the demand for approaches to correctly label patches based on their local quality. The proposed framework, which is inspired by the way subjective human opinions are collected and processed, provides quality score distributions from which a quality label is derived. We use the rich state-of-the-art of 2D-IQA to achieve this. To estimate the quality score distribution of image patches, traditional and deep-learning models designed for 2D content are utilized as observers. A screening of the 2D-IQA models and a validation step are conducted on the KonPatch-32k [6] to formalize the proposed framework. The validation is performed according to the ability to binary classify a patch quality into low or high quality. The performance of the proposed framework to adequately label patches for 360-IQA CNN training is validated on two 360-IQA datasets, namely OIQA [7] and MVAQD [8]. For this purpose, we sampled patches from the spherical representation of 360-degree images based on visual scan-paths. By doing so, we account for the exploration trajectories of users. Then, two CNN models, including ResNet-50 [9] and EfficientNet-B5 [10] are used to validate the efficiency of the proposed patch labeling framework. The models are trained to predict the assigned quality of the patch as well as the quality of the corresponding 360-degree image. As a result, the model is able to associate local quality with global quality through a multi-regression training scheme.

2. 360-PATCH QUALITY LABELING FRAMEWORK

With the aim to solve the redundancy problem among quality labels when using MOS for several patches, regardless of their relevance or amount of information, the proposed labeling framework describes local quality based on quality score distribution. It involves two primary steps, quality score distribution estimation and quality score labels generation. These two steps are described in the following.

2.1. 2D-IQA Models as Virtual Observers

We choose m 2D-IQA models to act as virtual observers based on how subjective experiments are conducted and opinion scores from

This work is funded by the Nouvelle Aquitaine regional Council under projects SIMOREVA360 2018-1R50112 and PERCEIVE.

observers are gathered. The selected models are summarized in Table 1. Patches extracted from 360-degree content are considered as a kind of 2D images. So, by exploiting the rich state-of-the-art on 2D-IQA, we are able to generate quality score distributions at patch-level for 360-IQA.

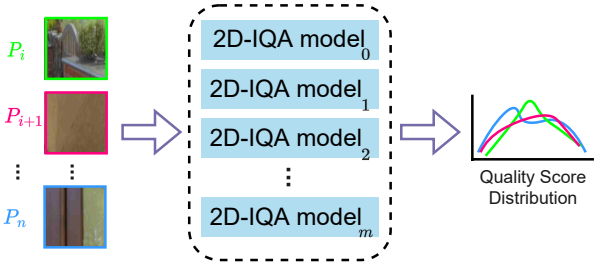

Fig. 1: Structure of the proposed patch labeling framework.

The structure of the proposed labeling framework is depicted in Fig. 1. It consists of applying m 2D-IQA models to each patch P_i with $i \in \{1, n\}$, generating a quality score distribution QSD composed of m scores. This step is formalized as follows:

$$QSD(P_i)_{i\in\{1,n\}} = \{\mathrm{M}_0(P_i), \mathrm{M}_1(P_i), ..., \mathrm{M}_m(P_i)\}, \quad (1)$$

where M_j represent the j^{th} 2D-IQA model. The computed distributions (QSD) are then validated in terms of ability to classify patches into low- or high-quality. Each score composing QSD is considered as a feature, focusing and describing distinct attributes within the input patch. For instance, DB-CNN assesses jointly authentic and synthetic distortions, while CPBD focuses on quantifying the blur. Hence, the quality scores associated with P_i composing QSD_i are used as features to a logistic regression (LR) model. Here, the task is to effectively classify the quality class of P_i starting from $QSD_i \in \mathbb{R}^m$. The set of features composing QSD are used under three different configurations:

ALL: refers to using all m features obtained from 2D-IQA models.

Traditional: refers to features obtained from traditional models.

DL: refers to features associated to deep-learning-based models.

Additionally, as every 2D-IQA model has a distinct quality scale, a normalization step is required prior to applying the LP model. The $MAX(.)$ scaling method is applied by dividing the set of features of M_j by the respective maximum value. By doing so, the extent of the values obtained by the M_j model is retained with respect to the mean.

Table 1: 2D-IQA models for estimation of the quality score distributions at patch level.

Type	Model
Traditional	NIQE [11], PIQUE [12], ILNIQE [13], BRISQUE [14], BIQAA [15], BLIINDS [16], CPBD [17], FADE [18]
Deep learning	DB-CNN [5], MUSIQ [19], Kang-CNN [20], NIMA [21], PaQ-2-PiQ [22], DipIQ [23]

The objective of the above-described operations is to identify the composition of QSD that accurately classifies patch quality. The configuration with the best performance is then used to derive quality labels. To select the most descriptive features, we use the well-known principal component analysis (PCA). It provides insights about which set of features contributes the most to the computation of the resulting component. This allows to screen the 2D-IQA models with the intention to achieve the most accurate data representation. Therefore, by applying $PCA(QSD(P_i))$, a new distribution, called $QSD'_i \in \mathbb{R}^k$, is generated and composed of k scores. QSD' serves to derive the quality label QL_i for P_i. At this stage, a simple arithmetic mean can be applied. However, the arithmetic mean is highly affected by extreme values. Therefore, we adopt the Minkowski mean instead. The latter has been widely used in IQA for pooling purposes. For each P_i, QL can be obtained as follows:

$$QL = (\frac{1}{k} \sum_{j \in QSD'} S_j^p)^{\frac{1}{P}}, \quad (2)$$

where S_j represents the j^{th} quality score obtained by the 2D-IQA model retained in QSD'. P corresponds to the Minkowski parameter. The latter is defined empirically as the $1/8$.

2.2. 360-Patch Quality Labeling for CNN Training

Let us remind that the proposed labeling framework is used to label patches from 360-degree images. It assigns a quality label to each P_i, representing its local quality. This step is used in a patch-based CNN framework for 360-IQA as a pre-processing step prior to training. First, patches are selected using the visual scan-path model ScanGAN360 [24]. The predicted exploration trajectories, twelve in total, are composed of thirty gaze fixations each, representing gaze points. These points correspond to salient regions of the scenes. This way, we account for (a) the way the human observer explored scenes, and (b) the variability of the exploration among observers as several scan-paths are considered. Therefore, 12×30 fixations are generated for each 360-degree image I. These fixations are taken as the center of 128×128-pixel patches, as illustrated in Fig. 2.

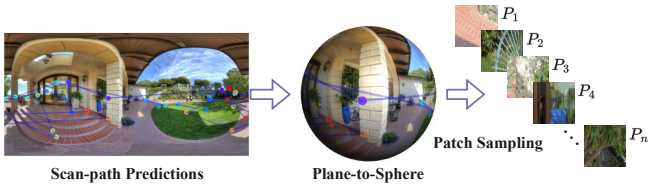

Fig. 2: Visual scan-path based patch sampling on the sphere.

The use of MOS (global quality) to label patches as done in the literature assumes equal contribution of patch qualities (local quality) to the MOS. This assumption does not leave any considirations of non-uniformity among patches. To avoid such an inadequacy, the proposed framework assigns a distinct label to each patch P_i by taking into account its objective quality QL_i and MOS_I.

To train a CNN model to predict both QL and MOS, a multi-regression training strategy is adopted. By doing so, the model achieves a global-local quality combination by regressing the learned representation to pQL and $pMOS$. Feature extraction, encoding, and representation are performed by a CNN. Here, we use ResNet-50 [9] and EfficientNet-B5 [10] pre-trained models. The choice of these models is motivated by their success and popularity with transfer learning in various image processing tasks. The output of the CNN model is fed to a regression block composed of a global average pooling (GAP), a fully connected layer (FC) with 512 neurons

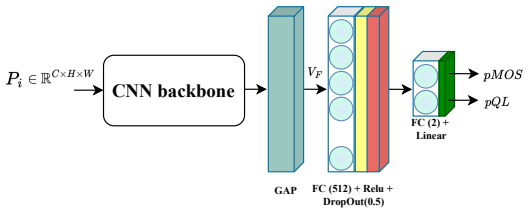

Fig. 3: Adopted training strategy. $V_{\mathbf{F}} \in \mathbb{R}^{C\times 1\times 1}$ represents the feature vector learned by the CNN backbone.

followed by a ReLU, and a dropout layer for regularization, and a final FC layer with two units to deliver $pMOS$ and Sp_i, representing the global/local quality respectively. The predicted scores associated to the same 360-degree image are averaged in order to compute the final objective quality scores. The architecture of training strategy is depicted in Fig. 3. The appended regression block is trained to minimize the following loss equation:

$$Loss = \alpha L_{MOS} + (1-\alpha)L_{QL} \tag{3}$$

where L corresponds to the Huber loss [25] and $\alpha \in \mathbb{R}^+$ is a parameter balancing both losses. It was fixed to 0.5 in this study.

3. RESULTS AND DISCUSSION

3.1. Experimental Setup:

Datasets: The KonPatch-32K is used to benchmark the proposed labeling framework, and to effectively derive the QLs. This dataset contains 32k patches sampled from other datasets and assessed subjectively according to whether a patch is sampled from a low- or high-quality image. Therefore, binary classes are provided with the dataset, allowing us to train the Logistic Regression (LR) model. In total, the dataset comprises, 22432 and 9062 patches with 0 (low-quality) and 1 (high-quality) labels, respectively. Prior to performing the experiments, the classes are balanced to avoid altering the accuracy of the classification model.

Table 2: Summary of the used 360-degree image datasets.

dataset	OIQA	MVAQD
Ref images	16	15
Distorted images	320	300
Distortion type (Distortion levels)	JPEG (5) / JPEG2000 (5) / BLUR (5) / WGN (5)	BLUR (4) / HEVC (4) / JPEG (4) / JP2K (4) / WGN (4)

Regarding the training and evaluation of the ResNet-50 and EfficientNet-B5 models according to the training strategy explained in Sec. 2.2, we use the OIQA and MVAQD datasets. A summary of each one is provided in Table 2.

Implementation Details: The training of the CNN models is conducted on a server with Intel Xeon Silver 4208 2.1GHz, 192G RAM, and a GPU Nvidia Telsa V100S 32G. The batch size is set to 32 and the Adam optimizer is used with a learning rate $Lr = 0.001$. Five-fold cross-validation is used within each of the selected datasets. Each fold is trained for 100 epochs. We split the datasets into $60\%/20\%/20\%$ for training, validation, and testing, respectively. To ensure complete separation of the training and testing sets, the distorted samples associated with the same pristine image are allocated to the same set.

Evaluation Criteria: The evaluation of the proposed labeling framework during its design and validation is performed using the confusion matrix (MC) and the area under the curve (AUC) [26]. To evaluate the performance of the CNN models, three statistical metrics are used, including the Pearson linear correlation coefficient (PLCC) to evaluate accuracy, the Spearman rank-order correlation coefficient (SRCC) for monotonicity, and the root mean squared error (RMSE) to evaluate prediction errors.

3.2. Classification on KonPatch-32k

The proposed labeling framework is validated on the KonPatch-32k. This validation is carried out by evaluating the performance of binary classifying patches into "sampled from low-quality images" and "sampled from high-quality images". The results achieved by the three configurations **ALL**, **Traditional**, and **DL** are compared using CMs and AUC scores.

From the CMs plots in Fig. 4, one can observe a good performance is achieved when using all scores composing the QSD (**ALL** configuration). 85% and 87% correct classifications of class 0 and 1, respectively. This showcases the representativeness of the used quality score distributions QSD. For the **Traditional**, the classification performance dropped to $77\%/82\%$. Such a behavior may be due to the variance among the used traditional models. It is acknowledged that less variability among the used features leads to better performance when it comes to quality assessment. Whereas in our case, the selected traditional models may assess and describe different aspects, as mentioned in Sec. 2.1. Regarding the performances of **DL**, $83\%/85\%$ correct classification is obtained, outperforming the **Traditional** subgroup.

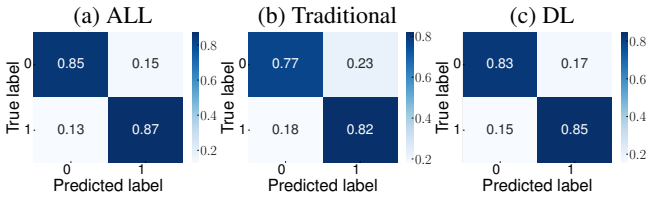

Fig. 4: Confusion matrix (CM) for both classes using QSD as features under three configurations: ALL, Traditional, and DL.

Table 3: AUC scores using QLs for the binary classification on KonPatch-32K. Best performances are highlighted in **bold**.

Configuration	Arithmetic mean	Five-number summary	Minkowski mean
All	0.528	0.524	0.566
Traditional	0.556	0.521	0.664
DL	**0.652**	**0.666**	**0.836**

As the purpose of the conducted classification experiment is to determine the composition of the QSD that effectively describes the local quality of each patch, the Minkowski mean is used to derive the QLs. The same analysis is performed to determine the efficacy of the derived QLs in describing patches quality. The LR model

is then trained to classify patches on KonPatch-32k using only the obtained QLs. The performances are compared with a simple arithmetic mean, and the five-number summary [27] methods. The latter makes use of information on (i) the location given by the median, (ii) the spread of the values given by the $Q1$ and $Q3$ quartiles representing the 25% and 75% percentiles respectively, and (iii) the range of values expressed by the minimum and maximum, making it effective in describing a QSD.

The AUC scores of each method with **ALL**, **Traditional**, and **DL** configurations are provided in Table 3. The first observation that emerges, is that the performances drastically dropped compared to using QSD as features for the classification. Using the arithmetic mean and the five-number summary, the classification performances become as random as a naive method, except for **DL** achieving $AUC = 0.652$ and $AUC = 0.666$, respectively. It appears that, by merging each QSD into a single value, the description of local quality losts its representativeness. This is not the case for the Minkowski mean. As it can be seen, an $AUC = 0.836$ is obtained with the **DL** configuration, supporting the previous observations. The other configurations are slightly improved as well. This behavior showcases the effectiveness of using the Minkowski mean to derive the QLs across the configurations over simply averaging or summarizing the $QSDs$.

Table 4: AUC scores using QSD' after applying the PCA of the QLs based on the Minkowski mean under **DL**. Best performances are highlighted in **bold**.

	All	Traditional	DL
AUC	0.605	0.748	**0.878**

The representativeness of QSD can be improved by the use of PCA as explained in Sec. 2.1. To validate this, we further analyze the classification performances of the Minkowski mean after applying the PCA on QSD, resulting in a new distribution QSD'. The performances are reported in Table 4, where AUC scores shifted from 0.566, 0.664, and 0.836 prior to using PCA to 0.606, 0.748, 0.878 after applying it. This actively showcases the utility of selecting and screening the 2D-IQA models to achieve a robust and representative quality score distribution for each patch. Another noteworthy observation is that QLs derived from diverse QSD do not benefit the quality classification, as demonstrated by the **ALL** configuration. This typically agrees with the way MOS is obtained through subjective experiments, where only scores that concur are considered to generate the MOS.

3.3. Quality Prediction on 360-IQA Datasets

With the intent to assess the effectiveness of the proposed quality labeling framework, two CNN models are trained to (i) predict the global quality $pMOS$ and (ii) predict the global/local quality $pMOS + pQL$. In addition, two sampling strategies are used, scan-path based sampling discussed in Sec. 2.2 and standard sampling. The latter is performed on the sphere to be consistent with 360-degree image characteristics. The performance results are gathered in Table 5.

According to the overall performances, an improvement is achieved when jointly using global and local qualities to train the CNN models. This can be seen regardless of the used sampling method for ResNet-50, but not with EfficientNet-B5. For the latter, the performances dropped on OIQA with standard sampling and on MVAQD with scan-path based sampling. Here, by using the assigned QLs along with MOS, the model behaves slightly worse than using only MOS. The multi-regression in this case did not improve the accuracy. Besides, the OIQA dataset appears to be more challenging compared to MVAQD, where on the latter, the improvements are higher compared to those on OIQA. An improvement within 1% is observed on OIQA, whereas 2% on MVAQD. As the model maps the learned representation to local and global qualities simultaneously, the regularization parameter (see Eq. 3) could be impacting the learning of the MOS. Regulating the computation of the overall loss should be in favor of the $pMOS$, and therefore, effectively learning the global quality while considering the local one to each patch. Regarding the sampling methods, using visual scan-paths appears to be less effective than the standard sampling. With the latter, all the content is considered, accounting for the whole scene. However, with scan-path-based sampling, a part of the content is neglected. This should not affect the performances, as relevant patches are selected. By considering this, the sampling method is not influencing the proposed labeling framework.

Table 5: Performances comparison among training modes ($pMOS$ for using only global quality and $pMOS$+pQL when considering the local quality labels). The best performances are highlighted in **bold** for each sampling and dataset.

	Sampling	Training mode	ResNet-50			EfficientNet-B05		
			PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
OIQA	Scan-path	$pMOS$	0.9622	**0.9504**	0.0384	0.9286	0.9055	0.0539
		$pMOS$+pQL	**0.9624**	0.9498	**0.0383**	**0.9335**	**0.9221**	**0.0501**
	Standard	$pMOS$	0.9632	0.9562	**0.0382**	**0.9308**	**0.9135**	**0.0519**
		$pMOS$+pQL	**0.9644**	**0.9582**	0.0384	0.9216	0.9101	0.0542
MVAQD	Scan-path	$pMOS$	0.9068	0.8969	0.0937	**0.8351**	**0.8135**	**0.1181**
		$pMOS$+pQL	**0.9150**	**0.9039**	**0.0876**	0.8251	0.8053	0.1213
	Standard	$pMOS$	0.9137	0.8824	0.0873	0.8582	0.8668	**0.1097**
		$pMOS$+pQL	**0.9148**	**0.8897**	**0.0867**	**0.8728**	**0.8672**	0.1100

4. CONCLUSIONS

This paper presented a patch quality labeling framework for 360-IQA CNN-based models utilizing a set of 2D-IQA models. Quality labels, which describe local attributes of individual patches, are derived from quality score distributions. The proposed methodology is evaluated on the KonPatch-32K dataset, where deep learning-based 2D-IQA models outperformed traditional methods in determining quality labels. The classification study performed on KonPath-32k helped in the validation of the local quality labels. The latter are then examined using a multi-regression strategy with ResNet-50 and EfficientNet-B5 and two patch sampling techniques. The performance on OIQA and MVAQD datasets demonstrated the advantages of categorizing patches based on their local attributes using the suggested framework, rather than using MOS to label patches. Nonetheless, training a patch-based CNN model to predict global quality by taking the local qualities of patches into account proved challenging. A precise regularization of the training process appears to be a solution for allocating appropriate attention to overall quality (MOS) during training. In the future, a more appropriate training strategy will be investigated by means of primary/auxiliary tasks training.

5. REFERENCES

[1] A. Perkis, C. Timmerer, S. Baraković, J. Husić, and *et al.*, "Qualinet white paper on definitions of immersive media experience (imex)," *ENQEMSS, 14th QUALINET meeting (online). Online: https://arxiv.org/abs/2007.07032*, 2020.

[2] A. Sendjasni, MC. Larabi, and FA. Cheikh, "Convolutional neural networks for omnidirectional image quality assessment: A benchmark," *IEEE TCSVT*, vol. 32, no. 11, pp. 7301–7316, 2022.

[3] TQ. Truong, TH. Tran, and TC. Thang, "Non-reference quality assessment model using deep learning for omnidirectional images," in *IEEE ICAST*, Morioka, Japan, 2019, pp. 1–5.

[4] L. Yang, M. Xu, X. Deng, and B. Feng, "Spatial attention-based non-reference perceptual quality prediction network for omnidirectional images," in *IEEE ICME*, Shenzhen, China, 2021, pp. 1–6.

[5] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE TCSVT*, vol. 30, no. 1, pp. 36–47, 2020.

[6] O. Wiedemann, V. Hosu, H. Lin, and D. Saupe, "Disregarding the big picture: Towards local image quality assessment," in *10^{th} QoMEX*, Cagliari, Italy, 2018, pp. 1–6.

[7] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual Quality Assessment of Omnidirectional Images," in *IEEE ISCAS*, Florence, Italy, 2018, pp. 1–5.

[8] X. Zheng, G. Jiang, M. Yu, and H. Jiang, "Segmented spherical projection-based blind omnidirectional image quality assessment," *IEEE Access*, vol. 8, pp. 31647–31659, 2020.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.

[10] T. Mingxing and L. Quoc, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[11] A. Mittal, R. Soundararajan, and AC. Bovik, "Making a "completely blind" image quality analyzer," *IEEE SPL*, vol. 20, no. 3, pp. 209–212, 2013.

[12] N. Venkatanath and D. Praneeth *et al.*, "Blind image quality evaluation using perception based features," in *IEEE NCC*, 2015, pp. 1–6.

[13] L. Zhang, L. Zhang, and AC. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE TIP*, vol. 24, no. 8, pp. 2579–2591, 2015.

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE TIP*, vol. 21, no. 12, pp. 4695–4708, 2012.

[15] G. Salvador and C. Gabriel, "Blind image quality assessment through anisotropy," *J. Opt. Soc. Am. A*, vol. 24, no. 12, pp. B42–B51, 2007.

[16] MA. Saad, AC. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE TIP*, vol. 21, no. 8, pp. 3339–3352, 2012.

[17] N. Narvekar and L. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *IEEE QoMEX*, San Diego, USA, 2009, pp. 87–91.

[18] LK. Choi, J. You, and AC. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE TIP*, vol. 24, no. 11, pp. 3888–3901, 2015.

[19] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *IEEE ICCV*, Montreal, Canada, 2021, pp. 5128–5137.

[20] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1733–1740.

[21] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE TIP*, vol. 27, no. 8, pp. 3998–4011, 2018.

[22] Z. Ying and H. Niu *et al.*, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *IEEE CVPR*, Seattle, USA, 2020, pp. 3572–3582.

[23] k. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "DdipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE TIP*, vol. 26, no. 8, pp. 3951–3964, 2017.

[24] D. Martin, A. Serrano, A. Bergman, G. Wetzstein, and B. Masia, "Scangan360: A generative model of realistic scanpaths for 360° images," *IEEE TVCG*, vol. 28, no. 5, pp. 2003–2013, 2022.

[25] PJ. Huber, *Robust statistics*, John Wiley & Sons, 2004.

[26] JA. Hanley and BJ. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[27] C. Zewdie, M. Pedersen, and Z. Wang, "A new pooling strategy for image quality metrics: Five number summary," in *5th EUVIP*, Paris, France, 2014, pp. 1–6.