BLIND OMNIDIRECTIONAL IMAGE QUALITY ASSESSMENT: INTEGRATING LOCAL STATISTICS AND GLOBAL SEMANTICS

Wei Zhou and Zhou Wang

Department of Electrical & Computer Engineering, University of Waterloo, Canada
Email: {wei.zhou, zhou.wang}@uwaterloo.ca

ABSTRACT

Omnidirectional image quality assessment (OIQA) aims to predict the perceptual quality of omnidirectional images that cover the whole $180\times360°$ viewing range of the visual environment. Here we propose a blind/no-reference OIQA method named Local Statistics and Global Semantics metric (LSGS) that bridges the gap between low-level statistics and high-level semantics of omnidirectional images. Specifically, statistic and semantic features are extracted in separate paths from multiple local viewports and the hallucinated global omnidirectional image, respectively. A quality regression along with a weighting process is then followed that maps the extracted quality-aware features to a perceptual quality prediction. Experimental results demonstrate that the proposed LSGS method offers highly competitive performance against state-of-the-art methods.

Index Terms— Omnidirectional image, blind image quality assessment, low-level statistics, high-level semantics

1. INTRODUCTION

The rapid recent advancement in virtual reality (VR) technologies makes it possible to create immersive multimedia quality-of-experience (QoE) for end-users. As a representative form of VR, omnidirectional content has increasingly emerged in our daily life. To evaluate and optimize the perceptual QoE of omnidirectional content, objective omnidirectional image quality assessment (OIQA) models play a critical roles in the development of modern VR systems.

In the literature, objective OIQA models have emerged that follow both full-reference (FR) and no-reference (NR) frameworks. FR-OIQA models assume full access to information of the reference image and are usually direct extensions of traditional FR methods developed for regular rectangular 2D image quality assessment (IQA). For example, based upon the peak signal-to-noise ratio (PSNR), Yu et al. [1] propose the spherical PSNR (S-PSNR) algorithm, where PSNR is calculated for uniformly distributed points on a sphere instead of projected rectangular image. In [2], the weighted-to-spherically uniform PSNR (WS-PSNR) method is presented, where a weighting map is created by considering the stretched
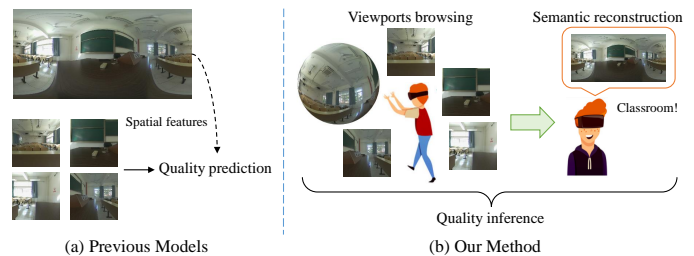
Fig. 1. Perceptual cues in omnidirectional image quality assessment. Existing models extract spatial information from various viewports and may obtain help from global projected maps, whereas the proposed method combines local image statistics and global semantic reconstruction.

degree. Zakharchenko et al. [3] propose the Craster parabolic projection PSNR (CPP-PSNR) approach, which maps the reference and distorted omnidirectional images on the Craster parabolic projection followed by PSNR computation.

NR-OIQA methods do not require access to the reference image and are more desirable in many application scenarios. Existing NR-OIQA approaches can generally be classified into two categories, depending on whether the conventional hand-crafted or learned deep features are employed for quality prediction. Multi-frequency information and local-global naturalness are applied to develop the MFILGN model [4]. More recent models employ deep convolutional neural networks (CNNs) or graph convolution networks (GCNs). These models demonstrate promising performance, including the multi-channel CNN for blind 360-degree image quality assessment (MC360IQA) [5], the viewport oriented graph convolution network (VGCN) [6], and its variant named adaptive hypergraph convolutional network (AHGCN) [7].

In a 360-degree viewing environment, e.g. using a head-mounted device, the observer is not able to visualize the whole omnidirectional content simultaneously, and thus an important step in the human subjective viewing experience is to establish or reconstruct a sense of the global semantics by browsing and integrating information from many viewports. During the course of image quality assessment, such global semantics are integrated with local observations on
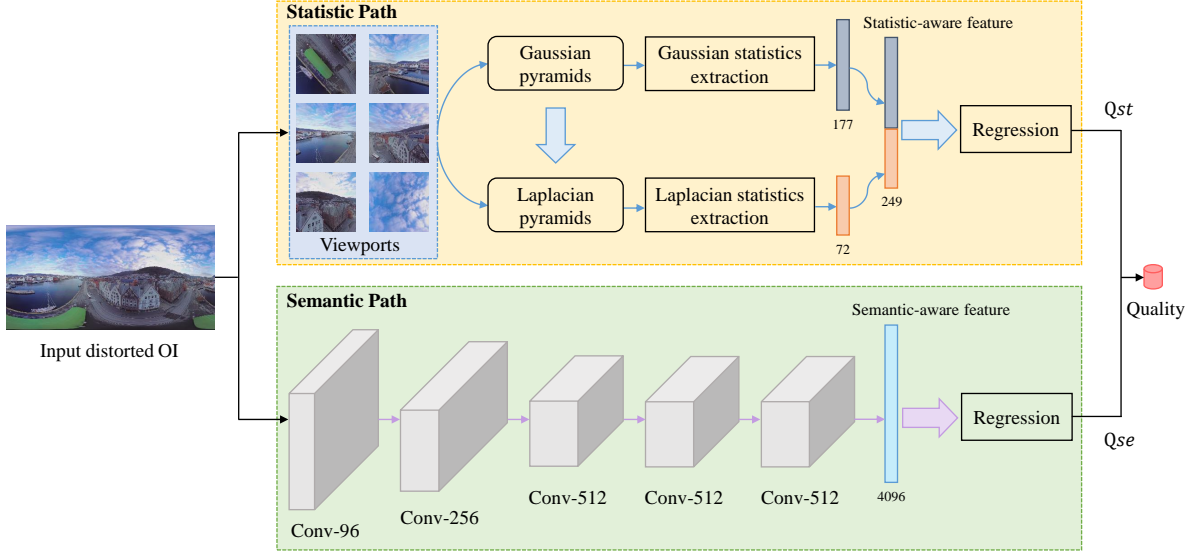
Fig. 2. Framework of the proposed LSGS method for blind OIQA.

image fidelity, naturalness, and/or artifacts to produce an overall quality evaluation. Motivated by this observation, we propose a statistic and semantic oriented quality prediction framework called LSGS for blind OIQA as illustrated in Fig. 1, by integrating features extracted from both low-level image statistics of multiple local viewports and high-level semantics of the hallucinated global omnidirectional image. A quality regression module is then leveraged to map the collection of the quality-sensitive features extracted from the two separate paths to an overall prediction of the subjective quality rating. Extensive experimental results demonstrate that the proposed method is superior to many state-of-the-art quality assessment models. In addition, we make some interesting observations on the relationship between semantic confidence and image distortions, as well as how the individual components affect the ultimate quality prediction performance in ablation studies.

2. PROPOSED METHOD

The overall framework of the proposed LSGS method is shown in Fig. 2, which consists of a statistic path, a semantic path, and a final quality regression step.

Since a variety of viewports are browsed by the viewers, we first convert the distorted omnidirectional image (OI) to multiple viewports. Given each input distorted OI denoted by D, we exploit the non-uniform viewport sampling strategy [8, 9] and obtain N viewports $V_n, n = 1, 2, ..., N$.

To capture the multi-scale characteristics of the human visual system [10], we construct pyramid representations [11, 12] of multiple local viewports. Specifically, multi-level Laplacian pyramids [13] are created by iterative Gaussian filtering, down-sampling, and subtracting, resulting in Gaussian

and Laplacian pyramids in the same process. For a specific viewport V_n, layers of the Gaussian pyramid are calculated as follows:

$$G_n^i(x,y) = \begin{cases} V_n, i = 1 \\ \sum_{u=-2}^{2} \sum_{v=-2}^{2} k(u,v) G_n^{i-1}(2x+u, 2y+v), i > 1 \end{cases},$$
(1)

where i is the layer index of the Gaussian pyramid, $x \in [0, X)$ and $y \in [0, Y)$ are the pixel position indices in which X and Y are the image dimensions, and $k(u, v)$ denotes the generating kernel that is typically defined by the coefficients of a low pass filter such as a 2D Gaussian filter.

We then interpolate each layer of the Gaussian pyramid by:

$$\hat{G}_n^i(x,y) = 4 \sum_{u=-2}^{2} \sum_{v=-2}^{2} k(u,v) G_n^i \left(\frac{u+x}{2}, \frac{v+y}{2} \right).$$
(2)

The residual between the current layer of the Gaussian pyramid and the interpolation result from the next layer defines the current layer of the Laplacian pyramid:

$$L_n^i = G_n^i - \hat{G}_n^{i+1}.$$
(3)

Since the computation of the i-th layer in the Laplacian pyramid requires the $(i + 1)$-th layer of the Gaussian pyramid, the number of layers in the Laplacian pyramid is one less than that in the Gaussian pyramid.

To extract features from the Gaussian pyramid, we compute the default uniform local binary pattern (LBP) descriptors, resulting in 59 statistics for each Gaussian layer. When a 3-layer Gaussian pyramid is employed, this leads to 177 Gaussian pyramid features denoted by f_{GP}. For a Laplacian pyramid, motivated by the success of natural scene statistics

(NSS) in IQA research [14, 15, 16], we extract mean subtracted and contrast normalized coefficients, leading to 36 features for each layer. When a 2-layer Laplacian pyramid is employed, this results in 72 Laplacian pyramid features denoted by f_{LP}. The full statistic feature set f_{st}, one for each viewport, is obtained by concatenating the statistical features extracted from the Gaussian and Laplacian pyramids as:

$$f_{st} = [f_{GP}, f_{LP}]. \tag{4}$$

We employ the VGGNet trained on the large ImageNet dataset [17] as the semantic feature extraction backbone, mainly for its simplicity and ability to capture image distortion-related representations [18]. In [19], three different structures of VGGNet have been proposed to balance between complexity and accuracy, namely fast VGG (VGG-F), medium VGG (VGG-M), and slow VGG (VGG-S). Each of them contains 5 convolutional (Conv) layers and 3 fully connected (FC) layers. The first two FC layers have 4,096 neurons, while the last one has 1,000 nodes indicating the 1,000 classes for image recognition. In our current implementation, we select the deep features from the first FC layer of VGG-M as our semantic feature set f_{se}:

$$f_{se} = FC_1(D). \tag{5}$$

To learn the mapping from features to quality labels, we feed the statistic features and semantic features separately to support vector regression (SVR) models [20], and denote the regressed statistic and semantic quality scores as Q_{st} and Q_{se}, respectively. The overall quality score is calculated by a weighted average:

$$Q_{\text{overall}} = wQ_{st} + (1-w)Q_{\text{se}}, \tag{6}$$

where w is a weighting factor that determines the relative importance of the statistic and semantic feature predictors.

3. VALIDATION

3.1. Experimental Setup and Performance Comparison

We evaluate the proposed approach on the CVIQD subjective database [21], which is so far a relatively large and widely adopted database containing both omnidirectional images and their corresponding quality labels given by human subjects. It consists of 16 original images and 528 distorted images produced by three classic image or video coding technologies, namely JPEG, AVC, and HEVC. The subjective quality ratings in the form of mean opinion score (MOS) are rescaled to the range of [0, 100], for which a higher MOS represents better perceptual image quality.

To compare the performance of various IQA models, we take Spearman Rank-Order Correlation Coefficient (SROCC), Pearson Linear Correlation Coefficient (PLCC) and Root Mean Squared Error (RMSE) as the evaluation criteria. Before calculating the PLCC and RMSE, a 5-parameter logistic

Table 1. Performance comparisons of objective models.

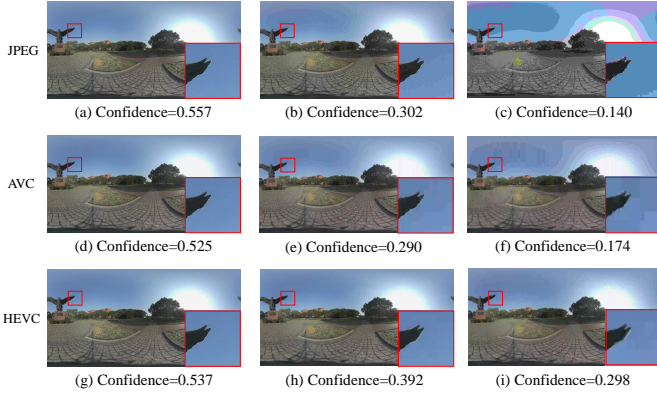| Types	Methods	SROCC	PLCC	RMSE
FR-IQA	PSNR	0.6239	0.7008	9.9599
	SSIM [23]	0.8842	0.9002	6.0793
	MS-SSIM [10]	0.8222	0.8521	7.3072
	FSIM [24]	0.9152	0.9340	4.9864
	DeepQA [25]	0.9292	0.9375	4.8574
FR-OIQA	S-PSNR [1]	0.6449	0.7083	9.8564
	WS-PSNR [2]	0.6107	0.6729	10.3283
	CPP-PSNR [3]	0.6265	0.6871	10.1448
NR-IQA	BRISQUE [26]	0.8180	0.8376	7.6271
	BMPRI [27]	0.7470	0.7919	8.5258
	DB-CNN [28]	0.9308	0.9356	4.9311
NR-OIQA	MFILGN [4]	0.9670	0.9751	3.1036
	MC360IQA [5]	0.9428	0.9429	4.6506
	VGCN [6]	0.9639	0.9651	3.6573
	AHGCN [7]	0.9623	0.9643	3.6990
	Proposed LSGS	**0.9710**	**0.9781**	**2.8945**

nonlinear fitting approach [22] is implemented to map the predicted quality into the subjective quality space.

The database is randomly divided into 80% data for training and the remaining 20% data for cross-validation. In order to relieve the uncertainty in training/testing splitting, we repeat this random-splitting and cross-validation process 100 times and report the median performance.

The performance of the proposed algorithm is compared against state-of-the-art quality assessment models, including five FR-IQA, three FR-OIQA, three NR-IQA, and four NR-OIQA methods. The results are shown in TABLE 1, where we observe that for FR-IQA metrics, the PSNR-based models are inferior to more advanced approaches such as structural (SSIM, MS-SSIM, FSIM) and deep learning (DeepQA) models. Somewhat surprisingly, the FR-OIQA methods do not help further improve upon FR-IQA approaches. By contrast, the NR-OIQA models show significant superiority over NR-IQA methods. This is likely due to their specific design to capture the characteristics of omnidirectional images. Among all metrics tested, the proposed LSGS method demonstrates highly competitive performance.

3.2. Semantic Confidence Versus Image Distortion

Since the proposed method contains a semantic path, it is interesting to observe the relationship between semantic confidence and image distortion. An example of distorted omnidirectional images with different JPEG, AVC and HEVC distortion levels is shown in Fig. 3, where from the first column to the third column, we observe that as the degree of distortion increases, the semantic confidence level decreases. This suggests that semantic information may be highly related to perceptual image quality. It is also interesting to see that semantic confidence shows various sensitivities to differ-

(a) Confidence=0.557 (b) Confidence=0.302 (c) Confidence=0.140

(d) Confidence=0.525 (e) Confidence=0.290 (f) Confidence=0.174

(g) Confidence=0.537 (h) Confidence=0.392 (i) Confidence=0.298

Fig. 3. The relationship between semantic confidence and image distortions. The first, second and third rows correspond to three distortion types (JPEG, AVC and HEVC compression, respectively). The first, second and third columns correspond to increasing distortion levels (low, medium and high, respectively) for each distortion type.
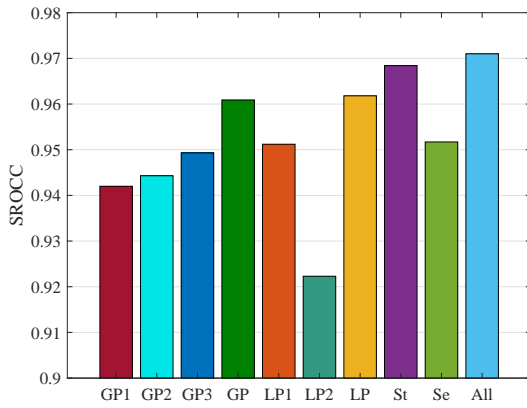
Fig. 4. Performance results of ablation experiments.

ent distortion types. In particular, the drop in semantic confidence levels is much less in more advanced image/video coding method HEVC than in earlier JPEG and AVC encoders.

3.3. Ablation and Parameter Sensitivity Tests

We evaluate the contributions from the statistic and semantic paths by ablation experiments, and the results are shown in Fig. 4, where GP1, GP2 and GP3, respectively, represent the cases of using the first, second, and third layers of Gaussian pyramid statistics only. GP denotes the case of using three layers of Gaussian pyramid statistics. We find that the performance increases gradually. Similarly, LP1 and LP2, respectively, correspond to the cases of using the first and second layers of Laplacian pyramid statistics only, while LP denotes the case of using 2-layer Laplacian pyramid statistics. The results show that LP produces the best performance among the three. The cases of adopting the statistic path and the se-

Table 2. Performance comparisons for different viewport numbers in the statistic path.

Numbers	SROCC	PLCC	RMSE
6	0.9684	0.9769	3.0083
20	0.9686	0.9777	2.9626
80	0.9683	0.9771	2.9501

Table 3. Performance comparisons for different neural network architectures in the semantic path.

Architectures	SROCC	PLCC	RMSE
VGG-F	0.9497	0.9537	4.2107
VGG-M	**0.9517**	**0.9576**	**4.0329**
VGG-S	0.9451	0.9486	4.4345

mantic path only are denoted by St and Se, respectively. It is observed that either path alone can achieve promising quality prediction performance, but adopting both paths (i.e. the All case) delivers the best performance. Relative speaking, the more dominant factor seems to be the statistic path. This may not be surprising as the statistic features come from different viewports directly visualized by human subjects while the global semantics offer complementary information for additional cues in quality assessment.

Because different parameter settings may be employed in the implementations of the proposed framework, here we test the sensitivity of our model with regard to various viewport numbers and semantic architectures. The results are reported in TABLE 2 and TABLE 3, respectively. We can see that the proposed model is insensitive to the viewport number. This allows us to reduce the number of viewports (for example, 6) to alleviate the computational complexity in real-world applications. The results also show that VGG-M outperforms the other neural network architectures in the semantic path. The possible reason may be that VGG-M achieves a preferable tradeoff between algorithm complexity and accuracy, making it a desired option for deep semantic backbone.

4. CONCLUSION

We propose a novel LSGS framework for blind omnidirectional image quality assessment that integrates both local low-level statistic and global high-level semantic features. Extensive experiments show that the proposed method achieves state-of-the-art performance. Observations on the relationship between semantic confidence and image distortion, and the ablation/sensitivity tests offer additional useful insights. Under the same framework, more advanced models for statistic and semantic analysis may be employed in the future, aiming for more accurate QoE assessment models that may help drive the advancement of immersive multimedia systems.

5. REFERENCES

[1] Matt Yu, Haricharan Lakshman, and Bernd Girod, "A framework to evaluate omnidirectional video coding schemes," in *IEEE International Symposium on Mixed and Augmented Reality*, 2015, pp. 31–36.

[2] Yule Sun, Ang Lu, and Lu Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1408–1412, 2017.

[3] Vladyslav Zakharchenko, Kwang Pyo Choi, and Jeong Hoon Park, "Quality metric for spherical panoramic video," in *Optics and Photonics for Information Processing X*. International Society for Optics and Photonics, 2016, vol. 9970, p. 99700C.

[4] Wei Zhou, Jiahua Xu, Qiuping Jiang, and Zhibo Chen, "No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[5] Wei Sun, Xiongkuo Min, Guangtao Zhai, Ke Gu, Huiyu Duan, and Siwei Ma, "MC360IQA: A multi-channel cnn for blind 360-degree image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 64–77, 2019.

[6] Jiahua Xu, Wei Zhou, and Zhibo Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2020.

[7] Jun Fu, Chen Hou, Wei Zhou, Jiahua Xu, and Zhibo Chen, "Adaptive hypergraph convolutional network for no-reference 360-degree image quality assessment," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 961–969.

[8] Jiahua Xu, Ziyuan Luo, Wei Zhou, Wenyuan Zhang, and Zhibo Chen, "Quality assessment of stereoscopic 360-degree images from multi-viewports," in *IEEE Picture Coding Symposium*, 2019, pp. 1–5.

[9] Zhibo Chen, Jiahua Xu, Chaoyi Lin, and Wei Zhou, "Stereoscopic omnidirectional image quality assessment based on predictive coding theory," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 103–117, 2020.

[10] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multi-scale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. IEEE, 2003, vol. 2, pp. 1398–1402.

[11] Zhou Wang and Eero P Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Human vision and electronic imaging X*. International Society for Optics and Photonics, 2005, vol. 5666, pp. 149–159.

[12] Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P Simoncelli, "Perceptual image quality assessment using a normalized laplacian pyramid," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–6, 2016.

[13] Peter J Burt and Edward H Adelson, "The laplacian pyramid as a compact image code," in *Readings in Computer Vision*, pp. 671–679. Elsevier, 1987.

[14] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[15] Yuming Fang, Kede Ma, Zhou Wang, Weisi Lin, Zhijun Fang, and Guangtao Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 838–842, 2014.

[16] Zhibo Chen, Wei Zhou, and Weiping Li, "Blind stereoscopic video quality assessment: From depth perception to overall experience," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 721–734, 2017.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[19] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[20] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[21] Wei Sun, Ke Gu, Siwei Ma, Wenhan Zhu, Ning Liu, and Guangtao Zhai, "A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison," in *IEEE 20th International Workshop on Multimedia Signal Processing*, 2018, pp. 1–6.

[22] Video Quality Experts Group et al., "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," *VQEG*, 2003.

[23] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[25] Jongyoo Kim and Sanghoon Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1676–1684.

[26] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[27] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.

[28] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.