Deep Ordinal Regression Framework for No-Reference Image Quality Assessment

Huasheng Wang, Yulin Tu, Xiaochang Liu, Hongchen Tan, and Hantao Liu

***Abstract*—Due to the rapid development of deep learning techniques, no-reference image quality assessment (NR-IQA) has achieved significant improvement. NR-IQA aims to predict a real-valued variable for image quality, using the image in question as the sole input. Existing deep learning-based NR-IQA models are formulated as a regression problem and trained by minimising the mean squared error. The error measurement does not consider the relative ordering between different ratings on the quality scale, which consequently affects the efficacy of the model. To account for this problem, we reformulate NR-IQA learning as an ordinal regression problem and propose a simple yet effective framework using deep convolutional neural networks (DCNN) and Transformers. NR-IQA learning is achieved by a deep ordinal loss and using a soft ordinal inference to transform the predicted probabilities to a continuous variable for image quality. Experimental results demonstrate the superiority of our proposed NR-IQA model based on deep ordinal regression. In addition, this framework can be easily extended with various DCNN architectures to build advanced IQA models.**

***Index Terms*—Convolutional neural networks, deep learning, image quality assessment, ordinal regression.**

I. Introduction

IN RECENT years, multimedia technologies have transformed our daily lives including the widespread use of digital cameras, smartphones, video surveillance systems, etc. These technologies produce large amounts of images, which typically exhibit different levels of perceived quality. It is critical to develop effective and reliable algorithms for image quality assessment (IQA) and use these algorithms to optimise image processing techniques, e.g., image retrieval [1], image denoising [2], and visual discomfort prediction [3]. According to the usage of the pristine/reference image, IQA models can be broadly classified into three genres: full-reference (FR) [4], reduced-reference (RR) [5], and no-reference (NR) [6], [7] models. Although FR-IQA and RR-IQA can achieve high performance, they are impractical since reference is often unavailable in many circumstances. Hence, NR-IQA that operates on distorted images directly has a great potential for real-world application scenarios.

The traditional NR-IQA models [8] utilise a two-stage framework including feature extraction and quality score regression. These models require prior knowledge of distortions to be able to extract relevant features; and their performance heavily depends on modelling of the natural scene statistics (NSS) [9] or the human visual system (HVS) properties [10]. To develop a universal approach for NR-IQA, recent research focuses on designing models using deep convolutional neural networks (DCNN) [11], [12]. Compared to the handcrafted feature-based approaches [13], [14], DCNN-based methods have a powerful capability to learn useful features for perceived image quality. These DCNN-based models have demonstrated satisfactory results for NR-IQA. In the literature, considerable effort has been made to learn a better feature representation. Some models employ a multi-task framework using the auxiliary information of a relevant sub-task (e.g., distortion classification [15] or semantic classification [16]) to enhance the features of the primary IQA sub-task. By exploiting the features of multiple sub-tasks and simultaneously optimising sub-tasks in an end-to-end fashion, these models can learn more discriminative feature representations from images. However, the challenge lies in obtaining adequate data for the auxiliary information, which limits the generalisation ability of this approach. Some models adopt multi-scale features to improve the prediction accuracy for the IQA task, e.g. in [17], [18], Transformers are used to extract multi-scale features, with the aim to fuse high-level semantic information and low-level texture information.

Notwithstanding the significant progress made in the NR-IQA problem, there is still room for improvement. A great deal of attention has been paid to improving the DCNN-based model's ability to learn features for quality prediction. However, little attention has been paid to the loss function. In order to predict a continuous variable of image quality, existing models are formulated as a regression problem and trained by minimising the loss function of the mean squared error (i.e., L2 norm). The L2 norm, however, ignores the relative ordering between different scores on the quality scale, which affects the model's performance in predicting human judgements. To account for this issue and to enable a model to produce quality scores in agreement with subjective ratings, we reformulate NR-IQA learning as an ordinal regression problem. To this end, we build a simple yet effective framework integrating convolutional neural

Manuscript received 27 February 2023; revised 28 March 2023; accepted 1 April 2023. Date of publication 18 April 2023; date of current version 24 April 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sheng Li. *(Corresponding author: Yulin Tu.)*

Huasheng Wang and Hantao Liu are with the School of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, U.K. (e-mail: WangHS@cardiff.ac.uk; LiuH35@cardiff.ac.uk).

Yulin Tu is with the Yonsei University, Wonju 26493, Korea (e-mail: tyl941215@163.com).

Xiaochang Liu is with the School of Materials, Sun Yat-sen University, Guangzhou 510275, China (e-mail: liuxiaochang8012@163.com).

Hongchen Tan is with the Institute of Artificial Intelligence, Beijing University of Technology, Beijing 100124, China (e-mail: tanhongchenphd@bjut.edu.cn).

Digital Object Identifier 10.1109/LSP.2023.3265569

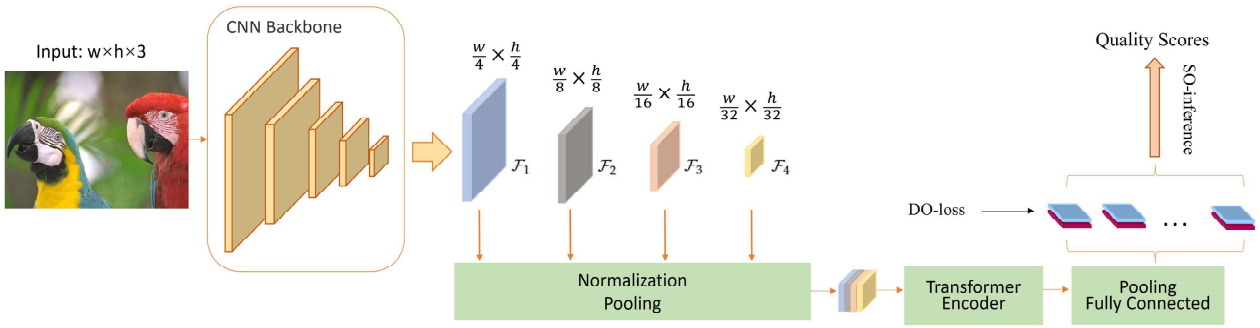

Fig. 1. The network architecture of the proposed no-reference image quality assessment (NR-IQA) model. The input image is first processed by the CNN encoder. Then the contextual information of feature maps is enhanced by the Transformer encoder. Finally, a deep ordinal loss (DO-loss) combined with a soft ordinal inference (SO-inference) is used to predict the image quality score.

networks (CNNs) and Transformers. In this framework, learning is achieved by a deep ordinal loss and using a soft ordinal inference to transform the predicted probabilities to a continuous variable for image quality. Experimental results show that this can give significant improvements for NR-IQA.

II. Proposed Method

We describe the proposed deep ordinal regression NR-IQA (DOR-IQA) in detail below.

A. Motivation

Inspired by the way human subjects rate image quality on a given scale [19], we could interpret this as a two-step process including placing the image first in one of the quality categories (e.g., intervals representing "bad" through "excellent" on a continuous scale from 0 to 10) based on overall perception of the given image space and then refining the quality score (i.e., producing a decimal number) by comparing the current image with other images falling into the locality of perceived quality. Based on this interpretation, our new proposal towards DCNN-based NR-IQA is to formulate the model as an ordinal regression (or ordinal classification) problem, which is better suited to address the human levels of preference.

It should be noted that optimising a standard regression network for the IQA problem could pose challenges such as slow convergence. Given that the image quality score ranges from 0 to 10 and retains three decimal places, a regression problem can be understood as a classification problem of predicting 10,000 categories, which causes a model to converge slow and to a poor solution. Also, the loss function (i.e., L2 norm and its variants) used by a regression network does not take into account the ordering of quality ratings, and consequently cannot capture human preference in the ground truth. Alternatively, if we treat this as an ordinal regression problem by dividing the entire range of quality scores into a series of intervals (e.g., ten discrete bins) in a sequential order, a DCNN-based model is trained to predict an ordinal variable (i.e., categorical quality score range). Then, the output probability distribution can be used to infer a decimal representing the final quality score. This will provide a plausible solution for faster convergence and higher consistency with ground truth.

B. IQA Framework

Based on above concept, we devise an ordinal regression framework for NR-IQA. This is a simple yet effective DCNN architecture without specific constrains or multiple sub-tasks. The proposed network architecture is illustrated in Fig. 1, which combines convolutional neural networks (CNNs) with Transformers. The model aims to train a feature encoder that can learn effective confidence of the quality categories. The resolution of input image I is $w \times h \times 3$ (w and h represent width and height, respectively). Let f_ϕ represent the proposed model with learnable parameters ϕ, which include all network parameters of the CNN backbone, Transformer and the last fully connected (FC) layer. Let $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$ denote the feature maps obtained by the last layers of the CNN backbone (i.e., higher-level features relevant for saliency), with their dimensions being $\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$ of the size of the input image, respectively. In order to integrate multi-scale features and capture the interactions of local and global information, we let $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$ through normalisation and pooling layers. We use Euclidean norm to normalise these feature maps before they enter the pooling layer, as the same approach taken in [20]. By doing this, image features are treated as a sequential input to the Transformer model, which allows capturing long-range dependencies and correlations between different parts of the image. More specifically, the sizes of $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ are unified to the same size of $\mathcal{F}_4$. Then they are concatenated to form a new feature map $\tilde{\mathcal{F}}$. Since $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$ are from different layers of CNNs representing different image properties, such as texture, edge and semantics, this makes $\tilde{\mathcal{F}}$ carry rich information about the image content. In addition, $\tilde{\mathcal{F}}$ is sent to a Transformer encoder, which contains multi-head attention mechanism. Transformers are used to extract features with enhanced contextual information. The Transformer encoder is implemented as per [21] and we define $\bar{\mathcal{F}}$ as the output of the module. Finally, $\bar{\mathcal{F}}$ is delivered to FC layer to obtain the prediction of image quality categories.

C. Deep Ordinal Regression

To reformulate IQA as an ordinal regression, we regard the quality score as a continuous value which also contains ordering information between different quality ratings. Therefore, we discretize the entire range of ground truth image quality scores into K sub-intervals of equal size,

$$d^* = \left\lfloor \frac{s^* - s_{\min}}{s_{\max} - s_{\min}} \times K \right\rfloor \tag{1}$$

where $\lfloor \cdot \rfloor$ denotes the floor function, $d^* \in \{0, 1, \ldots, K-1\}$ is the discrete label of ground truth, and s^* is the original continuous value of image quality score, $s_{\min}, s_{\max}$ represent the minimum and maximum scores in a given IQA database, respectively. Hence the ordinal thresholds $c^k \in \{0, 1, \ldots, c^{K-1}\}$ are computed as follows,

$$c^k = s_{\min} + \frac{s_{\max} - s_{\min}}{K-1} * k \tag{2}$$

where k denotes a specific bin. Now, we adopt the concept in [22] to design a deep ordinal loss (DO-loss) for the network. More specifically, this transforms a multi-class classification problem to a set of binary classification sub-problems: for each instance $c^k \in \{0, 1, \ldots, c^{K-1}\}$, a binary classifier is applied to predict whether the ordinal value of a sample is larger than c^k; and the ordinal value of an unseen sample is predicted on the basis of the classification results of the $K-1$ binary classifiers. We define the output of the model as $Y = f_\phi(I)$, where Y belongs to a 2K-dimensional vector. the DO-loss can be computed as follows,

$$L_{DO} = -\sum_{k=0}^{d^*-1} \ln \mathcal{P}^k - \sum_{d^*}^{K-1} \left(1 - \ln \mathcal{P}^k\right) \tag{3}$$

where $\mathcal{P}^k = \mathcal{P}(d > k) = \frac{e^{y_{2k+1}}}{e^{y_{2k}} + e^{y_{2k+1}}}$, d is the predicted discrete label, $\mathcal{P}^k$ is the ordinal probability when d is larger than k. Compared with the conventional use of cross-entropy loss to train a classification network, DO-loss can update DCNN network parameters more effectively [22]. Due to the innate ordinal properties contained in the quality ratings, the ordinal loss is more responsive to predictions that are inconsistent with the ordinal properties of ground truth.

Now, based on the output probabilities of K binary classification instances, the predicted image quality score s can be computed as,

$$\begin{aligned} s &= \frac{c^d + c^{d+1}}{2} \\ d &= \sum_{k=0}^{K-1} \eta\left(\mathcal{P}^k \geq 0.5\right) \end{aligned} \tag{4}$$

where $\eta(\cdot)$ denotes an indicator function where $\eta(true) = 1$ and $\eta(false) = 0$. The above operation (so-called hard ordinal inference [23]) involves the use of a hard threshold, which will lead to sudden changes (i.e., step effect [24]) in the transition region of the model. To take the advantage of the probability (or confidence) predicted by the network, we devise a soft ordinal inference (SO-inference) by adapting the method in [22] to our IQA context. The SO-inference can transform the output probabilities to a continuous variable for image quality,

$$s = \frac{c^d + c^{d+1}}{2} * (1 - \mathcal{D}) + \frac{c^{d+1} + c^{d+2}}{2} * \mathcal{D} \tag{5}$$

where $d = \lfloor h \rfloor, \mathcal{D} = h - d, h = \sum_{k=0}^{K-1} \mathcal{P}^k$. $\mathcal{D}$ is between 0 and 1, indicating the extent to which the predicted category is close to $d+1$. In this soft reference, due to the introduction of the adaptation factor $\mathcal{D}$, the predicted score will adapt to the sparsity of the discretization intervals, which enables the inferred score to be closer to the ground truth.

III. Experimental Results

A. Experimental Protocols

To demonstrate the superiority of our proposed approach, we rely on seven widely recognised IQA databases, including TID2013 [33], LIVE [34], LIVE-FB [31], KonIQ-10K [35], CSIQ [36], KADID-10K [37], and CLIVE [38]. We employ two commonly used criteria, the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SROCC) to evaluate the performance of IQA models. Both PLCC and SROCC range from 0 to 1 with a higher value indicating better performance. Our experiments are implemented on an NVIDIA GeForce RTX 3060 with PyTorch 1.8.0 and CUDA 11.2 for training and testing. Following the popular training strategy used in existing DCNN- and Transformer-based IQA methods [21], [29], we select 50 patches of 224×224 pixels each randomly from each input image. The parameters of CNN backbone (i.e., ResNet50 use in our model) are pre-trained on the ImageNet classification task [39]. The model is then trained end-to-end by using Adam [40] optimizer with learning rate 1×10^{-5}. We set at most 5 epochs and mini-batch size of 16 during traning process.

B. Performance Evaluation

Table I lists the performance of the proposed DOR-IQA and other 12 state-of-the-art DCNN-based IQA metrics, where the best and second- and third-best results are labelled in red, blue and green colours, respectively. As shown in the table, our proposed DOR-IQA achieves superior performance on PLCC and SROCC. Our proposed model outperforms the existing methods by a significant margin on both the KADID and KonIQ databases, which represent large-scale IQA databases. More specifically, our model's performance is 2.7%, 2.4% (PLCC, SROCC) higher than that of TRes (second-best) on the KADID database; and is 1.5%, 1.6% (PLCC, SROCC) higher than that of TRes on the KonIQ database. This demonstrates our model's ability in handling complex and diverse natural scenes (as the feature of KADID and KonIQ databases).

To verify the superiority of using DO-loss rather than L2-loss (as per the plausible reason described in Section II-A) and demonstrate that DO-loss can be effectively embedded in other DCNNs, we conduct an ablation study. In our experiments, we use two popular baselines VGG [41] and Resnet-50 [42]; and implement both L2-loss and DO-loss by modifying the last fully

TABLE I
PERFORMANCE COMPARISION OF IQA METRICS

	LIVE		CSIQ		TID2013		KADID		CLIVE		KonIQ		LIVEFB	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
DIIVINE [25]	0.908	0.892	0.776	0.804	0.567	0.643	0.435	0.413	0.591	0.588	0.558	0.546	0.187	0.092
BRISQUE [26]	0.944	0.929	0.748	0.812	0.571	0.626	0.567	0.528	0.629	0.629	0.685	0.681	0.341	0.303
ILNIQE [14]	0.906	0.902	0.865	0.822	0.648	0.521	0.558	0.534	0.508	0.508	0.537	0.523	0.332	0.294
BIECON [27]	0.961	0.958	0.823	0.815	0.762	0.717	0.648	0.623	0.613	0.613	0.654	0.651	0.428	0.407
MEON [15]	0.955	0.951	0.864	0.852	0.824	0.808	0.691	0.604	0.710	0.697	0.628	0.611	0.394	0.365
WaDIQaM [6]	0.955	0.960	0.844	0.852	0.855	0.835	0.752	0.739	0.671	0.682	0.807	0.804	0.467	0.455
DBCNN [28]	0.971	0.968	0.959	0.946	0.865	0.816	0.856	0.851	0.869	0.869	0.884	0.875	0.551	0.545
TIQA [29]	0.965	0.949	0.838	0.825	0.858	0.846	0.855	0.850	0.861	0.845	0.903	0.892	0.581	0.541
MetaIQA [30]	0.959	0.960	0.908	0.899	0.868	0.856	0.775	0.762	0.802	0.835	0.856	0.887	0.507	0.540
P2P-BM [31]	0.958	0.959	0.902	0.899	0.856	0.862	0.849	0.840	0.842	0.844	0.885	0.872	0.598	0.526
HyperIQA [32]	0.966	0.962	0.942	0.923	0.858	0.840	0.845	0.852	0.882	0.859	0.917	0.906	0.602	0.544
TReS [21]	0.968	0.969	0.942	0.922	0.883	0.863	0.858	0.859	0.877	0.846	0.928	0.915	0.625	0.554
DOR-IQA	0.978	0.977	0.961	0.945	0.901	0.887	0.885	0.883	0.891	0.871	0.943	0.931	0.643	0.573

TABLE II
PERFORMANCE OF IQA BASED ON L2-LOSS VERSUS DO-LOSS

Dataset	Method	PLCC	SROCC	Dataset	Method	PLCC	SROCC
LIVE	VGG+L2	0.868	0.839	CSIQ	VGG+L2	0.823	0.801
	VGG+DO-loss	0.903	0.894		VGG+DO-loss	0.865	0.839
	Resnet-50+L2	0.911	0.895		Resnet-50+L2	0.886	0.867
	Resnet-50+DO-loss	0.954	0.941		Resnet-50+DO-loss	0.926	0.914
KonIQ	VGG+L2	0.822	0.805	TID2013	VGG+L2	0.803	0.786
	VGG+DO-loss	0.853	0.841		VGG+DO-loss	0.851	0.842
	Resnet-50+L2	0.865	0.849		Resnet-50+L2	0.842	0.821
	Resnet-50+DO-loss	0.903	0.886		Resnet-50+DO-loss	0.875	0.954
CLIVE	VGG+L2	0.817	0.796	KADID	VGG+L2	0.774	0.776
	VGG+DO-loss	0.849	0.834		VGG+DO-loss	0.821	0.832
	Resnet-50+L2	0.843	0.821		Resnet-50+L2	0.827	0.825
	Resnet-50+DO-loss	0.867	0.856		Resnet-50+DO-loss	0.851	0.858

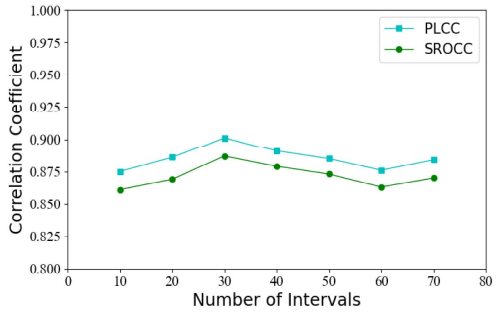

Fig. 2. Model performance versus different choices of K on the TID2013 dataset.

connected layer of the baseline network. Table II shows the performance of IQA variants using L2-loss versus DO-loss on popular IQA databases. It can be seen that by equipping IQA with our proposed DO-loss, both baselines can achieve more than a 4% increase in PLCC/SROCC.

We also explain the choice of K-discretization intervals (see (1)) in our model. To reveal the impact of the choice of K on the model's performance, we conduct experiments using different choices of K (i.e., K=10, 20,...70). Fig. 2 illustrates the model performance versus different choices of K on the TID2013 dataset. Note, the results on other databases show the same trend therefore not visualised here. As shown in Fig. 2, the model's performance peaks at K=30 and tends to be saturated onwards. Therefore, K=30 is used in our model.

IV. CONCLUSION

In this letter, we have proposed a new framework for no-reference image quality assessment based on deep ordinal regression. The training process is regularised by a deep ordinal loss, aiming to learn efficient representations of the probability (confidence) of the quality categories. Then a soft ordinal inference is used to transform the discrete prediction output to a continuous variable for image quality. To the best of our knowledge, this is the first attempt to build an ordinal regression NR-IQA model with an effective solution. Extensive experimental results on popular databases demonstrate the superior performance of our proposed model against the state-of-the-art. The proposed framework can be easily extended to design advanced IQA models in the future.

REFERENCES

[1] Y. Guo, G. Ding, and J. Han, "Robust quantization for general similarity search," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 949–963, Feb. 2018.

[2] L. Li, Y. Yan, Y. Fang, S. Wang, L. Tang, and J. Qian, "Perceptual quality evaluation for image defocus deblurring," *Signal Process.: Image Commun.*, vol. 48, pp. 81–91, 2016.

[3] Q. Jiang, F. Shao, W. Gao, H. Li, and Y.-S. Ho, "A risk-aware pairwise rank learning approach for visual discomfort prediction of stereoscopic 3D," *IEEE Signal Process. Lett.*, vol. 26, no. 11, pp. 1588–1592, Nov. 2019.

[4] Y. Wang, "Image quality assessment based on gradient complex matrix," in *Proc. IEEE Int. Conf. Syst. Inform.*, 2012, pp. 1932–1935.

[5] X. Min, K. Gu, G. Zhai, M. Hu, and X. Yang, "Saliency-induced reduced-reference quality index for natural scene and screen content images," *Signal Process.*, vol. 145, pp. 127–136, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168417303857

[6] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[7] Q. Jiang, Z. Peng, S. Yang, and F. Shao, "Authentically distorted image quality assessment by learning from empirical score distributions," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1867–1871, Dec. 2019.

[8] L. Li, H. Zhu, G. Yang, and J. Qian, "Referenceless measure of blocking artifacts by Tchebichef kernel analysis," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 122–125, Jan. 2014.

[9] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[10] Q. Li, W. Lin, J. Xu, and Y. Fang, "Blind image quality assessment using statistical structural and luminance features," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2457–2469, Dec. 2016.

[11] K. Zhang, Y. Fang, W. Chen, Y. Xu, and T. Zhao, "A display-independent quality assessment for HDR images," *IEEE Signal Process. Lett.*, vol. 29, pp. 464–468, 2022.

[12] P. Chen, L. Li, Q. Wu, and J. Wu, "SPIQ: A self-supervised pre-trained model for image quality assessment," *IEEE Signal Process. Lett.*, vol. 29, pp. 513–517, 2022.

[13] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[14] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[15] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[16] D. Li, T. Jiang, and M. Jiang, "Exploiting high-level semantics for no-reference image quality assessment of realistic blur images," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 378–386.

[17] K. Gu et al., "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.

[18] J. Guan, S. Yi, X. Zeng, W.-K. Cham, and X. Wang, "Visual importance and distortion guided deep image quality assessment framework," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2505–2520, 2017.

[19] B. Keelan, *Handbook of Image Quality: Characterization and Prediction*. Boca Raton, FL, USA: CRC, 2002.

[20] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2020.

[21] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1220–1230.

[22] Y. Chen, H. Zhao, Z. Hu, and J. Peng, "Attention-based context aggregation network for monocular depth estimation," *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 6, pp. 1583–1596, 2021.

[23] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3174–3182, Nov. 2018.

[24] W. Zhang, R. R. Martin, and H. Liu, "A saliency dispersion measure for improving saliency-based image quality metrics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1462–1466, Jun. 2018.

[25] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

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

[27] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[28] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2018.

[29] J. You and J. Korhonen, "Transformer for image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 1389–1393.

[30] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14143–14152.

[31] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3575–3585.

[32] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3667–3676.

[33] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Process.: Image Commun.*, vol. 30, pp. 57–77, 2015.

[34] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[35] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10K: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.

[36] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, 2010, Art. no. 011006.

[37] H. Lin, V. Hosu, and D. Saupe, "Kadid-10K: A large-scale artificially distorted IQA database," in *Proc. IEEE 11th Int. Conf. Qual. Multimedia Experience*, 2019, pp. 1–3.

[38] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

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

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.