frontiers in Neuroscience

ORIGINAL RESEARCH
published: 14 January 2022
doi: 10.3389/fnins.2021.767977

Cross-Domain Feature Similarity Guided Blind Image Quality Assessment

Chenxi Feng[1], *Long Ye*[2]* *and Qin Zhang*[2]

[1] Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China, Beijing, China, [2] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

This work proposes an end-to-end cross-domain feature similarity guided deep neural network for perceptual quality assessment. Our proposed blind image quality assessment approach is based on the observation that features similarity across different domains (e.g., Semantic Recognition and Quality Prediction) is well correlated with the subjective quality annotations. Such phenomenon is validated by thoroughly analyze the intrinsic interaction between an object recognition task and a quality prediction task in terms of characteristics of the human visual system. Based on the observation, we designed an explicable and self-contained cross-domain feature similarity guided BIQA framework. Experimental results on both authentical and synthetic image quality databases demonstrate the superiority of our approach, as compared to the state-of-the-art models.

Keywords: cross-domain feature similarity, image quality assessment, deep learning, transfer learning, human visual system

OPEN ACCESS

Edited by:
Guangtao Zhai,
Shanghai Jiao Tong University, China

Reviewed by:
Xiongkuo Min,
University of Texas at Austin, United States
Huiyu Duan,
Shanghai Jiao Tong University, China
Wei Sun,
Shanghai Jiao Tong University, China

***Correspondence:**
Long Ye
yelong@cuc.edu.cn

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

Received: *31 August 2021*
Accepted: *08 November 2021*
Published: *14 January 2022*

Citation:
Feng C, Ye L and Zhang Q (2022) Cross-Domain Feature Similarity Guided Blind Image Quality Assessment. Front. Neurosci. 15:767977. doi: 10.3389/fnins.2021.767977

1. INTRODUCTION

Objective image quality assessment (IQA) aims to enable computer programs to predict the perceptual quality of images in a manner that is consistent with human observers, which has become a fundamental aspect of modern multimedia systems (Zhai and Min, 2020). Based on how much information the computer program could access from the pristine (or reference) image, objective IQA could be categorized into full-reference IQA (FR-IQA) (Wang et al., 2003, 2004; Sheikh and Bovik, 2006; Larson and Chandler, 2010a; Li et al., 2011; Zhang et al., 2011, 2014; Liu et al., 2012; Chang et al., 2013; Xue et al., 2013), reduced-reference IQA (RR-IQA) (Wang and Simoncelli, 2005; Wang and Bovik, 2011; Rehman and Wang, 2012), and no-reference (or blind) IQA (NR-IQA/BIQA) (Kim and Lee, 2016; Liu et al., 2017; Ma et al., 2017a; Lin and Wang, 2018; Pan et al., 2018; Talebi and Milanfar, 2018; Sun et al., 2021). The absence of reference information in most real-world multimedia systems calls for BIQA methods, which are more applicable but also more difficult.

Deep neural network (DNN) has significantly facilitated various image processing tasks (Fang et al., 2017; Park et al., 2018; Casser et al., 2019; Ghosal et al., 2019) in recent years due to its powerful capacity in feature abstraction and representation. It is also worth noting that the success of deep-learning techniques is derived from large amounts of training data, which is often leveraged to adjust the parameters in the DNN architecture to guarantee that both the accuracy and generalization ability are satisfying. Unfortunately, image quality assessment is typically a small-sample problem since the annotation of the ground-truth quality labels calls for time-consuming subjective image quality experiments (Zhang et al., 2018a). Inadequate quality

annotations severely restrict the performance of DNN-based BIQA models in terms of both accuracy and generalization ability.

In order to address the problem caused by limited subjective labels, data augmentation is firstly employed to increase the training labels (e.g., Kang et al. (2014)) proposed to split the image with quality labels into multiple patches. and each of the patches is assigned with a quality score which is the same with the whole image. However, some distortion types are inhomogeneous, i.e., the perceptual quality of local patches might differ from the overall quality of the whole image. Therefore, transfer learning has gained more attention to relieve the small-sample problem (Li et al., 2016). Specifically, the BIQA framework is comprised of two stages: which are pre-training and fine-tuning. In the pre-training stage, the parameters in the DNN architecture are trained by other image processing tasks such as object recognition, whilst in the fine-tuning stage, images with subjective labels are employed as training samples. Such a transfer-learning scheme is feasible since the low-level feature extraction procedure across different image processing tasks are shared (Tan et al., 2018).

More recently, various sources of external knowledge are incorporated to learn a better feature representation for the BIQA issue. For example, hallucinated reference (Lin and Wang, 2018) is generated via a generative network and employed to guide the quality-aware feature extraction. The distortion identification is incorporated as the auxiliary sub-task in MEON model (Ma et al., 2017b), by which the distortion type information is transparent to the primary quality prediction task for better quality prediction. Visual saliency is employed in Yang et al. (2019) to weight the quality-aware features more reasonably. Semantic information is also employed for better understanding of the intrinsic mechanism of quality prediction, e.g., multi-layer semantic features are extracted and aggregated through several statistical structures in Casser et al. (2019). An effective hyper network is employed in Su et al. (2020) to generate customized weights from the semantic feature for quality prediction, i.e., the quality perception rule differs as the image content changes.

Unlike other studies, this paper employs the cross-domain feature similarity as an extra restraint for better quality-aware feature representation. Specifically, the transfer-learning based BIQA approach is pre-trained in one domain (say, object recognition in the semantic domain) and is fine-tuned in the perceptual quality domain with similar DNN architectures, we have observed that the cross-domain (Semantic vs. Quality) feature similarity would, in turn, contribute to the quality prediction task (as shown in **Figure 1**).

By thoroughly analyzing the intrinsic interaction between object recognition task and quality prediction task, we think the phenomenon represented in **Figure 1** is sensible. As shown in **Figure 2**, previous works (Larson and Chandler, 2010b) have revealed that human observers would take different strategies to assess the perceptual quality when viewing images with different amounts of degradation: when judging the quality of a distorted image containing near-threshold distortions, one tends to rely primarily on visual detection of any visible local differences, in such a scenario, semantic information is instructive for quality perception since distortion in the semantic-sensitive area would contribute more in the quality decision and vice versa. On the other hand, when judging the quality of a distorted image with clearly visible distortions, one would rely much less on visual detection and much more on the overall image appearance, in such a scenario, the quality decision procedure is much more independent with semantic information.

Considering the effectiveness of cross-domain feature similarity (CDFS), this work leverages CDFS as an extra restraint to improve the prediction accuracy of BIQA models. As shown in **Figure 3**, the parameters in our CDFSNet are updated according to both the basic loss and the extra loss, which would restrain the network yielding quality predictions as similar as the ground-truth label whilst maintaining that the CDFS also correlates well with the perceptual quality, in such a manner that, the accuracy of the DNN architecture would get improved according to the experimental results presented in section 3.

Compared to the aforementioned works, the superiority of the cross-domain feature similarity guided BIQA framework is embodied in the following aspects:

(1) The proposed cross-domain feature similarity is self-contained for transfer-learning based BIQA models since the transfer-learning procedure itself is comprised of the training in two different domains (i.e., object recognition and quality prediction). Therefore, no extra annotation procedure (such as distortion identification in Ma et al., 2017b and visual saliency in Yang et al., 2019) is needed.
(2) The proposed cross-domain feature similarity is more explicable since it is derived from the intrinsic characteristic of interactions between semantic recognition and quality perception.
(3) In addition to general-purpose IQA, the performance of our proposed CDFS guided BIQA framework is also evaluated on other specific scenarios such as screen content (Xiongkuo et al., 2021) and dehazing oriented (Min et al., 2018b, 2019) IQA. The experimental results indicate that CDFS guided BIQA has significant potential toward diverse types of BIQA tasks (Min et al., 2020a,b).

The rest part of the paper is organized as follows: Section 2 illustrates the details of our CDFS-based BIQA framework and section 3 shows the experimental results; Section 4 is the conclusion.

2. MATERIALS AND METHODS

2.1. Problem Formulation

Let x denote the input image, conventional DNN based BIQA works usually leverage an pre-trained DNN architecture $f(\cdot;\theta)$ (with learnable parameters θ) to predict the perceptual quality of x via $\hat{q} = f(x;\theta)$, where $\hat{q}$ denotes the prediction of perceptual quality q.

Our work advocates employing the cross-domain feature similarity to supervise the update of parameters in a quality prediction network. Specifically, let $f(\cdot;\theta_{Smtc})$ denotes the DNN with fixed and pre-trained parameters oriented toward semantic recognition, and $f(\cdot;\theta_{Qlty})$ denotes the DNN with learnable

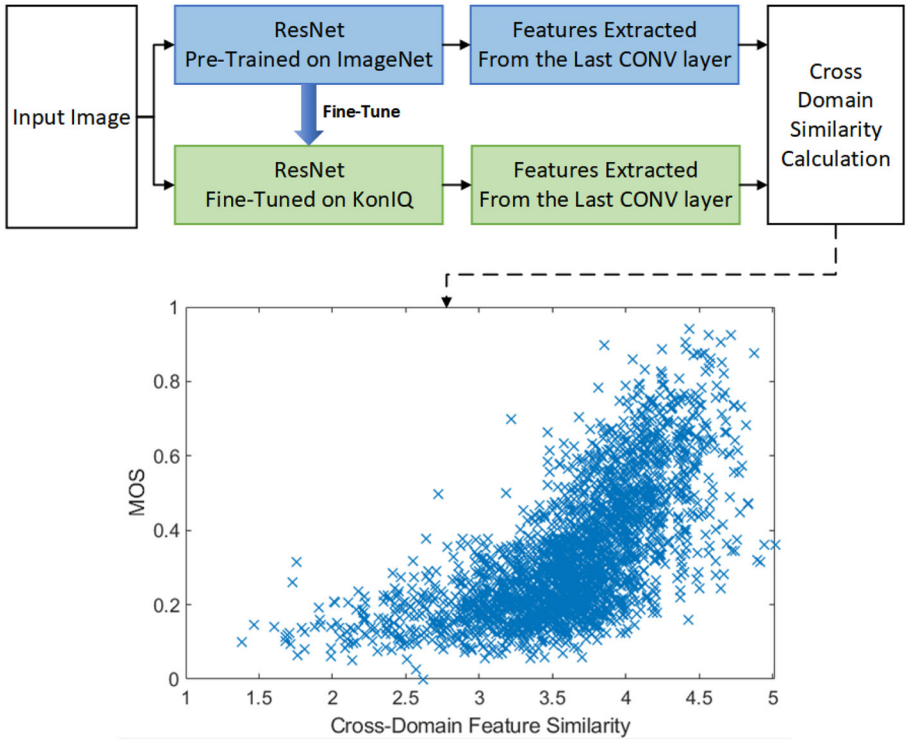

FIGURE 1 | The overall framework of our proposed CDFS guided BIQA approach. As shown in the lower part, the cross-domain feature similarity is highly correlated with the perceptual quality. The 'cross-domain similarity calculation' is obtained by: (1) Extractinged the features from the last convolutional layer of pre-trained ResNet (denoted as R_s) and fine-tuned ResNet (denoted as R_q); (2) Calculatinge the similarity matrix W according to Equation 1; (3) Obtaining the eigen values of W by $\vec{v} = eig(W)$; (4) The similarity Sim is calculated by $Sim = \frac{1}{std(v)}$, in which $std(\cdot)$ denotes the standard deviation operator.

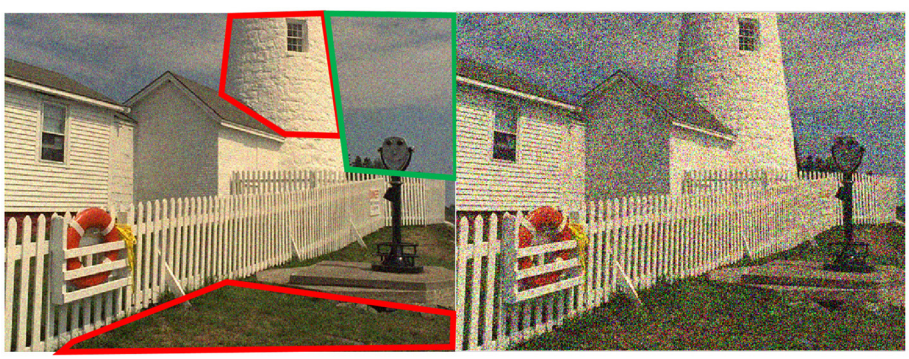

FIGURE 2 | Illustration of different strategies that the human visual system would take to assess the perceptual quality when viewing images with different amounts of degradation. Specifically, when judging the quality of a distorted image containing near-threshold distortions (Left), one tends to rely primarily on visual detection of any visible local differences, e.g., the distortions in red boxed are slighter than that in the green box even though the noise intensity is the same. On the other hand, when judging the quality of a distorted image with clearly visible distortions (Right), one would rely much less on visual detection and much more on overall image appearance e.g., the distortions in each image area are roughly the same.

parameters oriented toward quality prediction. It should be noticed that $f(\cdot; \theta_{Smtc})$ and $f(\cdot; \theta_{Qlty})$ share the same architectures whilst having own different parameters. This work attempts to further improve the quality prediction accuracy by analyzing the similarity between the features extracted for different tasks, i.e., features extracted for semantic recognition $ft_s = f(x; \theta_{Smtc})$, and features extracted for quality regression $ft_q = f(x; \theta_{Qlty})$.

Given three-dimensional features ft_s and ft_q with size $[C, H, W]$, where C, H, W denotes the channel size, height, and width of the features, respectively, ft_s and ft_q are firstly reshaped

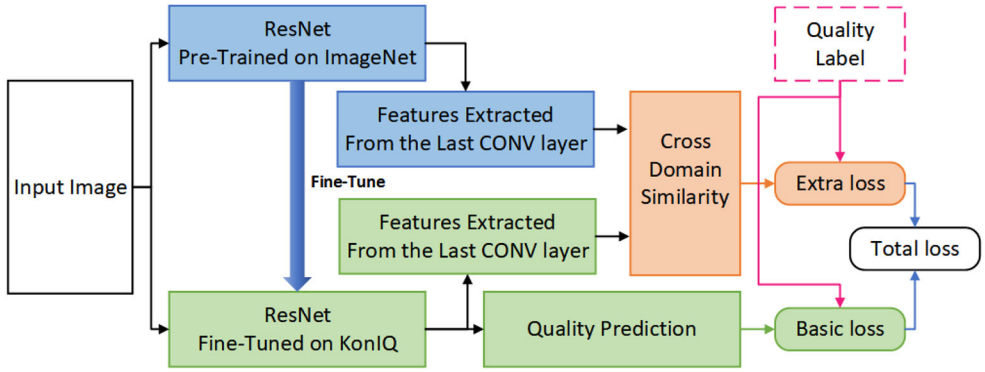

FIGURE 3 | The overall pipeline of our proposed CDFS-based IQA approach.

into R_q and R_s with size $[C, H \times W]$. The similarity Sim between R_q and R_s is obtained via the following steps.

Step 1, employ linear regression to express R_q via R_s, i.e., $R_q = W \times R_s + e$, where W denotes the weighting matrix and e denotes the prediction error of linear regression. Therefore, W could be obtained by

$$W = (R_s^T \times R_s)^{-1} \times R_s^T \times R_q \tag{1}$$

Step 2, a learnable DNN architecture $g(\cdot; \gamma)$ is employed to yield the similarity between ft_s and ft_q given W, i.e., $Sim = g(W; \gamma)$

2.2. Network Design

The architecture of our proposed network is shown in **Figure 4**, which mainly consists of a semantically oriented feature extractor, perceptual-quality oriented feature extractor, and cross-domain feature similarity predictor. More details are described as follows.

2.2.1. Semantic Oriented Feature Extractor

The DNN pre-trained in large-scale object recognition datasets (e.g., ImageNet Deng et al., 2009) are leveraged as the semantic oriented feature extractor.

Specifically, this work employs the activations of the last convolutional layers in ResNet50 to represent the semantic-aware features ft_s of an specific image, i.e., $ft_s = f(x; \theta_{Smtc})$.

It is worth noting that θ_{Smtc} is fixed during the training stage since the proposed DNN framework will be fine-tuned in IQA datasets in which the semantic label is unavailable.

2.2.2. Perceptual-Quality Oriented Feature Extractor

The architecture of perceptual-quality oriented feature extractor $f(\cdot; \theta_{Qlty})$ is quite similar with semantic oriented feature extractor. However, the parameters θ_{Qlty} in $f(\cdot; \theta_{Qlty})$ are learnable and independent with θ_{Smtc}.

The quality-aware features $ft_q = f(x; \theta_{Qlty})$ are further leveraged to aggregate the prediction of subjective quality score, i.e., $\hat{q} = h(ft_q; \delta)$, in which q denotes the subjective quality score (MOS), $\hat{q}$ is the prediction of q, and $h(\cdot; \delta)$ stands for the MOS prediction network given quality-aware features with learnable parameters δ.

2.2.3. Cross-Domain Feature Similarity Predictor

As illustrated in section 1, the cross-domain feature similarity would contribute to the prediction of perceptual quality. However, directly evaluating the similarity between ft_s and ft_q via Minkowski-Distance or Wang-Bovik metric (Wang et al., 2004) is not as efficient, as shown in **Figure 6**. We think the invalidation of the Wang-Bovik metric is mainly attributed to its pixel-wise sensitivity, i.e., any turbulence during the parameter initializing and updating of the DNN framework would result in a significant difference between ft_s and ft_q.

To this end, this work proposes to depict the cross-domain feature similarity through a global perspective. Specifically, the similarity is derived from the weighting matrix W which is employed to reconstruct ft_q given ft_s via linear regression. Since the W is derived from the features amongst all channels, it is less likely to suffer from the instability of the DNN during initializing and updating. The experiments reported in section 3.3 also demonstrate the superiority of our proposed similarity measurement for cross-domain features. In our CDFS-guided BIQA framework, the CDFS is incorporated as follows:

Linear regression is employed for the reconstruction and the weighting matrix W could be obtained according to equation 1 and ***Step 1*** in section 2.1

A stack of convolutional layers (denoted as $g(\cdot;\gamma)$) is followed to learn the cross-domain feature similarity given W.

During the training stage, the cross-domain similarity is employed as a regularization item to supervise the quality prediction network.

2.2.4. Loss Function

The loss function L of our proposed network is designed as

$$L_1 = argmin_{[\theta_{Qlty}, \delta]} \left\| q - h(f(x; \theta_{Qlty}); \delta) \right\| \tag{2}$$

$$L_2 = argmin_{[\theta_{Qlty}, \gamma]} \left\| q - g(W; \gamma) \right\| \tag{3}$$

and

$$L = L_1 + \lambda L_2 \tag{4}$$

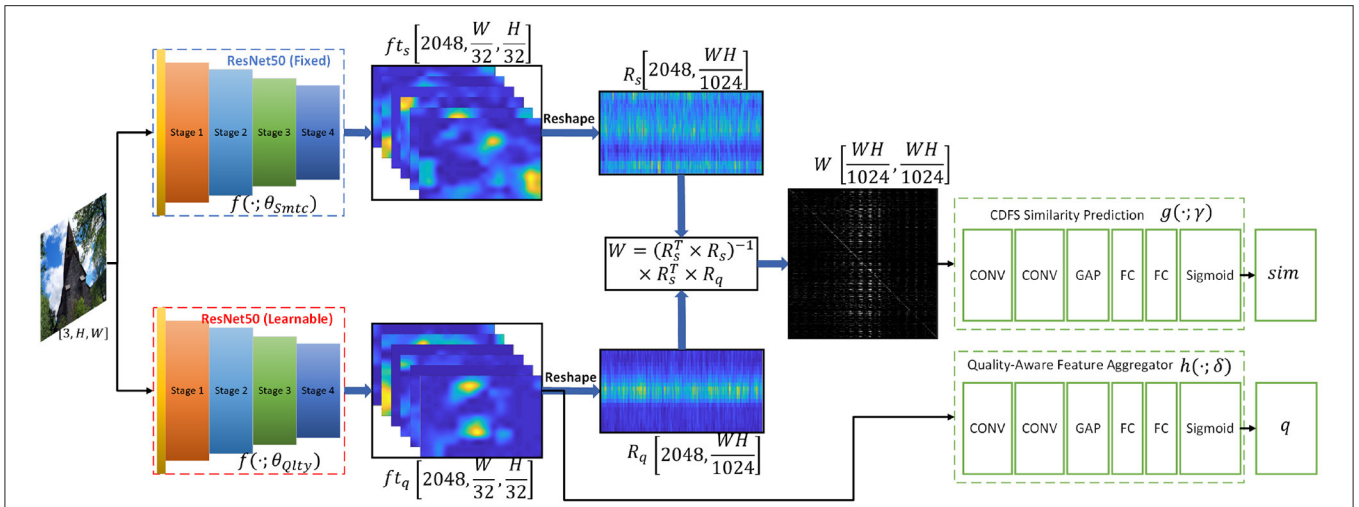

FIGURE 4 | The detailed architecture of our proposed Cross-Domain Feature Similarity Guided Network. The "CONV" denotes convolutional layers followed by batch normalization and ReLU layer, the "FC" denotes the fully-connected layer, and the 'GAP' denotes the global average pooling layer.

where $\|\cdot\|$ denotes the $L1$ norm operator, W is calculated according to equation 1, and λ is a hyper parameter controlling the weights of L_1 and L_2.

2.3. Implementation Details

We use ResNet50 (He et al., 2016) as the backbone model for both the semantically oriented feature extractor and the perceptual-quality oriented feature extractor. As aforementioned, the pre-trained model on ImageNet (Deng et al., 2009) is used for network initialization. During the training stage, the θ_{Smtc} is fixed whilst θ_{Qlty} is learnable. In our network, the last two layers of the origin ResNet50, i.e., an average pooling layer and a fully connected layer, are removed to output features ft_s and ft_q.

For quality regression, a global average pooling (GAP) layer is used to pool the features ft_q into one-dimensional vectors, then three fully -connected (FC) layers are followed with size 2048-1024-512-1 and activated by ReLu, except for the last layer (activated by sigmoid).

The $g(\cdot;\gamma)$ in cross-domain feature similarity predictor is implemented by 3 three stacked convolutional layers, a GAP layer, and three FC layers. The architectures of convolutional layers are $in(1)-out(32)-k(1)-p(0)$, $in(32)-out(64)-k(3)-p(1)$, and $in(64)-out(128)-k(3)-p(1)$, respectively, where $in(\alpha)-out(\beta)-k(x)-p(y)$ denotes the input channel size and output channel size is α and β, the kernel size is x, and the padding size is y. Each of the convolutional layers is followed by a batch normalization layer and a ReLu layer. The GAP layer and the FC layers are the same with quality regression except that the size of FC layers is 128-512-512-1.

The experiment is conducted on Tesla V100P GPUs, while the DNN modules are implemented by Pytorch. The size of minibatch is 24. Adam (Kingma and Ba, 2014) is adopted to optimize the loss function with weight decay 5×10^{-4} and learning rate 1×10^{-5} for parameters in baseline (ResNet) and 1×10^{-4} for other learnable parameters. As mentioned, the parameters in semantic oriented feature extractor is are fixed, i.e., the learning rate is 0 for θ_{Smtc}.

3. EXPERIMENTAL RESULTS

3.1. Datasets and Evaluation Metrics

Three image databases including KonIQ-10k (Hosu et al., 2020), LIVE Challenges (LIVEC) (Ghadiyaram and Bovik, 2015), and TID2013 (Ponomarenko et al., 2015) are employed to validate the performance of our proposed network. The KonIQ-10k and LIVEC are authentically distorted image databases containing 10,073 and 1,162 distorted images, respectively, and the TID2013 is a synthetic image database containing 3,000 distorted images.

Two commonly used criteria, Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC), are adopted to measure the prediction monotonicity and the prediction accuracy. For each database, 80% images are used for training, and the others are used for testing. The synthetic image database is split according to reference images. All the experiments are under five times random train-test splitting operation, and the median SRCC and PLCC values are reported as final statistics.

3.2. Comparison With the State-of-the-Art Methods

Ten BIQA methods are selected for performance comparison, including five hand-crafted based (BRISQUE Mittal et al., 2012, ILNIQE Xu et al., 2016, HOSA Zhang et al., 2015, BPRI Min et al., 2017a, BMPRI Min et al., 2018a) and five DNN-based approaches (SFA Li et al., 2018, DBCNN Zhang et al., 2018b, HyperIQA Su et al., 2020, SDGNet Yang et al., 2019). The experimental results are shown as in **Table 1**.

As shown in **Table 1**, our method outperforms all the SOTA methods on the two authentic image databases in terms of SRCC. As for PLCC measurement, our method achieves

TABLE 1 | Performance comparison in terms of PCLL and SRCC on KonIQ, LIVEC, and TID2013, respectively.

SRCC	**KonIQ**	**LIVEC**	**TID2013**
BRISQUE	0.665	0.608	0.572
ILNIQE	0.507	0.432	0.521
HOSA	0.671	0.640	0.688
BPRI	–	–	0.899
BMPRI	–	–	**0.929**
SFA	0.856	0.812	–
DBCNN	0.875	0.851	–
HyperIQA	0.906	0.859	–
SGDNet	0.903	0.851	0.843
DeepFL	0.877	0.734	0.858
ours	**0.918**	**0.865**	0.899
PLCC	**KonIQ**	**LIVEC**	**TID2013**
BRISQUE	0.681	0.645	0.651
ILNIQE	0.523	0.508	0.648
HOSA	0.694	0.678	0.764
BPRI	–	–	0.892
BMPRI	–	–	**0.947**
SFA	0.872	0.833	–
DBCNN	0.884	0.869	–
HyperIQA	0.917	**0.882**	–
SGDNet	0.920	0.872	0.861
DeepFL	0.887	0.769	0.876
ours	**0.928**	0.875	0.880

Values in bold represents the highest value.

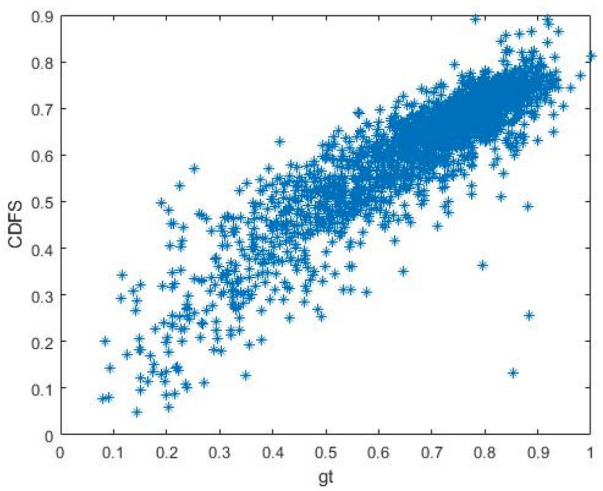

FIGURE 5 | The scatter plot of CDFS vs. MOS on KonIQ.

TABLE 2 | Ablation results in terms of SRCC and PLCC on KonIQ.

Modules	**BaseLine**	**+SP_wang**	**+SP_W**
SRCC	0.842	0.895	0.918
Gain(%)	–	6.3	9.0
PLCC	0.849	0.913	0.928
Gain(%)	–	7.5	9.3

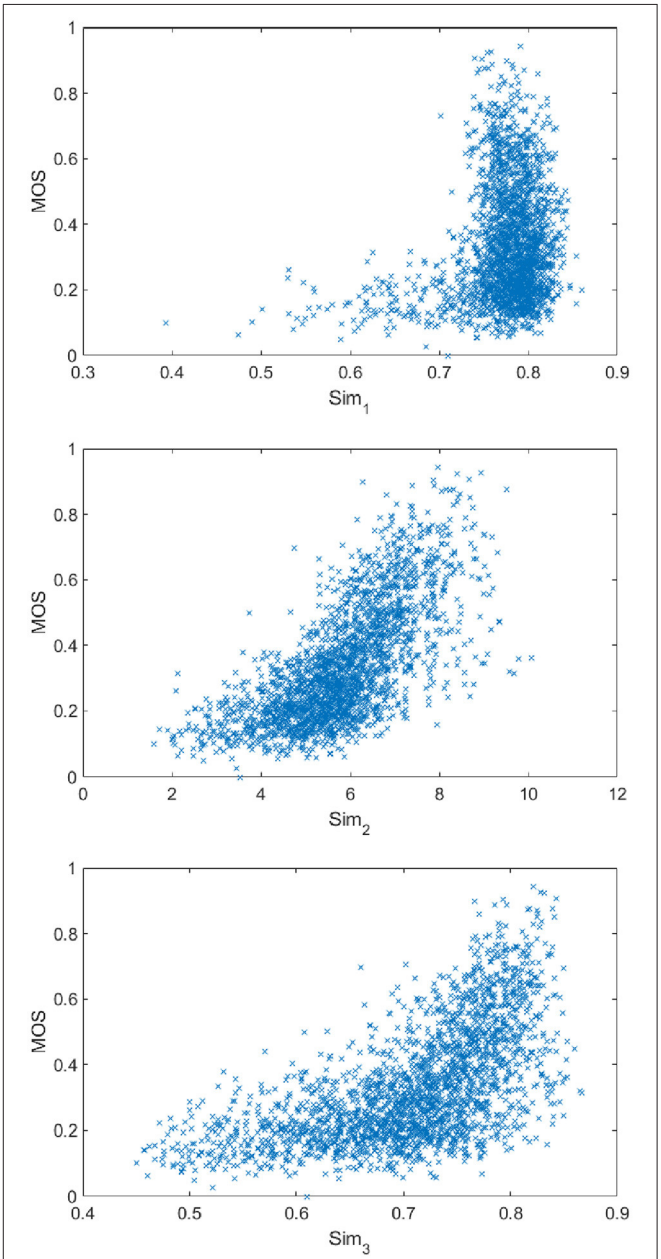

FIGURE 6 | The scatter plot of Sim_1, Sim_2, and Sim_3 vs. MOS on KonIQ.

the best performance on KonIQ and competing (the second) performance on LIVEC. This suggests that calculating cross-domain feature similarity for quality prediction refinement is effective. Though we do not especially modify the networks for synthetic image feature extraction, the proposed network has achieved competing performance in TID2013. Specifically, the proposed approach achieves the second-highest performance in terms of SRCC and the third-highest performance in terms of PLCC on TID2013.

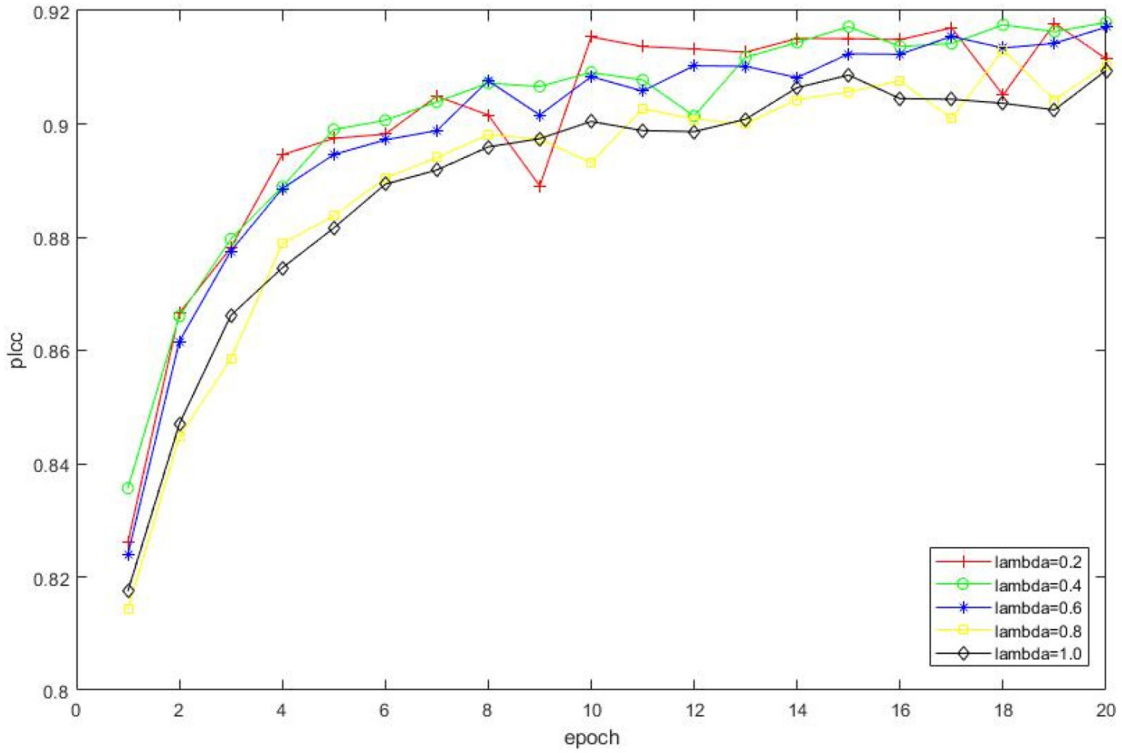

FIGURE 7 | Impact on selections of different λ. The experimental result is conducted on KonIQ, and a total of 20 epochs are involved.

3.3. Cross-Domain Feature Similarity Visualization

In order to further illustrate the superiority of our proposed CDFS, we firstly present the scatter plot of CDFS vs. MOS on KonIQ in **Figure 5**, indicating the CDFS is well correlated with perceptual quality.

In addition, we also investigate several non-learnable approaches for calculating CDFS: (1) $Sim_1 = mean(\frac{2\times ft_s \times ft_q + C}{ft_s^2 + ft_q^2 + C})$, where C denotes the constant to avoid numerical singularity; and (2) $Sim_2 = std(eig(W))$; (3) $Sim_3 = mean(\frac{2\times \vec{v}\times \vec{1} + C}{\vec{v}^2 + \vec{1}^2 + C})$, and $\vec{v} = eig(W)$, in which $\vec{1}$ denotes the vectors with the same size as $\vec{v}$ whilst whose elements are all 1.

Therefore, the calculation of Sim_1 is directly comparing the difference between ft_s and ft_q, and the calculation of Sim_2 and Sim_3 is based on the W derived according to equation 1. As shown in **Figure 6**, Sim_2 and Sim_3 is more correlated with the subjective score, demonstrating that measuring the cross-domain feature similarity based on W is more effective.

3.4. Ablation Study

Ablation study is conducted on KonIQ-10k to validate the efficiency of our proposed components, including the ResNet50 backbone (BaseLine), the similarity predictor (SP) obtained by Wang-Bovik metric (SP_wang, similar as Sim_1 in section 3.3), and the similarity predictor derived from the weighting metric W (SP_W). The results are shown in **Table 2**, indicating that incorporating a cross-domain similarity predictor could significantly improve the accuracy of quality prediction. Our proposed similarity measurement has achieved a great PLCC improvement (1.8%) compared to SP_wang and a more significant SRCC improvement (2.7%).

The impact of λ in equation 4 is also investigated, i.e., we set $\lambda = [0.2, 0.4, 0.6, 0.8, 1.0]$, respectively and observe the corresponding performance as shown in **Figure 7**. Therefore, we select $\lambda = 0.4$ for performance comparison and the following experiments.

3.5. Cross-Database Validation

In order to test the generalization ability of our network, we train the model on the entire KonIQ-10k and test on the entire LIVEC. The four most competing IQA models in terms of generalization ability are involved in the comparison, which are PQR (Zeng et al., 2017), DBCNN, HyperIQA, and DeepFL. The validation results are shown in **Table 3**, indicating the generalization ability of our approach is higher than existing SOTA methods for assessing authentically distorted images.

However, if the network is trained on KonIQ-10k and directly applied for a synthetic image database, its generalization ability is not satisfactory, and the SRCC on TID2013 is only 0.577. That is mainly because the distortion mechanisms between synthetic and authentically distorted image databases are widely different. Training the network solely on authentically

TABLE 3 | Cross data base validation (Trained on KonIQ-10k and Tested on LIVEC).

Modules	DeepFL	DBCNN	HyperIQA	PQR	Ours
SRCC	0.704	0.755	0.770	0.785	0.817
Gain(%)	–	7.2	9.4	11.5	16.1

TABLE 4 | SRCC and PLCC performance on CCT, DHQ, and SHRQ.

		SRCC	PLCC
CCT	20-%Test	0.9655	0.9672
	100-%Test	0.5758	0.6193
DHQ	20-%Test	0.9533	0.9223
	100-%Test	0.6819	0.6678
SHRQ	20-%Test	0.8875	0.9082
	100-%Test	0.4233	0.4761

-distorted image databases could not learn the specific synthetic distortion patterns such as JPEG compression, transmission errors, or degradation caused by denoising, *etc*.

3.6. Further Validation on Other Specific IQA Tasks

In order to further validate the robustness of our BIQA framework toward other specific IQA tasks, the performance of CDFS guided BIQA network is evaluated on CCT (Min et al., 2017b), DHQ (Min et al., 2018b), and SHRQ (Min et al., 2019). The CCT contains 1,320 distorted images with various types of images including natural scene images (NSI), computer graphic images (CGI), and screen content images (SCI); The DHQ contains 1,750 dehazed images generated from 250 real hazy images.; The SHRQ database consists of two subsets, namely: regular and aerial image subsets, which include 360 and 240 dehazed images created from 45 and 30 synthetic hazy images using 8 eight image dehazing algorithms, respectively.

The training pipeline is similar with section 3.1, i.e., 80% of the CCT, DHQ, or SHRQ are involved as the training set and the other 20% is the testing set. Considering that the scale of the subset is not adequate for the training of DNN, we merge the subsets in each datasets. For example, the NSI, CGI, and SCI are merged as the training set of CCT.

As shown in **Table 4**, the predictions of our CDFS guided BIQA framework shows significant consistency with subjective scores, indicating that our proposed BIQA approach is feasible to be generalized into other types of IQA tasks.

Furthermore, if the network is trained on KonIQ-10k and directly applied on CCT, DHQ, and SHRQ, the accuracy is not satisfactory, as shown in **Table 4**. Such phenomenon is similar to the cross-database validation results discussed in section 3.5, indicating that training the network solely on authentically-distorted natural image databases could not sufficiently learn the quality-aware features for CGI, SCI, etc.

4. CONCLUSION

This work aims to evaluate the perceptual quality based on cross-domain feature similarity. The experimental results on KonIQ, LIVEC, and TID2013 demonstrate the superiority of our proposed methods.

We would further investigate such CDFS-incorporated BIQA framework in the following aspects: (1) investigating more efficient approaches of CDFS measurement; (2) investigating more types of DNN baselines in addition to ResNet.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CF established the BIQA framework and adjusted the architecture for better performance. LY and CF conducted the experiments and wrote the manuscripts. QZ designed the original method, and provided resource support (e.g., GPUs) for this manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work is supported by the National Key R&D Program of China under Grant No. 2021YFF0900503, the National Natural Science Foundation of China under Grant Nos. 61971383 and 61631016, and the Fundamental Research Funds for the Central Universities.

ACKNOWLEDGMENTS

We would like to thank Li Fang, Wei Zhong, and Fei Hu for some swell ideas.

REFERENCES

Casser, V., Pirk, S., Mahjourian, R., and Angelova, A. (2019). Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos. *Proc. AAAI Conf. Artif. Intell.* 33, 8001–8008. doi: 10.1609/aaai.v33i01.33018001

Chang, H.-W., Yang, H., Gan, Y., and Wang, M.-H. (2013). Sparse feature fidelity for perceptual image quality assessment. *IEEE Trans. Image Proc.* 22, 4007–4018. doi: 10.1109/TIP.2013.2266579

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 248–255.

Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). "RMPE: regional multi-person pose estimation," IN *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2334–2343.

Ghadiyaram, D., and Bovik, A. C. (2015). Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Proc*. 25, 372–387. doi: 10.1109/TIP.2015.2500021

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. (2019). Dialoguegcn: a graph convolutional neural network for emotion recognition in conversation. *arXiv preprint* arXiv:1908.11540. doi: 10.18653/v1/D19-1015

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

Hosu, V., Lin, H., Sziranyi, T., and Saupe, D. (2020). Koniq-10k: an ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Proc*. 29, 4041–4056. doi: 10.1109/TIP.2020.2967829

Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 1733–1740.

Kim, J., and Lee, S. (2016). Fully deep blind image quality predictor. *IEEE J. Sel. Top. Signal Process*. 11, 206–220. doi: 10.1109/JSTSP.2016.2639328

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint* arXiv:1412.6980.

Larson, E. C., and Chandler, D. M. (2010a). Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* 19, 011006.

Larson, E. C., and Chandler, D. M. (2010b). Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* 19, 011006. doi: 10.1117/1.3267105

Li, D., Jiang, T., Lin, W., and Jiang, M. (2018). Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Trans. Multimedia* 21, 1221–1234. doi: 10.1109/TMM.2018.2875354

Li, S., Zhang, F., Ma, L., and Ngan, K. N. (2011). Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Trans. Multimedia* 13, 935–949. doi: 10.1109/TMM.2011.2152382

Li, Y., Po, L.-M., Feng, L., and Yuan, F. (2016). "No-reference image quality assessment with deep convolutional neural networks," in *2016 IEEE International Conference on Digital Signal Processing (DSP)* (Beijing: IEEE), 685–689.

Lin, K.-Y., and Wang, G. (2018). "Hallucinated-iqa: No-reference image quality assessment via adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City: IEEE), 732–741.

Liu, T.-J., Lin, W., and Kuo, C.-C. J. (2012). Image quality assessment using multi-method fusion. *IEEE Trans. Image Proc*. 22, 1793–1807. doi: 10.1109/TIP.2012.2236343

Liu, X., Van De Weijer, J., and Bagdanov, A. D. (2017). "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 1040–1049.

Ma, K., Liu, W., Liu, T., Wang, Z., and Tao, D. (2017a). dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Trans. Image Proc*. 26, 3951–3964. doi: 10.1109/TIP.2017.2708503

Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., and Zuo, W. (2017b). End-to-end blind image quality assessment using deep neural networks. *IEEE Trans. Image Proc*. 27, 1202–1213. doi: 10.1109/TIP.2017.2774045

Min, X., Gu, K., Zhai, G., Liu, J., Yang, X., and Chen, C. W. (2017a). Blind quality assessment based on pseudo-reference image. *IEEE Trans. Multimedia* 20, 2049–2062. doi: 10.1109/TMM.2017.2788206

Min, X., Ma, K., Gu, K., Zhai, G., Wang, Z., and Lin, W. (2017b). Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Trans. Image Proc*. 26, 5462–5474. doi: 10.1109/TIP.2017.2735192

Min, X., Zhai, G., Gu, K., Liu, Y., and Yang, X. (2018a). Blind image quality estimation via distortion aggravation. *IEEE Trans. Broadcast*. 64, 508–517. doi: 10.1109/TBC.2018.2816783

Min, X., Zhai, G., Gu, K., Yang, X., and Guan, X. (2018b). Objective quality evaluation of dehazed images. *IEEE Trans. Intell. Transport. Syst*. 20, 2879–2892. doi: 10.1109/TITS.2018.2868771

Min, X., Zhai, G., Gu, K., Zhu, Y., Zhou, J., Guo, G., et al. (2019). Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Trans. Multimedia* 21, 2319–2333. doi: 10.1109/TMM.2019.2902097

Min, X., Zhai, G., Zhou, J., Farias, M. C., and Bovik, A. C. (2020a). Study of subjective and objective quality assessment of audio-visual signals. *IEEE Trans. Image Proc*. 29, 6054–6068. doi: 10.1109/TIP.2020.2988148

Min, X., Zhou, J., Zhai, G., Le Callet, P., Yang, X., and Guan, X. (2020b). A metric for light field reconstruction, compression, and display quality evaluation. *IEEE Trans. Image Proc*. 29:3790–3804. doi: 10.1109/TIP.2020.2966081

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Proc*. 21, 4695–4708. doi: 10.1109/TIP.2012.2214050

Pan, D., Shi, P., Hou, M., Ying, Z., Fu, S., and Zhang, Y. (2018). "Blind predicting similar quality map for image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 6373–6382.

Park, S.-J., Son, H., Cho, S., Hong, K.-S., and Lee, S. (2018). "Srfeat: single image super-resolution with feature discrimination," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 439–455.

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., et al. (2015). Image database tid2013: Peculiarities, results and perspectives. *Signal Proc. Image Commun*. 30, 57–77. doi: 10.1016/j.image.2014.10.009

Rehman, A., and Wang, Z. (2012). Reduced-reference image quality assessment by structural similarity estimation. *IEEE Trans. Image Proc*. 21, 3378–3389. doi: 10.1109/TIP.2012.2197011

Sheikh, H. R., and Bovik, A. C. (2006). Image information and visual quality. *IEEE Trans. Image Proc*. 15, 430–444. doi: 10.1109/TIP.2005.859378

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., et al. (2020). "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 3667–3676.

Sun, W., Min, X., Zhai, G., and Ma, S. (2021). Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *arXiv preprint* arXiv:2105.14550.

Talebi, H., and Milanfar, P. (2018). Nima: Neural image assessment. *IEEE Trans. Image Proc*. 27, 3998–4011. doi: 10.1109/TIP.2018.2831899

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). "A survey on deep transfer learning," in *International Conference on Artificial Neural Networks* (Rhodes: Springer), 270–279.

Wang, Z., and Bovik, A. C. (2011). Reduced-and no-reference image quality assessment. *IEEE Signal Process Mag*. 28, 29–40. doi: 10.1109/MSP.2011.942471

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Proc*. 13, 600–612. doi: 10.1109/TIP.2003.819861

Wang, Z., and Simoncelli, E. P. (2005). Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. *Hum. Vision Electron. Imaging* 5666, 149–159. doi: 10.1117/12.597306

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2* (Pacific Grove, CA: IEEE), 1398–1402.

Xiongkuo, M., Ke, G., Guangtao, Z., Xiaokang, Y., Wenjun, Z., Callet, P. L., et al. (2021). *Screen Content Quality Assessment: Overview, Benchmark, and Beyond*. ACM Computing Surveys.

Xu, J., Ye, P., Li, Q., Du, H., Liu, Y., and Doermann, D. (2016). Blind image quality assessment based on high order statistics aggregation. *IEEE Trans. Image Proc*. 25, 4444–4457. doi: 10.1109/TIP.2016.2585880

Xue, W., Zhang, L., Mou, X., and Bovik, A. C. (2013). Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Trans. Image Proc*. 23, 684–695. doi: 10.1109/TIP.2013.2293423

Yang, S., Jiang, Q., Lin, W., and Wang, Y. (2019). "Sgdnet: an end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proceedings of the 27th ACM International Conference on Multimedia* (Nice), 1383–1391.

Zeng, H., Zhang, L., and Bovik, A. C. (2017). A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint* arXiv:1708.08190.

Zhai, G., and Min, X. (2020). Perceptual image quality assessment: a survey. *Sci. China Inf. Sci*. 63, 211301. doi: 10.1007/s11432-019-2757-1

Zhang, L., Shen, Y., and Li, H. (2014). Vsi: a visual saliency-induced index for perceptual image quality assessment. *IEEE Trans. Image Proc*. 23, 4270–4281. doi: 10.1109/TIP.2014.2346028

Zhang, L., Zhang, L., and Bovik, A. C. (2015). A feature-enriched completely blind image quality evaluator. *IEEE Trans. Image Proc*. 24, 2579–2591. doi: 10.1109/TIP.2015.2426416

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). Fsim: a feature similarity index for image quality assessment. *IEEE Trans. Image Proc*. 20, 2378–2386. doi: 10.1109/TIP.2011.2109730

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018a). "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 586–595.

Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. (2018b). Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circ. Syst. Video Technol*. 30, 36–47. doi: 10.1109/TCSVT.2018.2886771

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.