Teacher-Guided Learning for Blind Image Quality Assessment

Zewen Chen[1,2], Juan Wang[1], Bing Li[1]✉, Chunfeng Yuan[1], Weihua Xiong[4], Rui Cheng[4], and Weiming Hu[1,2,3]

[1] NLPR, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
[4] Zeku Technology (Shanghai) Corp
{chenzewen2022,jun_wang}@ia.ac.cn, {bli,cfyuan,wmhu}@nlpr.ia.ac.cn,
wallace.xiong@gmail.com, chengrui@zeku.com

Abstract. The performance of deep learning models for blind image quality assessment (BIQA) suffers from annotated data insufficiency. However, image restoration, as a closely-related task with BIQA, can easily acquire training data without annotation. Moreover, both image semantic and distortion information are vital knowledge for the two tasks to predict and improve image quality. Inspired by these, this paper proposes a novel BIQA framework, which builds an image restoration model as a teacher network (TN) to learn the two aspects of knowledge and then guides the student network (SN) for BIQA. In TN, multi-branch convolutions are leveraged for performing adaptive restoration from diversely distorted images to strengthen the knowledge learning. Then the knowledge is transferred to the SN and progressively aggregated by computing long-distance responses to improve BIQA on small annotated data. Experimental results show that our method outperforms many state-of-the-arts on both synthetic and authentic datasets. Besides, the generalization, robustness and effectiveness of our method are fully validated. The code is available in https://github.com/chencn2020/TeacherIQA.

Keywords: blind image quality assessment · image restoration · prior knowledge

1 Introduction

Image quality assessment (IQA) has been an active topic in image processing. Numerous applications, such as unmanned aerial vehicle, surveillance *et al.*, rise an urgent demand for IQA. Compared with full-reference and reduced-reference IQA, blind IQA (BIQA) receive more attention for removing the dependence on reference images which are even impossible to obtain in real-world applications.

In BIQA methods, the central idea is to extract features from images and map them to an IQA score. Traditional methods rely on handcrafted features

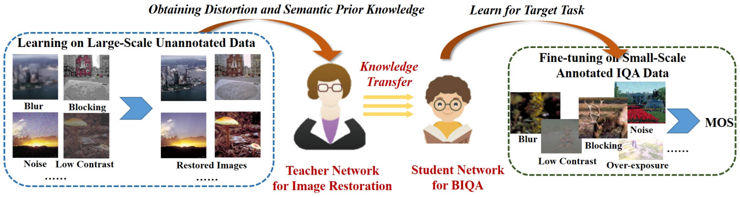

Obtaining Distortion and Semantic Prior Knowledge
Learn for Target Task
Learning on Large-Scale Unannotated Data
Blur
Blocking
Noise
Low Contrast
......
Restored Images
......
Knowledge Transfer
Teacher Network for Image Restoration
Student Network for BIQA
Fine-tuning on Small-Scale Annotated IQA Data
Blur
Low Contrast
Blocking
Noise
Over-exposure
......
MOS

Fig. 1. Workflow of the proposed teacher-guided learning framework for BIQA, which consists of prior knowledge learning and target task learning.

to construct BIQA models, which can be divided into nature scene statistic (NSS)-based models [35,29,28] and human visual system (HVS)-related models [19,44,46]. In recent years, Convolutional Neural Networks (CNNs) become the hotspot of various research fields. The performance of BIQA methods has also been greatly advanced by CNNs. Early CNNs-based BIQA methods [6,33] typically adopt shallow networks to extract low-level features, such as edges, textures and color, which explicitly reflect the distortion information.

However, training a successful deep CNN highly relies on an avalanche of annotated data. For IQA dataset, the labels, *e.g.*, mean opinion scores (MOSs), are obtained through psychophysical experiments. These experiments are expensive and time-consuming, making the acquisition of large-scale annotated data challenging. For this problem, one of seemingly plausible solutions is to combine multiple IQA datasets for training [52]. Unfortunately, IQA datasets have different perceptual scales due to the difference in psychophysical experiments. Using multi-scaled MOSs for training can lead to suboptimal performance. In addition, pre-training has been assumed to be an effective approach to address the lack of training data. Generally, the networks are pre-trained on other datasets or tasks to learn prior knowledge, and then fine-tuned for the target task. For example, some methods [22,25] pre-train the BIQA models to learn the quality rank on vast generated training samples first. Then they fine-tune the models to learn the quality score on standard IQA datasets. However, the models can only learn the perceptual scale of the same distortion type during the quality rank stage. Besides, the pre-trained image classification networks are also popularly used as the feature extractors in many BIQA models [39,18,48,42]. However, recent research [54] has shown that these networks are less adaptable to the BIQA task, since the classification task seldom considers the distortion information.

In order to overcome the bottleneck of insufficient annotated data, this paper proposes a new teacher-guided learning framework for BIQA to obtain the knowledge about image distortions and semantics from a large collection of unannotated data. The workflow of our framework is shown in Fig. 1, where an image restoration model is built as a teacher network (TN) to learn the two aspects of knowledge and then it guides the student network (SN) for the BIQA training on small annotated data. In TN, multi-branch convolutions (MC) are used for capturing fine-grained semantic and distortion information to achieve adaptive restoration and strengthen the knowledge learning. Then the learned knowledge is transferred from the TN to the SN through two paths. In SN, the transferred knowledge is progressively aggregated by computing long-distance responses and finally merged to a global quality score.

The motivation of these methods [20,23,34,31], which also employ image restoration as the auxiliary task, is to leverage the image restoration model to restore reference images to compare with distorted images. These methods are sensitive to the quality of restored images. However, our framework resorts to image restoration for feature learning since the image restoration and IQA tasks share similar knowledge. Our contributions are summarized as follows:

- We propose a new teacher-guided learning framework for BIQA, where a TN is presented to learn prior knowledge about image semantics and distortions from large unannotated data. By inheriting the knowledge from the TN, the SN can learn BIQA more efficiently by only using a small amount of annotated data.
- A multi-branch convolution is presented for the image restorer to capture the fine-grained features to achieve adaptive restoration for different types of distortions, and an attention mechanism is developed for the image quality predictor to aggregate the transferred knowledge for score estimate. Both network designs significantly improve the prediction accuracy for BIQA.
- Experimental results show that our method achieves state-of-the-art performance on both synthetic and authentic datasets. Besides, the generalization, robustness and effectiveness of our model are validated by cross-dataset evaluations, small training data experiments, ablation studies and group maximum differentiation (gMAD) competition.

2 Related Work

For the context of our work, we briefly review related work on blind image quality assessment and learning methods for insufficient data.

2.1 Blind Image Quality Assessment

In the early stage, researchers found out statistical characteristics vary when images are corrupted by different distortions. Thus, a quantity of NSS-based BIQA models were proposed. Saad *et al.* [35] and Moorthy *et al.* [29] respectively utilize the NSS model of discrete cosine transform and discrete wavelet transform coefficients to construct BIQA models. Mittal *et al.* [28] propose to utilize the NSS of locally normalized luminance coefficients to quantify quality scores. Other works attempt to extract HVS-related features, such as NRSL [19], LPSI [44] and M3 [46]. However, when it comes to complex and mixed distortions, the performance of the handcrafted-based BIQA models is far from satisfactory.

In recent years, benefit from the powerful representation ability of CNNs, BIQA methods have achieved impressive results. Kang *et al.* [12] are the pioneers to use the CNN to predict the quality score. They also propose a multi-task CNN [13], which predicts the quality score and distortion type simultaneously. These early CNNs-based BIQA methods adopt shallow networks to prevent the overfitting problem due to the lack of sufficient annotated data. To break through this limitation, some methods [1,39,48,42] adopt pre-trained networks (*eg.* ResNet [10] and VGG [38]), which are pre-trained on large-scale datasets like ImageNet [4], as the feature extractor. Benefit from the prior knowledge about image semantic information, these BIQA methods achieve a great progress on authentic IQA datasets. However, when it comes to synthetic IQA databases, the performance is far from satisfactory. This is because the pre-trained tasks seldom consider the image distortion information. To make BIQA models more aware of the distortions, some methods combine the image semantics with the distortions. For example, Zhang *et al.* [51] propose a DB-CNN model, where two networks respectively pre-trained for image classification and distortion classification are used as the feature extractor. Zhu *et al.* [54] propose a MetaIQA model. They adopt the meta learning approach to learn a prior knowledge model of various distortions, and then fine-tune the prior model with unknown distortions.

2.2 Learning Methods for Insufficient Data

Training a successful deep CNN largely relies on *supervised learning* that requires a huge number of annotations, which are expensive to obtain. Learning prior knowledge from other datasets or tasks has proven to be effective for improving the performance of target tasks in which the annotated data is not enough. As a result, a series of variant supervised learning methods are born. For example, in *semi-supervised learning* methods [3,47], the model is trained on a fraction of the dataset that is annotated manually first. Then the trained model is used to predict the remaining portion of the unannotated dataset. At last, the model is trained on the full dataset comprising of manually annotated and pseudo annotated data. In the *weakly-supervised learning* method [27], the recognition model is pre-trained with billions of Instagram images with noisy hashtags. Then the model is fine-tuned on annotated ImageNet dataset. For *self-supervised learning* methods, supervisory signals of the partial input is used to learn a better representation of the input. Generally, this is done via a pretext task that applies a transformation to the input image of the target task. The pretext tasks include image colorization [17], orientation [8] and counting visual primitives [30], *etc.* Though the learnt prior knowledge is effective in improving the performance of the target task, Shen *et al.* [37] and He *et al.* [9] demonstrate this benefit is reduced when the pre-training data belongs to a completely different domain.

In this paper, we resort to self-supervised learning (SSL) to address insufficient annotated data. We use image restoration as the pretext task to learn prior knowledge for BIQA. Compared with semi-supervised and weakly-supervised learning, SSL does not require any annotated data. Compared with other pretext tasks, image restoration can provide more related knowledge for BIQA.

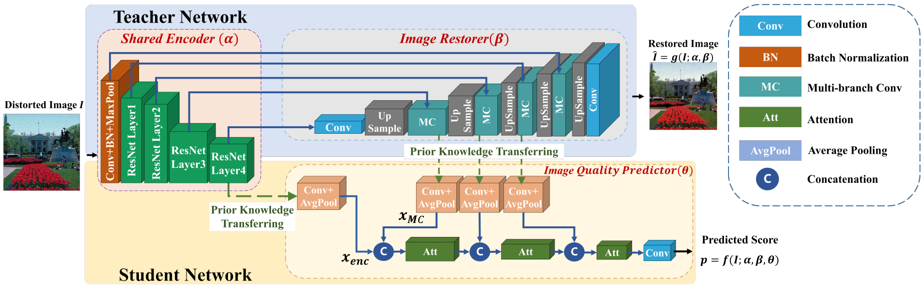


Fig. 2. Overview of our BIQA model, which consists of a teacher network (TN) and a student network (SN)

3 Proposed Method

Fig. 2 shows the overview of our framework, which consists of two networks: a teacher network (TN) and a student network (SN). The TN and the SN share a same encoder. In addition, the TN also includes an image restorer, while the SN includes an image quality predictor. The training of our framework is divided into two phases. In the first phase, the TN is trained to learn the knowledge about the semantics and distortions from the image restoration task. In the second phase, the SN inherits the prior knowledge from the TN to learn BIQA on the IQA datasets. In the following, we introduce our method in details.

3.1 Teacher Network Learning from Image Restoration

Pretext Task. Both image semantics and distortions are vital knowledge for BIQA. The purpose of TN is to obtain the two knowledge to guide BIQA learning. Although pre-trained classification networks are equipped with strong semantic perceptual ability, recent research has shown that these networks are less adaptable to BIQA. This can be attributed to that the classification task pays more attention to high-level semantics, which are less sensitive to distortions. In contrast, image restoration requires both high-level and low-level features. On the one hand, low-level details explicitly reflect the distortion type and level. On the other hand, high-level semantics help to infer the distortion information (*eg.*, semantic information helps to judge the smooth area is clean sky or blurry jeans). Both image restoration and BIQA rely on the two aspects of knowledge. From this perspective, the knowledge of the two tasks can be shared. We employ image restoration as the objective task of the TN training.

Distortion-Aware Image Restoration. The TN consists of a shared encoder and an image restorer parameterized by α and β. Given an input image I, the TN aims to recover a high-quality image $\hat{I}$ from I, which is denoted as $\hat{I} = g(I;\alpha,\beta)$. For the encoder, it should have powerful feature representation ability. To this end, we adopt ResNet-50 [10] as the encoder, since it has proven to be an excellent feature extractor by many computer vision tasks. For the image restorer, its responsibility is to infer clean image content from encoder features. However, because of the limited distortion perception and detail synthesis ability of the model, most existing image restoration models only perform well for a specific distortion. For this reason, many restoration models adopt complicated networks, such as multi-level wavelet CNNs [21], scale recurrent network [40] and residual channel attention CNNs [53]. Since image restoration is only used as an auxiliary task in this paper, we do not aim to design a cumbersome network to accurately restore images. We employ a simple effective multi-branch convolutions (MCs), which consist of multiple parallel paths with different sizes of convolution kernels, to deal with various distortions. We first add skip connections to transmit the low-level features of the encoder to the decoder to strengthen the distortion perception ability. Then, the MCs are used to extract the context features and synthesize fine-grained details in multi-resolution receptive fields. The operation can be formulated as follows:

$$\begin{aligned} &f_i = f_l \oplus f_h, \qquad f_o = B_5\left(B_1\left(f_i\right) \oplus B_2\left(f_i\right) \oplus B_3\left(f_i\right) \oplus B_4\left(f_i\right)\right), \\ &B_i \subseteq \{1\times 1\,\mathrm{conv}, 3\times 3\,\mathrm{conv}, 5\times 5\,\mathrm{conv}, \mathrm{maxPool}\}, \qquad (1 \le i \le 5) \end{aligned} \tag{1}$$

where f_l and f_h denote features of the encoder and decoder, respectively, the symbol $\oplus$ denotes concatenation, and B_i $(1 \le i \le 5)$ denotes the i-th branch[5].

Loss Function. The first phase of our method is to train the TN for image restoration. We combine three losses, including reconstruction loss $\mathcal{L}_{\mathrm{rec}}$, structure loss $\mathcal{L}_{\mathrm{stru}}$ and perceptual loss $\mathcal{L}_{\mathrm{percept}}$, to promote the consistency of the ground truth images I and the restored image $\hat{I}$ in pixel domain, low-level and high-level feature domains, respectively. The loss function is formulated as follows:

$$\begin{aligned} \mathcal{L}_{\mathrm{TN}}(I, \hat{I}; \alpha, \beta) &= \frac{1}{N}\sum_{n=1}^{N}[\varepsilon\mathcal{L}_{\mathrm{rec}}(I,\hat{I}) + \rho\mathcal{L}_{\mathrm{stru}}(I,\hat{I}) + \mu\mathcal{L}_{\mathrm{percept}}(I,\hat{I})], \\ &= \frac{1}{N}\sum_{n=1}^{N}[\varepsilon||I-\hat{I}||_1 + \rho(1-\mathrm{SSIM}(I,\hat{I})) + \mu\sum_{t=1}^{T}\frac{1}{\Theta_t}||\psi_t(I)-\psi_t(\hat{I})||_1], \end{aligned} \tag{2}$$

where the hyper-parameters ε, ρ and μ balance their trade-off, and each loss is normalized by the batch size N. In the second loss, SSIM denotes the Structural Similarity Index [43]. In the third loss, $\psi_t(x)$ denote the t-th layer output of the pretrained VGG-19 network [38] for input x, T is the total number of layers used to calculate $\mathcal{L}_{\mathrm{percept}}$, and Θ_t denote the number of elements in the t-th layer output. Concretely, we extract the 1st-5th pooling layer outputs of the VGG-19.

3.2 Student Network Learning for BIQA

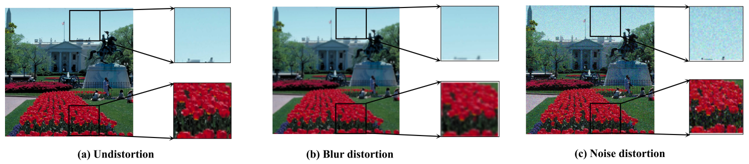
Given an input image I, our BIQA model aims to infer its quality score p, which is modeled as $p = f(I; \Phi)$, where $\Phi = \{\alpha, \beta, \theta\}$ denote the network parameters. Recall that α and β are the parameters of the encoder and the restorer, respectively, which have been well pre-trained on a large collection of distorted and

[5] More details about the architecture are provided in the supplementary material.

reference image pairs. While θ is the parameter of the image quality predictor in the SN. The objective of our method in the second phase is to learn θ and fine-tune α and β on the IQA dataset. Benefit from the prior knowledge in α and β, our model can learn Φ more effectively and efficiently only using a small amount of annotated IQA data compared with those which learn Φ from scratch. In the following, we introduce the prior knowledge guided learning in detail.

Prior Knowledge Transferring. During the second phase training, the prior knowledge learned in the first phase is transferred to the image quality predictor from two paths. As shown in Fig. 2, one path is from the encoder, where the features are extracted from its bottleneck layer, denoted as f_{enc}. Another path is from the image restorer. Recall that the MCs are adopted to capture fine-grained context features. Here, we extract its output to the SN, denoted as f_{MC}. The introduction of f_{MC} can effectively reinforce the information lacked in f_{enc} due to the pooling operations in the encoder. In the experimental section, we validate that both the prior knowledge transferring paths and the MCs promote the BIQA performance.

Global Knowledge Aggregation. Given the prior knowledge, the objective of the SN is to extract useful information from the knowledge and map it to the image quality score. It is worth noting that different semantic regions show different responses for the same distortion type and level. As shown in Fig. 3, when the image suffers from the uniformly distributed blur distortion, the textured regions (*e.g.,* flowering shrubs) are affected more seriously than the smooth regions (*e.g.,* sky). When the image suffers from the noise distortion, the responses of these regions are converse. Motivated by this fact, we propose to capture long-distance dependencies for merging the distortion level of each local regions to a global quality score. To achieve this goal, we adopt the self-attention module [41], where three projections are learned to compute distant responses by matrix multiplication. In our framework, f_{enc} is used as the basic image features of the SN, and f_{MC} output from the 1st-3rd MCs are successively transferred to concatenate with f_{enc}. The concatenated features are aggregated using the self-attention module to convert the local features to global quality-related features. By three feature aggregation operations, we obtain a 7×7 feature map (for a 224×224 input). Finally, we adopt three 1×1 convolutions to reduce the channel number and a 7×7 convolution to map the feature map to a quality score.


(a) Undistortion

(b) Blur distortion

(c) Noise distortion

Fig. 3. Different semantic regions show different responses for the same distortion. Smooth regions are sensitive to noise, while textured regions are vulnerable to blur.

Loss function. Let $g = \{g_n \mid n \in [1, N]\}$ and $p = \{p_n \mid n \in [1, N]\}$ denote the ground truth and predicted quality scores of N images, respectively, and the subscript n denotes the n-th image. The optimization objective of the SN is defined as minimizing the ℓ_1 distance between g_n and p_n, that is $\mathcal{L}_{\mathrm{SN}}(p, g; \alpha, \beta, \theta) = \frac{1}{N} \sum_n^N |g_n - p_n|$. Based on the loss function, we leverage the ADAM optimizer to update the parameter θ and fine-tune the parameters α and β with a small learning rate at the same time.

4 Experiments

In this section, we introduce experimental datasets, implementations and evaluation metrics first. Then we compare our method with state-of-the-art BIQA methods. Next, ablation studies are presented. Finally, we make a visual analysis and gMAD competitions to compare the robustness of the model.

4.1 Datasets

We trained and evaluated our model on four synthetic IQA datasets, including LIVE [36], CSIQ [16], TID2013 [32] and WED [24], and two authentic datasets, including LIVEC [6] and KonIQ-10K [11]. The LIVE consists of 779 distorted images by adding 5 distortion types on 29 reference images. The CSIQ possesses 866 distorted images derived from 30 reference images and 6 synthetic distortion types. The TID2013 contains 3,000 distortion images generated from 25 high-quality images, 24 distortion types and 5 levels. The WED contains 4,744 pristine natural images, and 94,880 distorted images are created from them with 4 distortion types and 5 levels. The LIVEC and KonIQ-10k contain 1,162 and 10,073 images, respectively, derived from the real world. These images have more complex distortions, making BIQA on the two datasets more challenging.

4.2 Implementations

The First Phase. The training of the TN for image restoration does not require any annotated data. We collect massive images from publicly available datasets. Specifically, We used 4,744 high-quality images of the WED dataset as reference images. Following [23,51,22], we manually added 16 out of a total of 24 distortions (including types of 1,2,5-10,14-19,22,23) defined by the TID2013 [32] and 4 levels (including levels of 1-4) on the reference images. We did not add other distortion types to evaluate the generalization ability of our model and we did not use the level 5, which are seriously distorted, leading the model difficult to converge. Consequently, we created 303,616 distorted images. Each distorted image and its corresponding reference image constitute an image pair. Besides, each image is randomly flipped and cropped into 20 patches with the size of 224×224. The TN is trained for 10 epochs with a learning rate of 5×10^{-5} and the batch size N is set to 80. The hyper-parameters of the three loss functions defined in (2) are set to $\varepsilon = 1.0$, $\rho = 0.08$ and $\mu = 1.0$.

The Second Phase. the SN is trained on individual IQA dataset for BIQA. Following existing methods [39,20,51], we randomly flip and crop the input image into 25 patches with the size of 224×224. Each patch keeps the same quality score as the source image. Each dataset is randomly divided into 80% for training and 20% for testing. To ensure that there is no overlapping images between the training and the testing set on synthetic datasets, the dataset is divided according to the reference image. For authentic datasets, there are no reference images, so we divide the dataset according to the distorted image. All the experiments on each dataset are conducted 10 times repeatedly and we choose the model with the lowest validation error. The final result is the median value of the 10 scores. We adopt the Adam optimizer with a learning rate of 6×10^{-6}. The model is optimized for 16 epochs with a batch size of 92. All the experiments are conducted on Pytorch with NVIDIA 3090 GPUs.

Evaluation Metrics. We utilize two metrics to measure the performance of BIQA model: Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SROCC). PLCC measures the linear correlation between the ground truth and predicted scores. SROCC measures the monotonicity between them. Both metrics range from -1 to 1. A higher value indicates a higher performance of the model.

4.3 Comparisons With State-of-the-Arts

A. Comparison Within Individual Datasets

We compare our model with four traditional BIQA, including BRISQUE [28], ILNIQE [50], HOSA [45] and FRIQUEE [7], and eleven CNNs-based BIQA, including BIECON [14], WaDIQaM-NR [2], PQR [49], RankIQA [22], DIQA [15],Hall-IQA [20], DB-CNN [51], MetaIQA [54], HyperIQA [39], AIGQA [23] and VCRNet [31] . The comparisons on five datasets are shown in Tab. 1, where the two highest scores are marked in black bold and blue bold respectively. Compared with 4 traditional models, our model achieves the best performance on all datasets. Compared with 11 CNNs-based models, our model achieves competitive results on 4 datasets. On the LIVE, our model obtains a slight lower performance, but it still achieves acceptable results with SROCC of 0.962 and PLCC of 0.965. We find that many of the compared models only perform well for authentic or synthetic datasets. For example, HyperIQA achieves excellent scores on authentic datasets, while its performance on synthetic datasets significantly degrades. By contrast, our model reports prominent scores on both datasets. In addition, the VCRNet [31] adopts a similar framework as our model, while its performance is not as excellent as ours. We attribute our advantage to the MC based image restorer and the self-attention based quality predictor, which effectively improve the performance of our model.

B. Comparison on Individual Distortions

In addition, we test the performance of our model on individual distortions. Tab. 2 lists SROCC on indivisual distortions on TID2013, and the best result for each distortion type is marked in bold. According to Tab. 2, our full model achieves the best performance on 9 out of 24 distortion types. Although our

Table 1. Performance comparison of BIQA methods on five IQA datasets

Dataset	LIVEC		KonIQ		TID		LIVE		CSIQ	
Model	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**	**PLCC**
BRISQUE [28]	0.608	0.629	0.665	0.681	0.651	0.573	0.939	0.935	0.746	0.829
ILNIQE [50]	0.432	0.508	0.507	0.523	0.519	0.640	0.902	0.865	0.806	0.808
HOSA [45]	0.640	0.678	0.671	0.694	0.688	0.764	0.946	0.947	0.741	0.823
FRIQUEE [7]	0.720	0.720	——	——	0.669	0.704	0.948	0.962	0.839	0.863
BIECON [14]	0.595	0.613	0.619	0.651	0.717	0.762	0.958	0.960	0.815	0.823
WaDIQaM [2]	0.671	0.680	0.797	0.805	0.787	0.761	0.954	0.963	——	——
PQR [49]	0.857	**0.882**	0.880	0.884	0.740	0.798	0.965	0.971	0.873	0.901
RankIQA [22]	——	——	——	——	0.780	0.799	**0.981**	**0.982**	**0.947**	**0.960**
DIQA [15]	0.703	0.704	——	——	0.825	0.850	0.975	0.977	0.884	0.915
Hall-IQA [20]	——	——	——	——	**0.879**	0.880	**0.982**	**0.982**	0.885	0.910
DB-CNN [51]	0.851	**0.869**	0.880	0.876	0.816	0.865	0.968	0.971	0.946	0.959
MetaIQA [54]	0.835	0.802	0.887	0.850	0.853	——	0.835	0.802	——	——
HyperIQA [39]	**0.859**	**0.882**	**0.906**	**0.917**	0.831	0.833	0.962	0.966	0.923	0.942
AIGQA [23]	0.751	0.761	——	——	0.871	**0.893**	0.960	0.957	0.927	0.952
VCRNet [31]	0.856	0.865	0.894	0.909	0.846	0.875	0.973	**0.974**	0.943	0.955
Ours	**0.861**	**0.882**	**0.910**	**0.916**	**0.920**	**0.932**	0.962	0.965	**0.950**	**0.961**

model does not perform best for the other 15 distortion types, it still surpasses most of the compared models. Moreover, our model obtains the highest average score (AVG) of 0.846 and the lowest standard deviation (STD) of 0.134. This indicates that our model has better stability for different types of distortion.

C. Cross-Dataset Evaluations

We further conduct cross-dataset experiments on three IQA datasets, including TID2013, CSIQ and LIVE. We use one dataset for training and the remaining two datasets for testing. Six state-of-the-art IQA models, including BRISQUE [28], FRIQUEE [7], HOSA [45], DB-CNN [51], HyperIQA [39] and VCRNet[31] are used for comparison. The evaluation results in terms of SROCC are shown in Tab. 3, where the two highest scores are marked in black bold and blue bold respectively. The TID2013 contains more distortion types and reference images than another two datasets. As shown from Tab. 3, most models achieve good performance when they are trained on TID2013 and tested on LIVE. On the contrary, when these models are trained on LIVE or CSIQ, their performance on TID2013 is much less satisfying. This indicates that training on the data with more diversities can effectively improve the generalization of the BIQA models. In addition, we observe that the models trained on CSIQ perform well on LIVE. However, the models trained on LIVE show a low performance on CSIQ. Nevertheless, our model ranks the top two among all the compared models. The result validates the good generalization ability of our model.

D. Comparison on Small Training Data

In this section, we conduct experiments to validate the proposed teacher-guided learning framework is effective in small data training. We randomly selected 20%, 40%, 60% and 80% images from the TID2013 and LIVEC dataset for training and 20% for testing. The four experiments were repeated 10 to overcome

Table 2. Performance comparison of individual distortions on TID2013 dataset in terms of SROCC. The bold distortion types are those used in TN learning

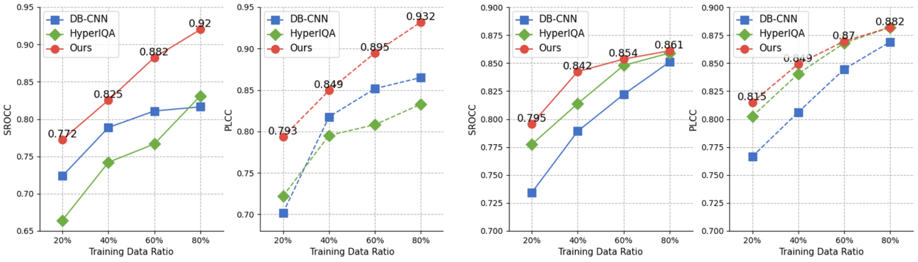
Type	BRISQUE	M3[46]	HOSA	RankIQA	Hall-IQA	DB-CNN	AIGQA	HyperIQA	Full	w/o TNL	w/o MC
#1	0.711	0.766	0.853	0.667	0.923	0.790	**0.932**	0.769	0.907	0.775	0.677
#2	0.432	0.560	0.625	0.620	0.880	0.700	**0.916**	0.613	0.855	0.557	0.490
#3	0.746	0.782	0.782	0.821	0.945	0.826	0.944	0.918	**0.967**	0.925	0.858
#4	0.252	0.577	0.368	0.365	0.673	0.646	0.662	0.448	**0.721**	0.203	0.191
#5	0.842	0.900	0.905	0.760	**0.955**	0.879	0.953	0.839	0.920	0.828	0.718
#6	0.765	0.738	0.775	0.736	0.810	0.708	**0.911**	0.758	0.906	0.736	0.653
#7	0.662	0.832	0.810	0.783	0.855	0.825	0.908	0.828	**0.909**	0.779	0.790
#8	0.871	0.896	0.892	0.809	0.832	0.859	0.917	0.873	**0.939**	0.848	0.808
#9	0.612	0.709	0.870	0.767	**0.957**	0.865	0.914	0.804	0.915	0.836	0.737
#10	0.764	0.844	0.893	0.866	0.914	0.894	**0.945**	0.860	0.913	0.852	0.777
#11	0.745	0.855	0.932	0.878	0.624	0.916	0.932	0.888	**0.950**	0.893	0.843
#12	0.301	0.375	0.747	0.704	0.460	0.772	0.858	0.723	**0.860**	0677	0.685
#13	0.748	0.718	0.701	0.810	0.782	0.773	**0.898**	0.846	0.881	0.817	0.821
#14	0.269	0.173	0.199	0.512	**0.664**	0.270	0.130	0.369	0.575	0.254	0.234
#15	0.207	0.379	0.327	0.622	0.122	0.444	**0.723**	0.428	0.598	0.440	0.418
#16	0.219	0.119	0.233	0.268	0.182	-0.009	**0.554**	0.424	0.434	0.404	0.032
#17	-0.001	0.155	0.294	0.613	0.376	0.548	**0.830**	0.740	0.779	0.617	0.478
#18	0.003	-0.199	0.119	0.662	0.156	0.631	0.689	0.710	**0.858**	0.676	0.395
#19	0.717	0.738	0.782	0.619	0.850	0.711	**0.948**	0.767	0.925	0.788	0.681
#20	0.196	0.353	0.532	0.644	0.614	0.752	**0.886**	0.786	0.855	0.692	0.618
#21	0.609	0.692	0.835	0.800	0.852	0.860	0.897	0.879	**0.938**	0.842	0.824
#22	0.831	0.908	0.855	0.779	**0.911**	0.833	0.908	0.785	0.878	0.740	0.702
#23	0.615	0.570	0.801	0.629	0.381	0.732	**0.889**	0.739	0.876	0.719	0.576
#24	0.807	0.893	0.905	0.859	0.616	0.902	0.908	0.910	**0.944**	0.903	0.873
AVG	0.538	0.597	0.668	0.691	0.681	0.714	0.836	0.738	**0.846**	0.700	0.620
STD	0.281	0.301	0.262	0.151	0.267	0.217	0.182	0.163	**0.134**	0.197	0.228

Table 3. Cross-dataset evaluation

Training	TID2013		CSIQ		LIVE	
Testing	LIVE	CSIQ	LIVE	TID2013	CSIQ	TID2013
BRISQUE	0.790	0.590	0.847	0.454	0.562	0.358
FRIQUEE	0.755	0.635	0.879	0.463	0.722	0.461
HOSA	0.846	0.612	0.773	0.329	0.594	0.361
DB-CNN	**0.891**	**0.807**	0.877	0.540	**0.758**	**0.524**
HyperIQA	0.834	0.686	0.848	0.481	0.707	0.504
VCRNet	0.822	0.721	**0.886**	**0.542**	**0.768**	0.502
Ours	**0.864**	**0.789**	**0.911**	**0.581**	0.738	**0.658**

Table 4. Ablation studies

Dataset	LIVEC		CSIQ		TID2013	
Methods	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
w/o path-1	0.853	0.871	**0.953**	**0.962**	0.876	0.894
w/o path-2	0.776	0.786	0.846	0.890	0.733	0.779
w/o TNL	0.854	0.875	0.921	0.942	0.764	0.801
w/o MC	0.799	0.821	0.887	0.895	0.692	0.759
w/o Att	0.850	0.871	0.941	0.955	0.865	0.890
Full Model	**0.861**	**0.882**	0.950	0.961	**0.920**	**0.932**



(1) TID 2013

(2) LIVEC

Fig. 4. Performance comparison of small data training on TID2013 and LIVEC.

the bias introduced by randomness. The median scores of SROCC and PLCC are reported. Two state-of-the-art IQA models (DB-CNN and HyperIQA) are used for comparison. The SROCC and PLCC curves with respect to the training data ratio on TID2013 and LIVEC datasets. As shown in Fig. 4, as the training data ratio increases, the performance of all models shows an upward trend. Moreover, our model always ranks the top one across all ratios on both datasets.

4.4 Ablation Studies

In order to validate contributions key components make to the proposed method, we train a series of variant models: i) *w/o path-1* and ii) *w/o path-2*, where the first and second prior knowledge path from the encoder and image restorer are removed, respectively; iii) *w/o TNL*, where teacher network learning (TNL) from image restoration is removed; iv) *w/o MC*, where the MC is replaced by three convolutions; v) *w/o Att*, where the self-attention is replaced by a convolution. All parameter settings are kept the same as the full model, as explained in Sec. 4.2. Experimental results on three datasets are presented in Tab. 4, where the highest scores are marked in bold.

i) For *w/o path-1* and *w/o path-2*, we can see that the removal of any path degrades the performance on LIVEC and TID2013, especially *w/o path-2*. This shows that the path 2 provides more useful information required by the IQA.

ii) For *w/o TNL*, we observe the two metrics show a significant decrease on TID2013 by 18.04% and 14.05%, respectively. We speculate that since the TN is pre-trained for learning the distortions defined by on TID2013, the learnt prior knowledge is more beneficial to the IQA performance on TID2013.

iii) For *w/o MC*, its performance shows a more obvious decease on TID2013, by 24.74% and 17.46% in terms of SROCC and PLCC, respectively. We speculate that this is because the TID2013 contains more various distortion types, which rises a higher requirement to the generalization ability of the model.

iv) For *w/o Att*, we observe a slight decrease on LIVEC and CSIQ but an obvious decrease on TID2013. We speculate that it is related to our pre-training samples, which are made by imposing distortions defined by TID2013. Consistent data distribution makes it easier to achieve knowledge aggregation.

4.5 The gMAD competition

To further evaluate the generalization of the proposed method, we conduct gMAD competition [26] on the SPAQ[5] dataset. There are two roles required: an attacker and a defender. The image pairs are selected when one model regards the image pairs with the same quality while the other regards them with different quality. If the image pairs are easy to distinguish, the attacker wins, otherwise, the defender wins. We choose two state-of-the-art BIQA methods HyperIQA [39] and DB-CNN [51] to compare with the proposed method.

As shown in Fig. 5, we can see that when our model is the defender (the leftmost four columns), there is no much perception difference in the image

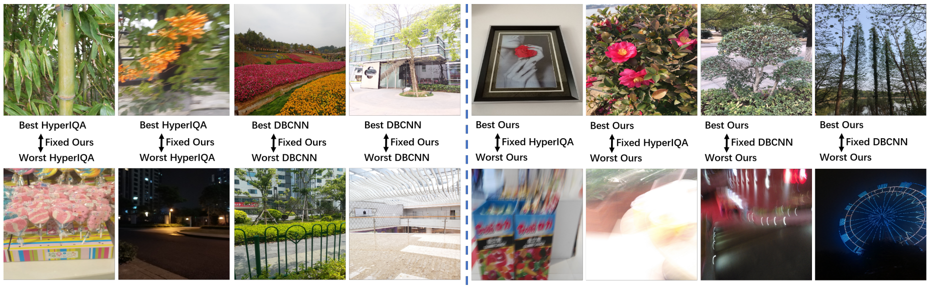


Fig. 5. The gMAD competition against HyperIQA[39] and DB-CNN[51] on SPAQ

pairs selected by the attacker. By contrast, when our model is the attacker (the rightmost four columns), it can easily select the image pairs with obvious quality differences, while the defenders regard these images with similar quality.

4.6 Analysis for the multi-branch convolutions

Benefit from the strengthened distortion information by the MC, the image restorer achieves adaptive image restoration for diverse distortions, which further improves the prediction accuracy of IQA scores. Both Tab. 2 and Tab. 4 have validated that removing MC degrades the IQA performance. In this section, we make a analysis on the MC and its effect on the image restoration.

Visual Analysis. Recall that f_{enc} and f_{MC} are the transferred knowledge from the TN to the SN. As their attention maps shown in Fig. 6, f_{enc} focuses more on the salient semantic regions, such as the persons in (b) and the elk in (d). In contrast, f_{MC} pays more attention on distorted regions, such as overexposure sky in (a), (b) and (c), noisy sky in (e) and motion blur in (f). The examples are consistent with our assumptions that f_{enc} has stronger semantic information while the distortion information is more prominent in f_{MC}.

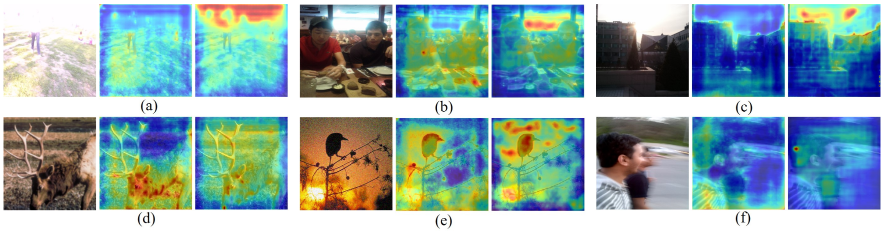


Fig. 6. Attention maps of the feature maps f_{enc} and f_{MC}. The images from left to right are respectively distorted images, attention maps of f_{enc} and f_{MC}.

Image Restoration Performance. In Fig. 7 (a), we qualitatively compare the restored results of our full model and the variant *w/o MC* for synthetic

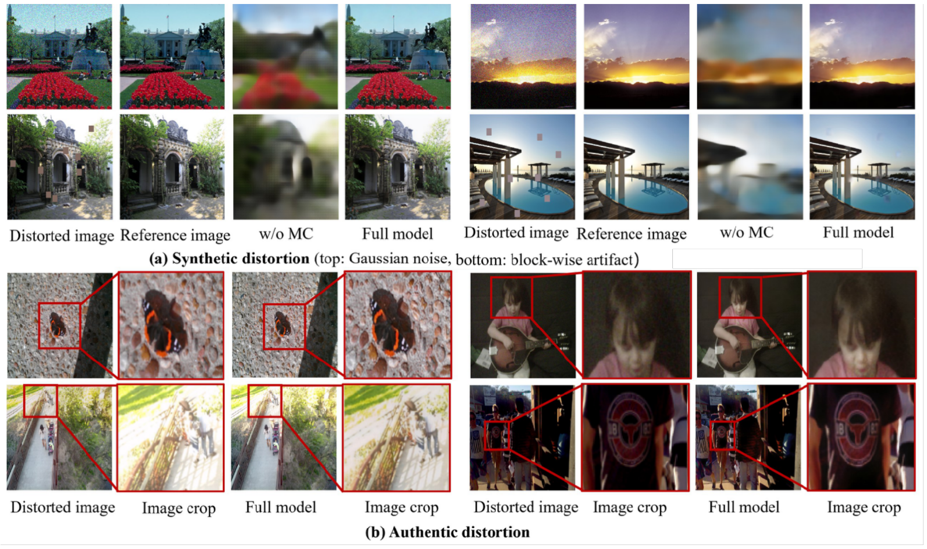


Fig. 7. Restored images by the TN of our variant model *w/o MC* and our full model for (a) synthetic distortion and (b) authentic distortion. (zoom in for more details)

distortions[6]. It can be seen that our full model restores visually pleasing images, while the results of the *w/o MC* suffer from serious blurry artifacts. Moreover, the results measured by PSNR and SSIM show that our full model achieves 28.22 and 0.881, while the *w/o MC* obtains 16.04 and 0.543. Both objective and subjective results validate the effectiveness of the MC in improving the quality of restored images. We also show restored images of our full model for authentic distortions, which are not presented in TN learning. As shown in Fig. 7 (b), our full model achieves excellent restoration for the authentic distorted images. Though authentic images are not pre-trained for image restoration, our model is still able to deal with them well, which validates the generalization of our model.

5 Conclusion

This paper proposes a new teacher-guided learning framework for BIQA to break the limitation of insufficient annotated data. In our framework, a multi-branch convolution based TN is presented to learn the prior knowledge from image restoration, and a SN is constructed to learn for BIQA by inheriting the prior knowledge from the TN using the attention mechanism. Experimental results show that our method surpasses many state-of-the-arts on both authentic and synthetic datasets. In addition, cross-dataset evaluations and gMAD competitions prove our method has a good generalization ability. Moreover, ablation studies validate the effectiveness of key components of our method.

[6] More restored results are shown in supplementary material.

References

1. Bianco, S., Celona, L., Napoletano, P., Schettini, R.: On the use of deep learning for blind image quality assessment. Signal, Image and Video Processing **12**(2), 355–362 (2018)
2. Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on image processing **27**(1), 206–219 (2017)
3. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks **20**(3), 542–542 (2009)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3677–3686 (2020)
6. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. IEEE Transactions on Image Processing **25**(1), 372–387 (2015)
7. Ghadiyaram, D., Bovik, A.C.: Perceptual quality prediction on authentically distorted images using a bag of features approach. Journal of vision **17**(1), 32–32 (2017)
8. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
9. He, K., Girshick, R., Dollár, P.: Rethinking imagenet pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4918–4927 (2019)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing **29**, 4041–4056 (2020)
12. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1733–1740 (2014)
13. Kang, L., Ye, P., Li, Y., Doermann, D.: Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: 2015 IEEE international conference on image processing (ICIP). pp. 2791–2795. IEEE (2015)
14. Kim, J., Lee, S.: Fully deep blind image quality predictor. IEEE Journal of selected topics in signal processing **11**(1), 206–220 (2016)
15. Kim, J., Nguyen, A.D., Lee, S.: Deep cnn-based blind image quality predictor. IEEE transactions on neural networks and learning systems **30**(1), 11–24 (2018)
16. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of electronic imaging **19**(1), 011006 (2010)
17. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6874–6883 (2017)

18. Li, D., Jiang, T., Lin, W., Jiang, M.: Which has better visual quality: The clear blue sky or a blurry animal? IEEE Transactions on Multimedia **21**(5), 1221–1234 (2018)
19. Li, Q., Lin, W., Xu, J., Fang, Y.: Blind image quality assessment using statistical structural and luminance features. IEEE Transactions on Multimedia **18**(12), 2457–2469 (2016)
20. Lin, K.Y., Wang, G.: Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 732–741 (2018)
21. Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-cnn for image restoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 773–782 (2018)
22. Liu, X., Van De Weijer, J., Bagdanov, A.D.: Rankiqa: Learning from rankings for no-reference image quality assessment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1040–1049 (2017)
23. Ma, J., Wu, J., Li, L., Dong, W., Xie, X., Shi, G., Lin, W.: Blind image quality assessment with active inference. IEEE Transactions on Image Processing **30**, 3650–3663 (2021)
24. Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., Zhang, L.: Waterloo exploration database: New challenges for image quality assessment models. IEEE Transactions on Image Processing **26**(2), 1004–1016 (2016)
25. Ma, K., Liu, W., Liu, T., Wang, Z., Tao, D.: dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. IEEE Transactions on Image Processing **26**(8), 3951–3964 (2017)
26. Ma, K., Wu, Q., Wang, Z., Duanmu, Z., Yong, H., Li, H., Zhang, L.: Group mad competition-a new methodology to compare objective image quality models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1664–1673 (2016)
27. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 181–196 (2018)
28. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
29. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE transactions on Image Processing **20**(12), 3350–3364 (2011)
30. Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning to count. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5898–5906 (2017)
31. Pan, Z., Yuan, F., Lei, J., Fang, Y., Shao, X., Kwong, S.: Vcrnet: Visual compensation restoration network for no-reference image quality assessment. IEEE Transactions on Image Processing **31**, 1613–1627 (2022)
32. Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: Color image database tid2013: Peculiarities and preliminary results. In: european workshop on visual information processing (EUVIP). pp. 106–111. IEEE (2013)
33. Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: Image database tid2013: Pe-

culiarities, results and perspectives. Signal processing: Image communication **30**, 57–77 (2015)

34. Ren, H., Chen, D., Wang, Y.: Ran4iqa: Restorative adversarial nets for no-reference image quality assessment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
35. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the dct domain. IEEE transactions on Image Processing **21**(8), 3339–3352 (2012)
36. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on image processing **15**(11), 3440–3451 (2006)
37. Shen, Z., Liu, Z., Li, J., Jiang, Y.G., Chen, Y., Xue, X.: Object detection from scratch with deep supervision. IEEE transactions on pattern analysis and machine intelligence **42**(2), 398–412 (2019)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: 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 (CVPR) (June 2020)
40. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8174–8182 (2018)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
42. Wang, J., Fan, H., Hou, X., Xu, Y., Li, T., Lu, X., Fu, L.: Mstriq: No reference image quality assessment based on swin transformer with multi-stage fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1269–1278 (2022)
43. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
44. Wu, Q., Wang, Z., Li, H.: A highly efficient method for blind image quality assessment. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 339–343. IEEE (2015)
45. Xu, J., Ye, P., Li, Q., Du, H., Liu, Y., Doermann, D.: Blind image quality assessment based on high order statistics aggregation. IEEE Transactions on Image Processing **25**(9), 4444–4457 (2016)
46. Xue, W., Mou, X., Zhang, L., Bovik, A.C., Feng, X.: Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. IEEE Transactions on Image Processing **23**(11), 4850–4862 (2014)
47. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019)
48. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1191–1200 (June 2022)
49. Zeng, H., Zhang, L., Bovik, A.C.: A probabilistic quality representation approach to deep blind image quality prediction. arXiv preprint arXiv:1708.08190 (2017)

50. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. IEEE Transactions on Image Processing **24**(8), 2579–2591 (2015)
51. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology **30**(1), 36–47 (2018)
52. Zhang, W., Zhai, K., Zhai, G., Yang, X.: Learning to blindly assess image quality in the laboratory and wild. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 111–115. IEEE (2020)
53. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)
54. Zhu, H., Li, L., Wu, J., Dong, W., Shi, G.: Metaiqa: Deep meta-learning for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14143–14152 (2020)