

Assessing UHD Image Quality from Aesthetics, Distortions, and Saliency

Wei Sun[✉], Weixia Zhang[✉], Yuqin Cao[✉], Linhan Cao[✉], Jun Jia[✉], Zijian Chen[✉], Zicheng Zhang[✉], Xiongkuo Min[✉], and Guangtao Zhai[✉]

Shanghai Jiao Tong University
{sunguwei,zwx8981,caoyuqin,caolinhan,jiajun0302,zijian.chen,
zzc1998,minxiongkuo,zhaiguangtao}@sjtu.edu.cn
[✉] Corresponding Authors

Abstract. UHD images, typically with resolutions equal to or higher than 4K, pose a significant challenge for efficient image quality assessment (IQA) algorithms, as adopting full-resolution images as inputs leads to overwhelming computational complexity and commonly used pre-processing methods like resizing or cropping may cause substantial loss of detail. To address this problem, we design a multi-branch deep neural network (DNN) to assess the quality of UHD images from three perspectives: **global aesthetic characteristics, local technical distortions, and salient content perception**. Specifically, aesthetic features are extracted from low-resolution images downsampled from the UHD ones, which lose high-frequency texture information but still preserve the global aesthetics characteristics. Technical distortions are measured using a fragment image composed of mini-patches cropped from UHD images based on the grid mini-patch sampling strategy. The salient content of UHD images is detected and cropped to extract quality-aware features from the salient regions. We adopt the Swin Transformer Tiny as the backbone networks to extract features from these three perspectives. The extracted features are concatenated and regressed into quality scores by a two-layer multi-layer perceptron (MLP) network. We employ the mean square error (MSE) loss to optimize prediction accuracy and the fidelity loss to optimize prediction monotonicity. Experimental results show that the proposed model achieves the best performance on the UHD-IQA dataset while maintaining the lowest computational complexity, demonstrating its effectiveness and efficiency. Moreover, the proposed model won **first prize in ECCV AIM 2024 UHD-IQA Challenge**. The code is available at <https://github.com/sunwei925/UIQA>.

Keywords: Ultra-high-definition images · Image quality assessment · Image aesthetics · Technical distortions · Image saliency

1 Introduction

In recent years, ultra-high-definition (UHD) images have been explosively growth on the Internet because of the rapid advancements of computational photography techniques and the popularity of social media networks. Image quality

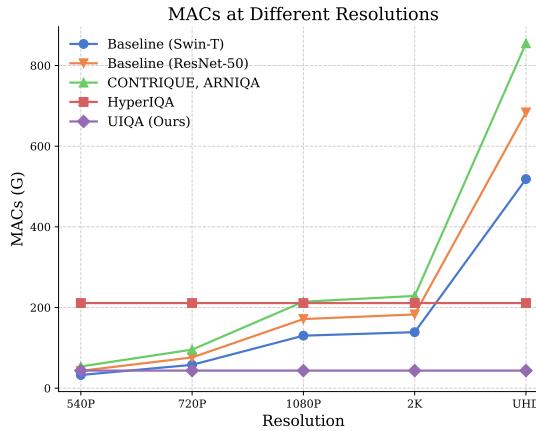


Fig. 1: The computational complexity (MACs) of existing IQA methods with different input resolutions.

assessment (IQA) is developed to provide quality scores of images aligned to human perception and plays a crucial role in various image processing systems. However, as illustrated in Fig. 1, the extremely high resolutions of UHD images (4K or higher) pose a significant challenge to IQA algorithms, as high resolution substantially raises computational complexity, making them impractical for real-time applications and quality monitoring of large-scale UHD images. Thus, it is necessary to develop an efficient and effective IQA model for UHD images.

Existing IQA methods [48] can be broadly classified into two categories: handcrafted-based methods [16, 17, 22–26, 28, 49, 50, 56–58] and deep neural network (DNN)-based methods [2, 4, 12, 14, 20, 21, 29, 32, 34, 39, 53–55, 59]. Handcrafted-based IQA methods primarily leverage prior knowledge of visual quality perception to design quality-related features to quantify image distortions. Popular handcrafted features include natural scene statistics (NSS) features [24–26, 28], free energy features [7, 49, 51], image texture [17, 50, 57], corner/edge [16, 22, 23, 56, 58], etc. Despite their utility, handcrafted features are limited in modeling the impact of varying content on visual quality. Hence, they perform well on synthetically distorted images, where quality degradation is mainly due to synthetic distortions, but perform poorly on authentically distorted images, where image quality is influenced by both image content and in-the-wild distortions [34]. DNN-based IQA methods [2, 4, 12, 21, 29, 34, 55, 59] automatically learn quality-aware features by training a carefully designed deep neural network (DNN) in a learning-based manner, which have demonstrated strong performance in evaluating the quality of both synthetically and authentically distorted images. However, as shown in Fig. 1, the computational complexity of DNN increases dramatically with the resolution of the input image, leading to extremely high computational demands when using full-resolution UHD images as inputs. Common image pre-processing techniques for DNN, such as downsampling and cropping, while sig-

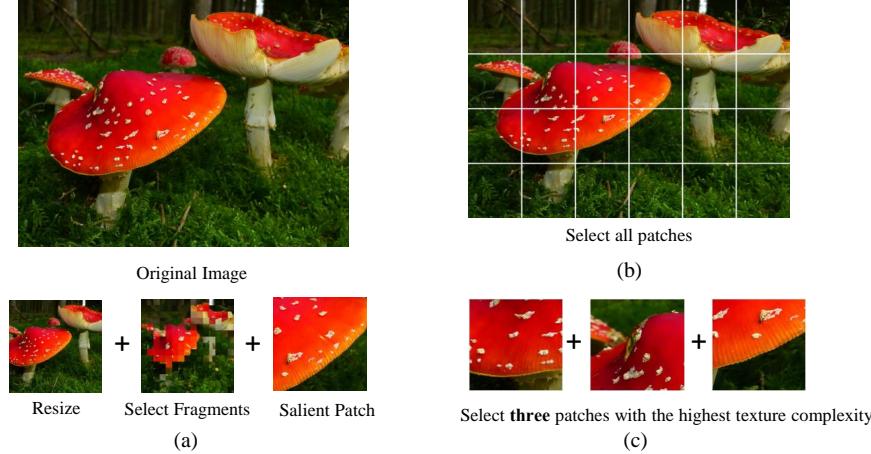


Fig. 2: The different image pre-processing methods for UHD images. (a) is the proposed method, which utilizes the resized image, the fragment image, and the salient patch to extract features of aesthetic, distortion, and salient content. (b) samples all non-overlapped image patches for feature extraction [40]. (c) selects three representative patches with the highest texture complexity for feature extraction [19, 60].

nificantly decreasing the computations, make the UHD image lose substantial details or lack content completeness, thus compromising the effectiveness of UHD image quality evaluation. Moreover, existing studies [10, 37] indicate that it is difficult to optimize DNN-based IQA methods with high-resolution inputs.

In general, there are two primarily factors that affect human perception on image quality: aesthetic characteristics [13, 43] and technical distortions [33, 34, 38, 45]. The aesthetic characteristics refer to human preferences for image content, color style, layout, etc., which are typically global image features and resolution-insensitive. In contrast, the technical distortions refer to image degradation types such as blur, noise, contrast, compression, etc., which are noticeable in local patches and resolution-sensitive. In addition, image saliency [3] is also crucial for image quality assessment, particularly for UHD images. That is because UHD images are more suitable for display on large screens, where human attention is more easily attracted to the salient regions of images. Thus, the perceived quality of UHD images is closely related to the visual quality of these salient regions of images [47, 52].

By individually evaluating the quality of these dimensions, we can effectively reduce computational complexity using appropriate preprocessing techniques. Specifically, for aesthetic quality evaluation, we can aggressively downsample the UHD image into a low-resolution one, which loses textural detail but still preserves the aesthetic-related features. For technical distortion measurement, we can extract local mini-patches from the original image, which compromises the completeness of semantic content but still effectively captures technical dis-

tortions. For salient content perception, we can use the saliency detection algorithm to crop the most salient region of the UHD image, representing the most critical content of the UHD image for human visual perception. As a result, we can use several low-resolution images instead of the full-resolution images for quality-aware feature extraction, thus achieving significantly lower computational complexity. We illustrate this process in Fig. 2 (a).

Therefore, we propose an efficient UHD IQA model that evaluates the perceptual quality of UHD images from three perspectives: **global aesthetic characteristics, local technical distortions, and salient content perception**. As depicted in Fig. 3, our model comprises three branches, each dedicated to one quality dimension. For the aesthetic branch, we resize and crop the original UHD image into a resolution of 480×480 for aesthetic analysis. For the distortion branch, we adopt the grid mini-patch sampling method [45] to uniformly sample local patches across the entire image. Specifically, it first divides the UHD image into 15×15 equal-sized patches and then randomly crops a mini-patch with the resolution of 32×32 from each patch. These mini-patches are further spliced into a composite image with a resolution of 480×480 for distortion analysis. For the branch of salient content perception, although the optimal approach is to utilize a state-of-the-art saliency detection algorithm to identify the most salient region, it would significantly increase computational complexity. Hence, we opt for a simpler approach by exploiting the center bias effect in saliency detection [3], which crops the center patch with a resolution of 480×480 from the UHD image to assess salient content quality. To keep the model simplistic, we utilize the standard Swin Transformer Tiny model as the feature extraction networks. We pre-trained the feature extraction network on AVA [27], a large-scale aesthetic image assessment dataset containing 250,000 images, to enhance the networks' capability for quality-aware feature analysis. We employ both the mean square error (MSE) loss and the fidelity loss [41] to optimize the proposed model. The MSE loss specifically reduces the absolute error relative to the ground truth, whereas the fidelity loss improves the rank consistency between two image pairs. The experimental results on the UHD IQA dataset [9] demonstrate that our model can effectively assess the quality of UHD images with acceptable computational complexity.

In summary, the contributions of this paper are concluded as follows:

- We propose an efficient and effective IQA model to evaluate the quality of UHD images from perspectives of global aesthetic characteristics, local technical distortions, and salient content perception.
- We aggressively downsample UHD images into three low-resolution ones for corresponding dimension feature extraction, thus significantly reducing the computational complexity of the proposed model.
- The experimental results show that the proposed model achieves the best performance on the UHD IQA dataset, even with a simplistic feature extraction network, demonstrating its effectiveness and its broad applicability in real-world scenarios.

2 Related Work

2.1 General-purpose IQA Methods

Handcrafted-based IQA methods. Natural scene statistics (NSS) is the most commonly used handcrafted features for IQA. The motivation behind NSS-based IQA methods is that high-quality images adhere to specific statistical properties, whereas quality degradations cause deviations from these statistics. For example, DIIVINE [26] employs a generalized Gaussian distribution to capture NSS features in the wavelet domain, which are subsequently used to identify the distortion types and regress distortion-specific quality scores. BLIINDS-II [28] and BRISQUE [24] further extract NSS features in the discrete cosine transform (DCT) domain and the spatial domain, respectively, and directly regress the extracted features into quality scores. NIQE [25] measures the distance between NSS features of distorted images and those of a set of high-quality images as quality scores. In addition to NSS, other handcrafted features such as free energy [7, 49, 51], image texture [17, 50, 57], corner/edge [16, 22, 23, 56, 58], etc. are also used in previous IQA models. Nevertheless, handcrafted-based IQA methods are usually designed for synthetically distorted images and perform poorly on in-the-wild images.

DNN-based IQA methods. A typical DNN-based IQA model consists of a feature extraction network to extract quality-aware features and a regression network to map the features into quality scores. For example, Kang *et al.* [12] use a shallow CNN model to estimate the patch scores and the image-level quality score is averaged by the evaluated patch scores of the corresponding image. Bosse *et al.* [4] further deepen the CNNs model by jointly learning the quality and weight of patches. Zhang *et al.* [53] utilize two kinds of CNN models to obtain a better representation, where two CNN are respectively pre-trained on the distortion type and level classification task and the image classification task. Su *et al.* [32] develop a self-adaptive hyper network to aggregate local distortion features and global semantic features. Ke *et al.* [14] propose a multi-scale image quality Transformer (MUSIQ), which utilizes the transformer architecture to solve problems of images with varying sizes and aspect ratios. Sun *et al.* [34] propose a staircase network to extract features from intermediate layers of CNN to enhance the quality-aware feature representation. Chen *et al.* [5] propose a top-down approach that uses high-level semantic information to guide the IQA network to focus on semantically important local distortion regions.

Generalization is a concern for DNN-based IQA methods. To enhance the generalization of IQA models, some studies [34, 54, 55] train feature extraction networks on multiple datasets to increase the diversity of training samples. Unsupervised learning methods [2, 21, 29, 59], such as contrastive learning, are also utilized to train feature extraction networks on large-scale unlabelled images. Recently, due to the powerful zero-shot learning ability of visual-language models, such as CLIP, large multi-modality models (LMMs), some studies [1, 42, 46, 55] fine-tune or directly inference visual-language models with the quality-related textual prompts for IQA tasks.

2.2 High-resolution IQA Methods

Existing high-resolution IQA methods primarily address two issues. First, with the rapid advancement of super-resolution techniques, there exists a significant number of fake UHD images/videos on the Internet. To improve the quality of experience (QoE) of users, it is essential to assess whether UHD images/videos were captured by cameras or upsampled by super-resolution methods, as the latter consume high transmission bandwidth but deliver a subpar QoE. Rishi *et al.* [31] propose a two-stage approach to determine whether a video frame has real or fake 4K resolution. In the first stage, local patches are classified using a lightweight CNN, and in the second stage, these local assessments are aggregated into a global image-level decision using logistical regression. However, this approach cannot evaluate the quality scores of 4K images, which limits its applicability. Zhu *et al.* [60] further develop an IQA metric to distinguish true and pseudo 4K images and simultaneously measure their quality. It extracts frequency domain features and NSS-based features from three representative patches with the highest texture complexity and then maps these features into the true/false 4K labels and the quality scores. Lu *et al.* [19] propose a grey-level co-occurrence matrix (GLCM) based texture complexity measure to select three representative image patches from a 4K image, which are subsequently utilized to extract quality-aware features from the intermediate layers of the convolutional neural network (CNN). The extracted features are regressed into the class probability and the quality score by two multilayer perceptron (MLP) networks via the multi-task learning manner.

Second, with the growing number of UHD images on social media platforms, it is also crucial to assess the quality of in-the-wild UHD images, which on the one hand can optimize the image processing algorithms in camera imaging systems, on the other hand, can help media providers further compress and transmit UHD images efficiently. Huang *et al.* [11] construct a high-resolution image quality database (HRIQ), consisting of 1,200 images with a resolution of 2880×2160 . They further propose a HR-IQA method [15], which employs a CNN to extract features from image patches, followed by training a recurrent neural network separately to estimate the quality scores. Hosu *et al.* [9] also build a large-scale UHD-IQA dataset containing 6,073 4K images with a fixed width of 3840 pixels, focusing on highly aesthetic photos of high technical quality. Tan *et al.* [40] propose a full-pixel covering (FuPiC) training strategy for 4K video quality assessment. They divide the full-resolution video into equal-sized patches, integrate frequency information into these patches, and then obtain patch-level scores and weights using a video Swin Transformer. Finally, the overall video quality score is computed as a weighted sum of the patch-level quality scores.

In summary, UHD image authentic assessment studies typically extract features from a few selected high-frequency patches, which results in low computational complexity but fails to capture the overall image quality characteristics. On the other hand, in-the-wild UHD IQA studies primarily divide the images into dense image patches to extract features of the entire image content, which are integrated and regressed into quality scores. However, the large number of

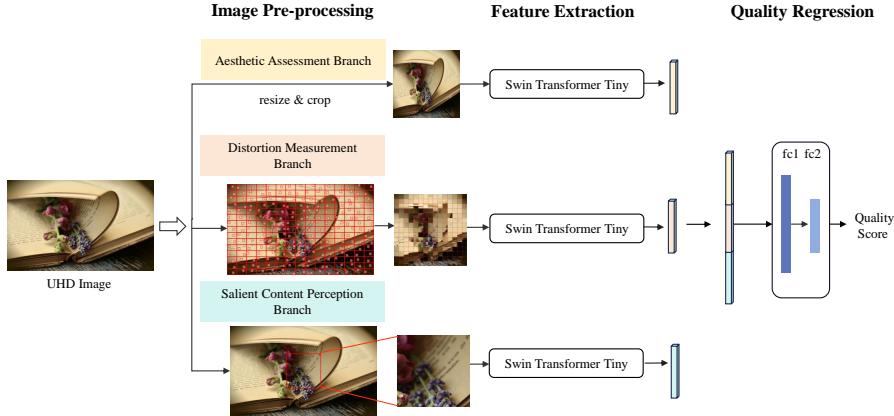


Fig. 3: The diagram of the proposed model. It consists of three modules: the image pre-processing module, the feature extraction module, and the quality regression module. We assess the quality of UHD images from three perspectives: global aesthetic characteristics, local technical distortions, and salient content perception, which are evaluated by the aesthetic assessment branch, distortion measurement branch, and salient content perception branch, respectively.

patches inevitably leads to high computational complexity. We illustrate these two pre-processing Fig. 2 (b) and (c).

3 Methods

As illustrated in Fig. 3, our method is composed of three modules: the image pre-processing module, the feature extraction module, and the quality regression module.

3.1 Image Pre-processing Module

The high-resolution characteristics of UHD images introduce significant computational complexity for IQA methods. Therefore, how to downsample high-resolution images into low-resolution ones in the proper manner is crucial for developing efficient UHD IQA methods. The motivation of the proposed method is to decompose the overall quality perception into three dimensions: aesthetic characteristics, technical distortions, and salient content perception. For each dimension, we perform an appropriate downsampling strategy to decrease the resolution while preserving the quality characteristics specific to that dimension.

Specifically, given a UHD image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, where H and W represent the height and the width, respectively, with the minimum of H and W being equal to or larger than 2160. For aesthetic feature extraction, we observe that the factors that affect aesthetic quality are normally global features, such as image

content, color, layout, luminance, etc. Therefore, downsampling the UHD image directly into a low-resolution one does not significantly impact the extraction of aesthetic features. We denote this process as:

$$\mathbf{x}_l = D(\mathbf{x}), \quad (1)$$

where $D(\cdot)$ represents the bilinear downsampling operator and $\mathbf{x}_l \in \mathbb{R}^{H_l \times W_l \times 3}$ is the low-resolution one of \mathbf{x} . Note that $H_l \ll H$ and $W_l \ll W$.

In contrast, technical distortions, such as noise and texture, are typically low-level features that are sensitive to the original resolution. So, the downsampling process can cause a substantial loss of detail, thus affecting distortion feature extraction. To overcome this problem, we adopt the grid mini-patch sampling strategy [45] to uniformly sample local mini-patches from the original resolution image. To be more specific, we first uniformly split the UHD image \mathbf{x} into $N_g \times N_g$ non-overlapped and equal-size image patches $\mathbf{G} = \{\mathbf{g}_{0,0}, \dots, \mathbf{g}_{i,j}, \dots, \mathbf{g}_{N_g-1,N_g-1}\}$, where $\mathbf{g}_{i,j}$ denotes the patch in the i -th row and j -th column. We formulate this process as:

$$\mathbf{g}_{i,j} = \mathbf{x} \left[\frac{i \times H}{N_g} : \frac{(i+1) \times H}{N_g}, \frac{j \times W}{N_g} : \frac{(j+1) \times W}{N_g} \right], \quad (2)$$

where $0 \leq i, j < N_g$. To preserve the local distortion information, we randomly crop a mini-patch $\mathbf{p}_{i,j}$ with the resolution of $H_p \times W_p$ from the patch $\mathbf{g}_{i,j}$, where $H_p \ll H/N_g$ and $W_p \ll W/N_g$. Then, we splice these mini-patches into a new fragment image \mathbf{x}_f according to their original position:

$$\begin{aligned} \mathbf{x}_f^{i,j} &= \mathbf{x}_f[i \times H_p : (i+1) \times H_p, j \times W_p : (j+1) \times W_p], \\ &= \mathbf{p}_{i,j}, \end{aligned} \quad (3)$$

$0 \leq i, j < N_g.$

We utilize the fragment image \mathbf{x}_f with the height of $H_p \times N_g$ and the width of $W_p \times N_g$ for the distortion feature extraction.

To effectively extract features of salient content, it is essential to select the most salient region of the image. Although numerous saliency detection methods have been proposed in the literature to tackle the task, we observe that state-of-the-art saliency detection methods significantly increase computational complexity. Therefore, we choose a simpler approach by leveraging the center bias effect in saliency detection studies [3], which crops the center patch with the resolution of $H_s \times W_s$ from the UHD image:

$$\mathbf{x}_s = \mathbf{x} \left[\frac{H - H_s}{2} : \frac{H + H_s}{2}, \frac{W - W_s}{2} : \frac{W + W_s}{2} \right], \quad (4)$$

where \mathbf{x}_s represents the salient image patch of the image \mathbf{x} and is used for extracting features of salient content.

3.2 Feature Extraction Module

As illustrated in Fig. 3, we design a multi-branch network to extract features of different dimensions. Specifically, we use Swin Transformer Tiny [18] as the

backbone network due to its excellent feature extraction capabilities as well as low computational complexity. Although previous studies [33–36] suggest that incorporating features extracted from intermediate layers of the backbone network can enhance quality-aware feature representation, we opt for the standard Swin Transformer to keep the model as simple as possible, making it more easily deployable on mobile or edge devices. To further enhance the feature representation, we pre-train the backbone network on the large-scale aesthetic quality assessment dataset AVA [27]. Since each branch of the proposed model is responsible for extracting different types of features, the branches do not share model weights. We formulate the feature extraction process as follows:

$$\begin{aligned}\mathcal{F}_{\text{aes}} &= \text{Swin}_{\text{aes}}(\mathbf{x}_l), \\ \mathcal{F}_{\text{dis}} &= \text{Swin}_{\text{dis}}(\mathbf{x}_f), \\ \mathcal{F}_{\text{sal}} &= \text{Swin}_{\text{sal}}(\mathbf{x}_s),\end{aligned}\tag{5}$$

where \mathcal{F}_{aes} , \mathcal{F}_{dis} , and \mathcal{F}_{sal} are the extracted aesthetic, technical distortion, and salient content features by the feature extractors Swin_{aes} , Swin_{dis} , and Swin_{sal} , respectively.

3.3 Quality Regression Module

After extracting three dimension features, we concatenate them into the final feature representation \mathcal{F} :

$$\mathcal{F} = \text{CAT}(\mathcal{F}_{\text{aes}}, \mathcal{F}_{\text{dis}}, \mathcal{F}_{\text{sal}}),\tag{6}$$

where CAT denotes the concatenation operator.

We then utilize a MLP network to regress \mathcal{F} into the overall quality scores \hat{q} :

$$\hat{q} = \text{MLP}(\mathcal{F}),\tag{7}$$

where MLP denotes the two-layer MLP network consisting of 128 and 1 neurons.

3.4 Loss Function

Mean square error (MSE) and mean absolute error (MAE) are two common loss functions used to minimize the discrepancy between model quality scores \hat{q} and ground-truth quality scores q . However, MSE and MAE cannot optimize the rank consistency between image pairs, which is crucial in real-world applications. Therefore, we adopt a combination of MSE loss and fidelity loss to optimize the proposed model, where fidelity loss [41] targets the rank consistency between image pairs.

Specifically, we randomly sample two images (\mathbf{x}, \mathbf{y}) as a pair and calculate the quality scores of two images $\hat{q}(\mathbf{x})$ and $\hat{q}(\mathbf{y})$ using the proposed model. We then calculate a binary label based on their ground-truth mean opinion scores (MOSs):

$$p(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{if } q(\mathbf{x}) \geq q(\mathbf{y}) \\ 1 & \text{otherwise} \end{cases},\tag{8}$$

Table 1: Performance of the compared models and the proposed model on the validation set of the UHD-IQA dataset. The best-performing model is highlighted in each column

Methods	SRCC \uparrow	PLCC \uparrow	KRCC \uparrow	RMSE \downarrow	MAE \downarrow	MACs (G) \downarrow
HyperIQA [32]	0.524	0.182	0.359	0.087	0.055	211
Effnet-2C-MLSP [44]	0.615	0.627	0.445	0.060	0.050	345
CONTRIQUE [21]	0.716	0.712	0.521	0.049	0.038	855
ARNIQA [2]	0.718	0.717	0.523	0.050	0.039	855
CLIP-IQA+ [42]	0.743	0.732	0.546	0.108	0.087	895
QualiCLIP [1]	0.757	0.752	0.557	0.079	0.064	901
Proposed	0.817	0.823	0.625	0.040	0.032	43.5

where $q(\mathbf{x})$ and $q(\mathbf{y})$ are the ground-truth MOSSs of the image pair (\mathbf{x}, \mathbf{y}) . We estimate the probability of \mathbf{x} perceived better than \mathbf{y} as

$$\hat{p}(\mathbf{x}, \mathbf{y}) = \Phi\left(\frac{\hat{q}(\mathbf{x}) - \hat{q}(\mathbf{y})}{\sqrt{2}}\right), \quad (9)$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function, and the variance is fixed to one. We then calculate fidelity loss for this pair:

$$\ell_{\text{fidelity}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = 1 - \frac{\sqrt{p(\mathbf{x}, \mathbf{y})\hat{p}(\mathbf{x}, \mathbf{y})}}{\sqrt{(1 - p(\mathbf{x}, \mathbf{y}))(1 - \hat{p}(\mathbf{x}, \mathbf{y}))}}, \quad (10)$$

where $\boldsymbol{\theta}$ represents the parameters of the proposed model. MSE loss are calculated as:

$$\ell_{\text{MSE}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = (q(\mathbf{x}) - \hat{q}(\mathbf{x}))^2 + (q(\mathbf{y}) - \hat{q}(\mathbf{y}))^2. \quad (11)$$

The overall loss is calculated as the weighted sum of fidelity loss and MSE loss:

$$\ell(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \alpha \ell_{\text{fidelity}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) + \beta \ell_{\text{MSE}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}), \quad (12)$$

where α and β are weights of fidelity loss and MSE loss respectively.

4 Experiment

4.1 Experimental Protocol

Implementation Details. We utilize Swin Transformer Tiny [18] as the backbone network of the three branches. To enhance its quality-aware feature representation capability, we first pre-train it on the AVA dataset [27], following the same training strategy as the proposed model. For the aesthetic branch, we resize the resolution of the minimum dimension of the UHD image as 512 pixels (*i.e.*, $\min(H_l, W_l) = 512$) while preserving its aspect ratios. During the training and test stages, we randomly crop and centrally crop the resized image to a

Table 2: Performance of the compared models and the proposed model on the test set of the UHD-IQA dataset. The best-performing model is highlighted in each column

Methods	SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	MAE ↓	MACs (G) ↓
HyperIQA [32]	0.553	0.103	0.389	0.118	0.070	211
Effnet-2C-MLSP [44]	0.675	0.641	0.491	0.074	0.059	345
CONTRIQUE [21]	0.732	0.678	0.532	0.073	0.052	855
ARNIQA [2]	0.739	0.694	0.544	0.052	0.739	855
CLIP-IQA+ [42]	0.747	0.709	0.551	0.111	0.089	895
QualiCLIP [1]	0.770	0.725	0.570	0.083	0.066	901
Proposed	0.846	0.798	0.657	0.061	0.042	43.5

resolution of 480×480 as the input, respectively. For the distortion branch, the number of mini-patches $N_g \times N_g$ is set as 15×15 , with each mini-patch having a resolution of 32×32 (*i.e.*, $H_p = W_p = 32$). Thus, the resolution of the fragment image is also 480×480 . For the salient perception branch, the resolution of the center patch is set to 480×480 . The Adam optimizer, with an initial learning rate 1×10^{-5} and a batch size 12, is used to train the proposed model on a server with 2 NVIDIA RTX 3090 GPUs. The learning rate is decayed by a factor of 10 after 10 epochs, with a total of 100 epochs. The weights of fidelity loss α and MSE loss β are set to 1 and 0.1 respectively.

Test Datasets. We evaluate the proposed model on the largest UHD IQA dataset, UHD-IQA [9], which focuses on highly aesthetic photos of high technical quality. UHD-IQA comprises 6,073 images, with 70%, 15%, and 15% for the training, validation, and test sets, respectively. Note that only the labels of the training set are accessible in the AIM 2024 UHD-IQA Challenge: Pushing the Boundaries of Blind Photo Quality Assessment ¹. Therefore, we randomly split the public training set into 80% and 20% for the training-private and validation-private sets respectively, which are used to train the proposed model and select the best one. We report the performance on the validation and test sets for model comparison.

Compared Models. We compare the proposed method with six popular IQA methods, including HyperIQA [32], Effnet-2C-MLSP [44], CONTRIQUE [21], ARNIQA [2], CLIP-IQA+ [42], and QualiCLIP [1].

Evaluation Criteria. We use five criteria to evaluate the performance of IQA models: Spearman Rank-order Correlation Coefficient (SRCC), Pearson’s Linear Correlation Coefficient (PLCC), Kendall’s Rank Correlation Coefficient (KRCC), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

4.2 Experimental Results

We present the performance of IQA models on the validation set in Table 1 and on the test set in Table 2, from which several conclusions can be drawn. First,

¹ <https://codalab.lisn.upsaclay.fr/competitions/19335>

Table 3: The results of AIM 2024 UHD-IQA Challenge

Methods	SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	MAE ↓
SJTU MMLab (ours)	0.846	0.798	0.657	0.061	0.042
CIPLAB	0.835	0.800	0.642	0.064	0.044
ZX AIE Vector	0.795	0.768	0.605	0.062	0.044
I^2 Group	0.788	0.756	0.598	0.066	0.046
Dominator	0.731	0.712	0.539	0.072	0.052
ICL	0.517	0.521	0.361	0.136	0.115

Table 4: The results of ablation studies on model design

Aesthetics	Distortions	Saliency	SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	MAE ↓
✓	✗	✗	0.721	0.738	0.535	0.047	0.037
✗	✓	✗	0.791	0.794	0.598	0.043	0.034
✗	✗	✓	0.696	0.702	0.506	0.050	0.039
✓	✓	✗	0.811	0.820	0.619	0.041	0.032
✓	✗	✓	0.782	0.793	0.588	0.043	0.034
✗	✓	✓	0.787	0.792	0.594	0.044	0.034
✓	✓	✓	0.817	0.823	0.625	0.040	0.032

we observe that existing IQA methods have high computational complexity for UHD images, with MACs exceeding 200G. In contrast, our method has significantly lower computational complexity, with 43.5G MACs ², which is nearly four times less than the fastest IQA method. Second, despite high computational complexity, existing IQA methods still perform poorly on the UHD-IQA dataset, indicating that these methods are unable to effectively capture high-resolution image quality representation. Third, the proposed method achieves the best performance on both the validation and test sets, outperforming the second-best method, QualiCLIP, by 8% on the validation set and 10% on the test set in terms of the SRCC metric, while reducing computational complexity by 20 times, which demonstrates its effectiveness in assessing the quality of UHD images.

We list the results of AIM 2024 UHD-IQA Challenge in Table 3. From Table 3, it is shown that the proposed model significantly outperforms other competing teams, which further demonstrates the superiority of the proposed method.

4.3 Ablation Studies

Network Design. The proposed model consists of three branches for extracting aesthetic, distortion, and salient content features. To validate the effectiveness of these branches, we train each individual branch and combinations of

² It takes 0.068 seconds to calculate the quality score of one image on an NVIDIA RTX 3090 GPU.

Table 5: The results of ablation studies on backbone networks and pre-trained datasets

Backbones	Pre-trained	SRCC \uparrow	PLCC \uparrow	KRCC \uparrow	RMSE \downarrow	MAE \downarrow
MobileNet v2	ImageNet-1k	0.625	0.631	0.444	0.054	0.043
ResNet-50	ImageNet-1k	0.600	0.613	0.427	0.054	0.043
Swin-T	ImageNet-1k	0.789	0.791	0.593	0.044	0.034
Swin-T	AVA	0.817	0.823	0.625	0.040	0.032
Swin-B	AVA	0.830	0.836	0.639	0.039	0.031

two branches on the UHD-IQA dataset. The performance on the validation set is listed in Table 4. For the single-branch network comparison, the distortion branch achieves the best performance, followed by the aesthetic branch, while the salient content perception branch performs moderately. This indicates that the distortion branch plays the most critical role in evaluating the quality of UHD images, suggesting that the grid mini-patch sampling strategy is effective for addressing the UHD IQA task. For the two-branch network comparison, the combination of the aesthetic branch and the distortion branch performs the best, followed by the combination of the distortion branch and the salient branch, and the combination of the aesthetic branch and the salient branch is slightly inferior to the combination of the distortion branch and the salient branch. This further confirms that the distortion branch contributes the most to the proposed model.

Backbones and pre-trained dataset. We investigate the effectiveness of different backbone networks and pre-trained datasets. Specifically, we evaluate two CNN networks: MobileNet v2 [30] and ResNet-50 [8] pre-trained on ImageNet-1k [6]. For the default backbone, Swin Transformer Tiny, we also evaluate its performance when pre-trained on ImageNet-1k and its larger counterpart, Swin Transformer Base, pre-trained on AVA. The results are presented in Table 5. First, by comparing the performance of Swin Transformer Tiny pre-trained on ImageNet-1k and AVA, we observe that pre-trained on AVA boosts model performance, indicating that AVA pre-training can significantly enhance the quality-aware feature representation ability. Second, the performance of MobileNet v2 and ResNet-50 is significantly inferior to that of Swin Transformer Tiny, which may be because CNN is difficult to handle fragment images due to the discontinuity characteristics of fragment images. Third, scaling up the Swin Transformer Tiny into the Base leads to a performance improvement, indicating a more powerful backbone network helps improve evaluation ability.

Resolutions. We study the impact of the resolution of inputs on model performance by setting the inputs of three branches to 224×224 and 384×384 . We train the proposed model with different input resolutions on the UHD-IQA datasets and list the results in Table 6. We observe that increasing the input resolutions consistently improves model performance, because higher resolutions contain more information for feature extraction. Despite this, the largest resolution in Table 6 is still far less than the original resolution of UHD images, demonstrating the model’s efficiency in assessing the quality of UHD images.

Table 6: The results of ablation studies on input resolutions

Resolution	SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	MAE ↓
224 × 224	0.765	0.773	0.570	0.045	0.035
384 × 384	0.809	0.810	0.614	0.042	0.033
480 × 480	0.817	0.823	0.625	0.040	0.032

Table 7: The results of ablation studies on loss function

Loss Function	SRCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	MAE ↓
Fidelity loss	0.804	0.814	0.613	0.042	0.033
MSE loss	0.804	0.813	0.614	0.042	0.033
Combination	0.817	0.823	0.625	0.040	0.032

Loss functions. The proposed model is optimized by a combination of fidelity loss and MSE loss. The effectiveness of each individual loss function is evaluated in Table 7. The results show that fidelity loss and MSE loss individually perform similarly but are both inferior to their combination, highlighting the effectiveness of the combined loss function.

5 Conclusion

In this paper, we design a multi-branch DNN model to evaluate the quality of UHD images from three perspectives: global aesthetic characteristics, local technical distortions, and salient content perception, while avoiding direct processing of high-resolution images. By decomposing the overall quality assessment of a high-resolution image into three quality dimension measurements of the low-resolution ones, our method effectively assesses the quality of UHD images with acceptable computational complexity. What’s more, we avoid complex model designs and use only the standard DNN structures, making it easy to implement in practical applications and optimize for hardware.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62301316, 62371283, 62225112, and 62271312, the China Postdoctoral Science Foundation under Grants 2023TQ0212 and 2023M742298, the Postdoctoral Fellowship Program of CPSF under Grant GZC20231618, the Fundamental Research Funds for the Central Universities, the National Key R&D Program of China (2021YFE0206700), the Science and Technology Commission of Shanghai Municipality (2021SHZDZX0102), and the Shanghai Committee of Science and Technology (22DZ2229005).

References

1. Agnolucci, L., Galteri, L., Bertini, M.: Quality-aware image-text alignment for real-world image quality assessment. arXiv preprint arXiv:2403.11176 (2024)
2. Agnolucci, L., Galteri, L., Bertini, M., Del Bimbo, A.: Arniqa: Learning distortion manifold for image quality assessment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 189–198 (2024)
3. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(1), 185–207 (2012)
4. 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)
5. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing* (2024)
6. 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)
7. Gu, K., Zhai, G., Yang, X., Zhang, W.: Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia* **17**(1), 50–63 (2014)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
9. Hosu, V., Agnolucci, L., Wiedemann, O., Iso, D.: Uhd-iqa benchmark database: Pushing the boundaries of blind photo quality assessment. arXiv preprint arXiv:2406.17472 (2024)
10. 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)
11. Huang, H., Wan, Q., Korhonen, J.: High resolution image quality database. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3105–3109. IEEE (2024)
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. Kao, Y., He, R., Huang, K.: Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing* **26**(3), 1482–1495 (2017)
14. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
15. Korhonen, J., Su, Y., You, J.: Consumer image quality prediction using recurrent neural networks for spatial pooling. arXiv preprint arXiv:2106.00918 (2021)
16. Li, C., Zhang, Z., Sun, W., Min, X., Zhai, G.: A full-reference quality assessment metric for cartoon images. In: 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6. IEEE (2022)
17. Liu, Y., Gu, K., Zhang, Y., Li, X., Zhai, G., Zhao, D., Gao, W.: Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception. *IEEE Transactions on Circuits and Systems for Video Technology* **30**(4), 929–943 (2019)

18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
19. Lu, W., Sun, W., Min, X., Zhu, W., Zhou, Q., He, J., Wang, Q., Zhang, Z., Wang, T., Zhai, G.: Deep neural network for blind visual quality assessment of 4k content. *IEEE Transactions on Broadcasting* **69**(2), 406–421 (2022)
20. Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., Zuo, W.: End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing* **27**(3), 1202–1213 (2017)
21. Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing* **31**, 4149–4161 (2022)
22. Min, X., Zhai, G., Gu, K., Fang, Y., Yang, X., Wu, X., Zhou, J., Liu, X.: Blind quality assessment of compressed images via pseudo structural similarity. In: 2016 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2016)
23. Min, X., Zhai, G., Gu, K., Liu, Y., Yang, X.: Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting* **64**(2), 508–517 (2018)
24. 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)
25. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012)
26. 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)
27. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2408–2415 (2012)
28. 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)
29. Saha, A., Mishra, S., Bovik, A.C.: Re-iqa: Unsupervised learning for image quality assessment in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5846–5855 (2023)
30. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
31. Shah, R.R., Akundy, V.A., Wang, Z.: Real versus fake 4k-authentic resolution assessment. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2185–2189. IEEE (2021)
32. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blinely 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. pp. 3667–3676 (2020)
33. Sun, W., Min, X., Lu, W., Zhai, G.: A deep learning based no-reference quality assessment model for ugc videos. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 856–865 (2022)

34. Sun, W., Min, X., Tu, D., Ma, S., Zhai, G.: Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing* **17**(6), 1178–1192 (2023)
35. Sun, W., Min, X., Zhai, G., Gu, K., Duan, H., Ma, S.: Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment. *IEEE Journal of Selected Topics in Signal Processing* **14**(1), 64–77 (2019)
36. Sun, W., Wang, T., Min, X., Yi, F., Zhai, G.: Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos. In: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–6. IEEE (2021)
37. Sun, W., Wen, W., Min, X., Lan, L., Zhai, G., Ma, K.: Analysis of video quality datasets via design of minimalistic video quality models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
38. Sun, W., Wu, H., Zhang, Z., Jia, J., Zhang, Z., Cao, L., Chen, Q., Min, X., Lin, W., Zhai, G.: Enhancing blind video quality assessment with rich quality-aware features. *arXiv preprint arXiv:2405.08745* (2024)
39. Sun, W., Zhang, W., Jiang, Y., Wu, H., Zhang, Z., Jia, J., Zhou, Y., Ji, Z., Min, X., Lin, W., et al.: Dual-branch network for portrait image quality assessment. *arXiv preprint arXiv:2405.08555* (2024)
40. Tan, X., Zhang, J., Quan, Y., Li, J., Wu, Y., Bian, Z.: Highly efficient no-reference 4k video quality assessment with full-pixel covering sampling and training strategy. *arXiv preprint arXiv:2407.20766* (2024)
41. Tsai, M.F., Liu, T.Y., Qin, T., Chen, H.H., Ma, W.Y.: Frank: a ranking method with fidelity loss. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 383–390 (2007)
42. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
43. Wang, T., Sun, W., Wu, W., Chen, Y., Min, X., Lu, W., Zhang, Z., Zhai, G.: A deep learning-based multidimensional aesthetic quality assessment method for mobile game images. *IEEE Transactions on Games* **15**(4), 658–668 (2022)
44. Wiedemann, O., Hosu, V., Su, S., Saupe, D.: Konx: cross-resolution image quality assessment. *Quality and User Experience* **8**(1), 8 (2023)
45. Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., Lin, W.: Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In: European Conference on Computer Vision. pp. 538–554 (2022)
46. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090* (2023)
47. Yang, S., Jiang, Q., Lin, W., Wang, Y.: Sgdnnet: 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. pp. 1383–1391 (2019)
48. Zhai, G., Min, X.: Perceptual image quality assessment: a survey. *Science China Information Sciences* **63**, 1–52 (2020)
49. Zhai, G., Min, X., Liu, N.: Free-energy principle inspired visual quality assessment: An overview. *Digital Signal Processing* **91**, 11–20 (2019)
50. Zhai, G., Sun, W., Min, X., Zhou, J.: Perceptual quality assessment of low-light image enhancement. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **17**(4), 1–24 (2021)

51. Zhai, G., Wu, X., Yang, X., Lin, W., Zhang, W.: A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing* **21**(1), 41–52 (2011)
52. Zhang, L., Shen, Y., Li, H.: Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing* **23**(10), 4270–4281 (2014)
53. 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 (2020)
54. Zhang, W., Ma, K., Zhai, G., Yang, X.: Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing* **30**, 3474–3486 (2021)
55. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14071–14081 (2023)
56. Zhang, Z., Sun, W., Min, X., Wang, T., Lu, W., Zhai, G.: A full-reference quality assessment metric for fine-grained compressed images. In: 2021 International Conference on Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2021)
57. Zhang, Z., Sun, W., Min, X., Zhu, W., Wang, T., Lu, W., Zhai, G.: A no-reference evaluation metric for low-light image enhancement. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
58. Zhang, Z., Sun, W., Min, X., Zhu, W., Wang, T., Lu, W., Zhai, G.: A no-reference deep learning quality assessment method for super-resolution images based on frequency maps. In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 3170–3174. IEEE (2022)
59. Zhao, K., Yuan, K., Sun, M., Li, M., Wen, X.: Quality-aware pre-trained models for blind image quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22302–22313 (2023)
60. Zhu, W., Zhai, G., Min, X., Yang, X., Zhang, X.P.: Perceptual quality assessment for recognizing true and pseudo 4k content. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2190–2194. IEEE (2021)