

Blind Image Quality Assessment via a Hierarchical Perceptual Modulation Network

LI YU, YIFAN BAO, and JUNYANG LI, School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China

WEI ZHOU, School of Computer Science and Informatics, Cardiff University, Cardiff, UK

ZHAOQING PAN, School of Electrical and Information Engineering, Tianjin University, Tianjin, China

MONCEF GABBOUJ, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

Blind Image Quality Assessment (BIQA) is crucial for perceptual optimization in image processing. To better align objective quality predictions with human perception, researchers have designed BIQA models inspired by the frequency-sensitive characteristics of the Human Visual System (HVS). However, existing methods usually treat High-Frequency (HF) cues as features for direct quality score regression, rather than exploiting them as perceptual modulation signals. In HVS, HF cues play a dynamic role in perceptual modulation, guiding the processing and integration of information across hierarchical stages. Inspired by this, we propose a Hierarchical Perceptual Modulation Network (HPM-Net) for BIQA. In the proposed network, we utilize an Octave Convolution-Based High-Frequency Module (OCHFM) to extract HF cues and then hierarchically modulate them into the quality estimation module through Hierarchical Feature Fusion (HFF). In addition, a Contrastive Learning-Based Distortion Module (CLDM) is integrated with the modulated features through an attentional feature fusion mechanism. Extensive experiments on five benchmarks show that HPM-Net achieves leading or competitive performance compared to state-of-the-art methods in the majority of datasets with both synthetic and authentic distortions.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Neural networks**;

Additional Key Words and Phrases: Blind image quality assessment, high-frequency, hierarchical perceptual modulation, human visual system

ACM Reference format:

Li Yu, Yifan Bao, Junyang Li, Wei Zhou, Zhaoqing Pan, and Moncef Gabbouj. 2026. Blind Image Quality Assessment via a Hierarchical Perceptual Modulation Network. *ACM Trans. Multimedia Comput. Commun. Appl.* 22, 10, Article 289 (October 2026), 19 pages.
<https://doi.org/10.1145/3840393>

This work was supported by the National Natural Science Foundation of China under Grants 62671320 and 62322116.

Authors' Contact Information: Li Yu, School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China; e-mail: li.yu@nuist.edu.cn; Yifan Bao, School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China; e-mail: 202412492116@nuist.edu.cn; Junyang Li, School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China; e-mail: jyongli@foxmail.com; Wei Zhou, School of Computer Science and Informatics, Cardiff University, Cardiff, UK; e-mail: zhouw26@cardiff.ac.uk; Zhaoqing Pan, School of Electrical and Information Engineering, Tianjin University, Tianjin, China; e-mail: zqpan3-c@my.cityu.edu.hk; Moncef Gabbouj (corresponding author), Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland; e-mail: moncef.gabbouj@tuni.fi.



This work is licensed under Creative Commons Attribution International 4.0.

<https://doi.org/10.1145/3840393>

1 Introduction

In recent years, with the advent of the digital age and the rapid development of technology, image data have shown explosive growth. However, the diversity of images and equipment has caused the problem of uneven image quality. At the same time, the image will be affected by a variety of factors in the process of acquisition, transmission, and so on, which leads to the distortion and quality change of the image. This ultimately affects human perception of images and the effectiveness of downstream vision tasks. In such a context, it is particularly important to develop an algorithm that can automatically evaluate the image quality. According to whether the reference information of the original image is needed or not, objective **Image Quality Assessment (IQA)** methods can be categorized into full-reference IQA, reduced-reference IQA, and **No-Reference IQA (NR-IQA)**. In real scenarios, it is almost impossible to obtain the original reference image. Therefore, NR-IQA has become a popular topic in the field of computer vision and image processing.

Traditional NR-IQA methods map distorted images to objective quality scores by extracting hand-crafted features such as **Natural Scene Statistics (NSS)** features [23, 24, 31]. The limitation of traditional methods is their inability to effectively handle complex and diverse image distortions, leading to significant inaccuracies in quality prediction under real-world conditions. To overcome this limitation, deep learning-based IQA methods [7, 27, 37, 39, 50, 55] have attracted widespread attention for their ability to learn complex features and improve accuracy in predicting image quality. However, while demonstrating superior performance, many deep learning-based methods remain as “black boxes” and provide limited incorporation of perceptually motivated mechanisms. Recent efforts have incorporated biologically inspired mechanisms derived from **Human Visual System (HVS)** to enhance perceptual consistency and interpretability, including MEON [19], MMMNet [16], VCRNet [25], Re-IQA [32], and ARNIQA [1]. These HVS-inspired strategies not only enrich feature representations from multiple perceptual levels but have also demonstrated promising generalization to authentic distortions. Thus, these results underscore the value of biologically inspired designs while also hinting at untapped potential in leveraging the frequency-sensitive characteristic of HVS for **Blind Image Quality Assessment (BIQA)**.

Current BIQA methods often treat **High-Frequency (HF)** cues as isolated descriptors for direct quality regression. In this article, HF cues refer to high-frequency perceptual signals related to fine-grained textures, edges, and local structural variations that are relevant to image quality perception. However, this view overlooks its dynamic role as an attentional modulator that guides cross-stage spatial integration. Neurobiological studies [5] suggest that the HVS is sensitive to spatial frequency, and HF cues are closely related to detail discrimination and perceptual sensitivity. This observation motivates us to explore whether HF cues can be used as conditioning signals for hierarchical quality representation learning. To bridge this gap, we propose a novel framework that models HF cues not as features for direct quality prediction, but as a hierarchical perceptual modulation signal.

In summary, we propose HPM-Net, a hierarchical perceptual modulation network for BIQA. Motivated by the hierarchical organization commonly observed in visual perception, HPM-Net differs from methods that treat frequency and spatial cues as parallel inputs: it uses HF cues as hierarchical modulators to guide and reweight perceptual processing, ranging from distortion encoding to context integration. Specifically, **Octave Convolution-Based High-Frequency Module (OCHFM)** decomposes features into low-frequency and HF components, extracts fine-grained HF cues, and injects these modulation signals into the backbone **Quality Estimation Module (QEM)** across layers to steer feature learning. **Contrastive Learning-Based Distortion Module (CLDM)**

is pre-trained via contrastive learning to learn distortion-aware representation and is integrated with the modulated stream through an **Attentional Feature Fusion (AFF)** mechanism, improving robustness to authentic distortions. Finally, **Hierarchical Feature Fusion (HFF)** adaptively integrates HF cues with globally contextualized quality representations using importance-aware weighting, balancing texture fidelity and semantic distortion awareness. Together, these components realize hierarchical perceptual modulation and yield coherent, robust BIQA predictions across diverse distortion types.

The main contributions of this article are as follows:

- We propose a novel perspective in utilizing HF cues for BIQA: from using it for direct quality regression to employing it as a perceptual modulation signal. Unlike methods that map handcrafted HF metrics directly to scores, our method leverages HF cues to guide and reweight the hierarchical processing of spatial features, following a perceptually motivated design.
- We realize this paradigm via HPM-Net, which includes an OCHFM for HF feature extraction, a CLDM for distortion-aware representation learning, and the HFF for attention-guided integration.
- Extensive experiments on five databases demonstrate that HPM-Net achieves leading or competitive performance across synthetic and authentic distortion datasets.

The remainder of this article is organized as follows: Section 2 describes the related work on the BIQA method. Section 3 describes our proposed HVS-inspired BIQA approach. Section 4 gives the experimental results with related analysis. The article is summarized in Section 5.

2 Related Work

2.1 BIQA Based on Traditional Hand-Crafted Features

Current generalized BIQA methods can be broadly classified into two categories: hand-crafted methods and learning based methods. Traditional hand-crafted IQA approaches manually extract features like NSS. In early NSS-based methods, features are usually extracted through multiple transform domains, such as wavelet transform domain and Discrete Cosine Transform (DCT) domain. Moorthy and Bovik [24] proposed DIIVINE, a wavelet-based method that first identifies the distortion type and then predicts image quality. Saad et al. [31] proposed BLINDS-II to use NSS with DCT coefficients, dividing images into blocks for DCT, but both have high computational cost. Mittal et al. [23] proposed BRISQUE, a spatial-domain method. It uses locally normalized luminance coefficients, fitting the image's MSCN coefficient distributions with an asymmetric generalized Gaussian model to quantify natural loss. CORNIA [46] uses a codebook-based approach to learn discriminative image features directly from raw image pixels, rather than using hand-crafted features. However, hand-crafted methods rely on NSS or prior knowledge, offer limited semantic modeling, generalize poorly to authentic or unseen distortions, are sensitive to design choices, and lack end-to-end optimization.

2.2 BIQA Based on Deep Learning

With the rise of deep learning, many deep learning methods are applied to IQA tasks. Liu et al. [18] proposed a Siamese network based on Rank Learning to alleviate the problem of limited training samples. Yang et al. [44] developed a joint distortion restoration and quality feature learning framework with multi-stage feature fusion to better capture multi-scale and semantic cues. Zeng et al. [49] designed a probabilistic quality representation that models both the **Mean Opinion Score (MOS)** and score variance, enabling faster and more stable training of deep BIQA models

under limited data. Bosse et al. [2] proposed WaDIQaM, which divides images into patches, predicts each patch's score with corresponding weights, and addresses non-uniform quality distribution among image patches. Su et al. [35] introduced hyperIQA, which fuses local distortion features with global semantic features to aggregate fine-grained details and global information, enabling adaptive IQA. Additionally, Yang et al. [45] designed an auxiliary information guided self-attention network, which leverages additional cues to reweight channels and spatial features for more reliable IQA. Zhang et al. [53] built a large-scale CGI quality database and a deep NR-IQA model that better exploits complex feature statistics for image quality prediction. Recent studies have also explored richer subjective quality information beyond the MOS. Min et al. [22] proposed RichIQA to exploit MOS, the standard deviation of opinion scores, and the distribution of opinion scores for in-the-wild IQA. Gao et al. [6] further investigated Gaussian mixture distribution for modeling opinion score distributions in BIQA. These methods demonstrate the importance of subjective score distributions for quality prediction. Different from these distribution-aware methods, our work focuses on modeling HF cues as hierarchical perceptual modulation signals to guide quality representation learning.

2.3 BIQA Based on HVS-Inspired Mechanisms

To improve the performance of BIQA and alleviate the limitations of small-scale datasets, recent studies have leveraged perceptual signals from related visual tasks to guide quality modeling. This strategy aims to enhance quality assessment by introducing auxiliary branches motivated by perceptual characteristics and fully utilizing feature information from related domains. For example, Lin and Wang [17] proposed Hallucinated-IQA, which synthesizes pseudo-references using GANs to mimic reference perception in the absence of ground-truth images. Yan et al. [43] adopted a two-stream scheme to decouple structure and intensity information, while MMM-Net [16] models saliency-aware attention guided by HVS priors. MBCNN [26] further introduced spatial and gradient domain branches to focus on structural degradation. Re-IQA [32] proposed an unsupervised Mixture of Experts framework that extracts semantic and quality features through dual encoders. DepictQA [48] introduced a multi-modal language model to provide fine-grained quality reasoning beyond scalar scores. Recent studies have further extended IQA from conventional 2D images to emerging visual media, such as light field images and omnidirectional images. For example, Zhang et al. [51] proposed a multi-dimension attention network for full-reference light field IQA to model spatial-angular quality degradation, while Zhang et al. [52] introduced a saliency-guided no-reference omnidirectional IQA method via scene content perceiving.

Despite the promising progress in incorporating spatial-domain human perceptual priors, a fundamental limitation persists: the lack of a BIQA framework that uses HF cues as perceptually motivated conditioning signals for hierarchical quality representation learning. While the seminal work of DeValois and DeValois [5] establishes the HVS's acute sensitivity to spatial frequencies and highlights that HF cues are paramount for detail discrimination, existing frequency-aware BIQA methods [36] treat HF cues in an isolated manner. They primarily employ handcrafted frequency metrics (e.g., energy ratio, variance) that are directly mapped to quality scores, failing to integrate HF cues into a hierarchical model of perception. This disconnect motivates our core hypothesis: HF cues in BIQA should not be treated as a fixed descriptor, but as a hierarchical perceptual modulation signal that guides representation learning across perceptual layers. This view is loosely motivated by the hierarchical nature of visual processing, where low-level cues and higher-level representations interact during perception. To operationalize this paradigm, we propose HPM-Net, which explicitly exploits HF cues to perform hierarchical modulation for perceptual quality estimation.

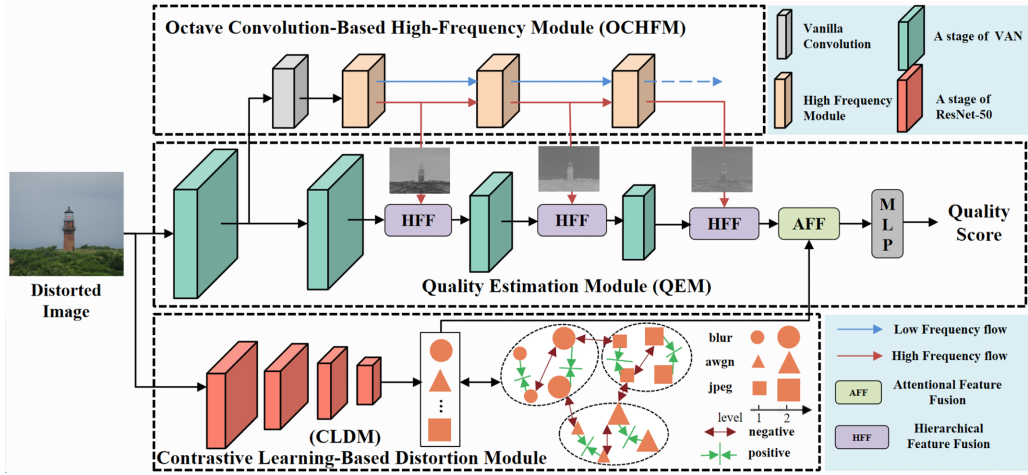


Fig. 1. The overall framework of the proposed HPM-Net method. Our model adopts the QEM as the backbone network and incorporates three key modules: an OCHFM, a CLDM, and a HFF. Specifically, the OCHFM consists of vanilla convolution and HF extraction components, which help guide the QEM to focus on key image regions by capturing fine-grained HF cues. The QEM itself employs the VAN network as its backbone to extract hierarchical feature representations from the input image. The CLDM is built upon ResNet-50 and is pre-trained using contrastive learning to improve the robustness of distortion feature representations. The HFF leverages an attention mechanism to adaptively fuse HF features with hierarchical quality features from QEM, thereby enhancing the quality-aware representation. Additionally, an AFF module is integrated to further refine the aggregation of distortion-aware information.

3 Proposed Method

Motivated by frequency-sensitive and hierarchical visual perception, we present HPM-Net, which treats HF cues as conditioning signals for hierarchical perceptual modulation in BIQA. In this article, HF cues refer to perceptually relevant HF signals, such as fine-grained textures, edges, and local structural variations. In the network implementation, these cues are represented by HF features extracted by the OCHFM. As shown in Figure 1, our framework comprises a QEM and three core modules for perceptually motivated hierarchical modulation, namely an OCHFM, a CLDM, and an HFF. The QEM provides the main prediction pathway, while the OCHFM extracts HF features that are injected into the intermediate representations of the QEM through HFF for stage-wise feature reweighting and refinement. Meanwhile, the CLDM provides auxiliary distortion-aware representations, which are integrated with the modulated quality features through AFF before final quality prediction.

Given a distorted image $I \in \mathbb{R}^{3 \times w \times h}$, where w and h denote the width and height, the objective is to predict its perceptual quality score. For the QEM, **Visual Attention Network (VAN)** [10] is employed as the backbone. The attention mechanism within this network operates through decomposed large kernel convolution, an approach that adeptly secures the representation of long-term dependencies inherent in image features, concurrently achieving a decrease in the computational requirements. Image I is first fed into the QEM. Let f_i denote the i th block of the backbone module and g_i denote the i th block of the HF module of the OCHFM. After the first block of backbone module, the feature F_1 is obtained, then F_1 is fed into the OCHFM. The whole process can be described as follows:

$$F_{i+1} = \text{HFF}_i(f_{i+1}(F_i), g_i(f_1(I))), \quad (1)$$

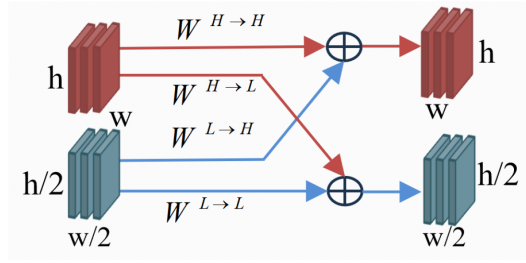


Fig. 2. OCHFM details: This module decomposes input features into HF (red) and low-frequency (blue) components, four convolutional paths are applied: $W^{H \rightarrow H}$ and $W^{L \rightarrow L}$ retain intra-branch information, while $W^{H \rightarrow L}$ and $W^{L \rightarrow H}$ enable inter-branch communication.

where $i \in \{1, 2, 3\}$, F_{i+1} denotes the output fused features, and $\text{HFF}_i(\cdot)$ denotes the HFF operation at the i th stage. Here, $g_i(f_i(I))$ denotes the HF representation generated by the i th OCHFM stage from the first-stage backbone feature $f_i(I)$. This operation enables the OCHFM to extract learnable HF feature representations from fine-grained textures, edges, and local structural variations. In HPM-Net, hierarchical perceptual modulation is formulated as an HF-conditioned, stage-wise feature reweighting, and injection process. At each stage, HFF takes the HF features as conditioning signals to reweight the corresponding QEM features and injects the refined representation back into the main backbone stream. The resulting reweighted quality features are further propagated to subsequent backbone layers, allowing HF cues to guide the evolution of quality-aware representations. In this sense, HFF serves as the concrete implementation of hierarchical perceptual modulation by using HF cues to adaptively reweight and refine QEM features. Finally, the modulated quality features are integrated with distortion-aware features from the CLDM and fed into the MLP head to predict perceptual quality scores.

3.1 OCHFM

The OCHFM is designed to adaptively capture perceptually salient HF cues, inspired by the role of early visual processing in highlighting semantically informative regions. Unlike methods relying on fixed filter banks (e.g., DCT, wavelet) that offer rigid frequency decomposition, we employ Octave Convolution [4], a learnable strategy that decomposes features into HF and low-frequency bands and enables interactive information exchange between them. This adaptive decomposition enhances HF cues associated with textures and edges, effectively generating a modulation signal that guides subsequent stages of processing. The architecture of the OCHFM is illustrated in Figure 2, which depicts the information flow between its HF and low-frequency branches. The flow of the HF extraction module is shown below:

$$\begin{cases} F_{i+1}^H = f(F_i^H; W_i^{H \rightarrow H}) + \text{up}(f(F_i^L; W_i^{L \rightarrow H}), 2), \\ F_{i+1}^L = f(F_i^L; W_i^{L \rightarrow L}) + f(\text{down}(F_i^H, 2); W_i^{H \rightarrow L}), \end{cases} \quad (2)$$

where F_{i+1}^L and F_{i+1}^H denote the low-frequency and HF features, respectively, and $f(F; W)$ denotes a convolution operation with parameter W . $\text{up}(F; 2)$ denotes upsampling using nearest interpolation method, and $\text{down}(F; 2)$ denotes twice the downsampling using the average pooling layer.

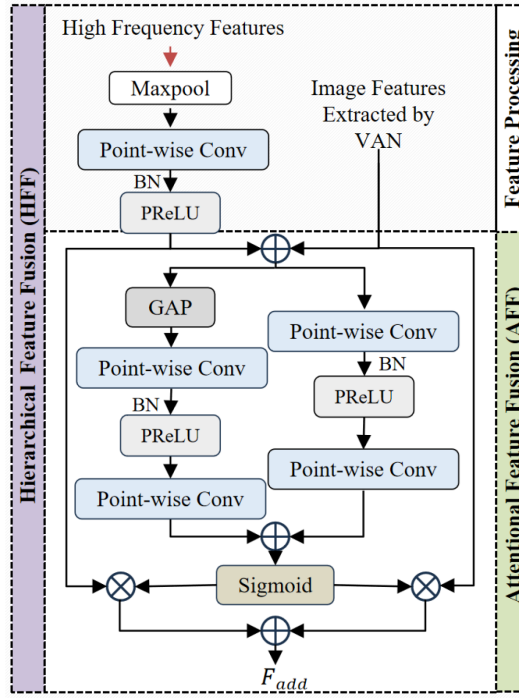


Fig. 3. The architecture of the HFF, consisting of feature processing and the AFF. The inputs to the module are the HF feature (top left) and the image quality feature extracted by VAN (top right). These are fused into F_{add} and then routed through two complementary context branches to compute fusion weights via a sigmoid function.

3.2 HFF

Motivated by the idea that visual quality perception involves adaptive integration of low-level details and higher-level contextual information, we design HFF to integrate HF features with hierarchical quality features through attention-guided reweighting.

Figure 3 depicts the structure of the HFF, which includes feature processing and AFF. For the HF features derived from the OCHFM, we downsample the features using a max-pooling layer to effectively retain the most salient HF cues of the image. First, the features are downsampled to the same size as the quality features using the max-pooling layer, followed by processing the channel dimensions using point-wise convolution of 1×1 size. Then, batch normalization [13] and PReLU layer [11] are used to increase the nonlinear relationship between the layers and accelerate the training process. Subsequently, the processed HF features and the quality features, which are derived from the QEM, are input into HFF, a hierarchical feature fusion that operates on the basis of the attention mechanism. In HFF, the two feature streams are first summed and then processed by two complementary context branches. Their outputs are aggregated and passed through a sigmoid to produce channel-wise attention weights, yielding context-aware channel attention. The whole process can be described as follows:

$$F_{hf} = f_{PReLU} \left(f_{BN} \left(f_{Conv} \left(f_{down} \left(F_{hf}^0 \right) \right) \right) \right), \quad (3)$$

$$F_{output} = w (F_{hf} \oplus F_{qf}) \otimes F_{hf} + (1 - w) (F_{hf} \oplus F_{qf}) \otimes F_{qf}, \quad (4)$$

where F_{hf}^0 denotes the HF features and F_{qf} denotes the image feature extracted by VAN. f_{down} denotes the max-pooling operation, f_{Conv} denotes the point-wise convolution, f_{BN} denotes the batch normalization operation, f_{ReLU} denotes the PReLU activation function, $\oplus$ denotes broadcast addition, and $\otimes$ denotes element-wise multiplication. w denotes the feature weights computed by the sigmoid function in the AFF, and the computation process of w can be expressed as follows:

$$w = f_{\text{sigmoid}}(f_{\text{Conv}}(f_{\text{BN}}(f_{\text{Conv}}(f_{\text{GAP}}(F_{\text{add}})))) \oplus f_{\text{Conv}}(f_{\text{BN}}(f_{\text{Conv}}(F_{\text{add}})))), \quad (5)$$

where f_{sigmoid} denotes the sigmoid activation function, and F_{add} denotes $F_{hf} \oplus F_{qf}$.

3.3 CLDM

The HVS assessment of image quality is an active and non-uniform process, rather than a uniform scan. Visual attention is automatically captured by distorted regions that break the statistical regularities of natural scenes, such as compression blocks in clear sky or noise speckles on smooth skin. Such involuntary shifts of focus effectively serve as a rapid assessment of both the type and severity of distortions. To inject analogous information into BIQA models, established methods [19, 42] train a distortion-type recognition sub-network in a supervised manner and use its predictions as auxiliary cues for quality estimation. However, this classification-based design is restricted to a small set of predefined synthetic distortion categories and therefore generalizes poorly to authentic, mixed distortions. Moreover, annotated IQA datasets with subjective scores are relatively small and expensive to collect, making purely supervised distortion classification difficult to scale. We therefore resort to an unsupervised training paradigm, specifically contrastive learning, which has been widely used in computer vision to learn distortion-aware image representations from unlabeled data. This strategy of leveraging contrastive pre-training to overcome data scarcity aligns with recent evidence in quality-driven image enhancement tasks [40]. The backbone network for the distortion-aware subtask uses the ResNet-50 encoder that has been trained through the processes outlined in CONTRIQUE [21]. For synthetic distorted images, the method employs known distortion types as classification labels and establishes associations by measuring the pairwise similarity of images in the batch. This approach is intended to enable the model to learn to capture the relationship between similar distortion types and similar quality grade levels, providing effective information for subsequent IQA tasks. The synthetic distortion loss function is the normalized temperature-scale cross-entropy (NT-Xent) loss, which is defined for the input image x_i as:

$$\mathcal{L}_i^{\text{syn}} = \frac{1}{|P(i)|} \sum_{j \in P(i)} -\log \frac{\exp(\phi(z_i, z_j)/\tau)}{\sum_{k=1}^N \mathbb{I}_{k \neq i} \cdot \exp(\phi(z_i, z_k)/\tau)}, \quad (6)$$

where N denotes the batch size of the image, z denotes the feature vector of the encoder output, $\phi(\cdot)$ denotes the vector dot product operation, $\mathbb{I}$ denotes the indicator function, τ denotes the temperature parameter, and $P(i)$ refers to the set of images belonging to the same class as x_i . For authentic distorted images in the wild, however, reliable distortion-type annotations are usually unavailable. Moreover, real-world degradations often arise from mixed, content-dependent, and spatially non-uniform factors, making it difficult to assign each image to a predefined distortion category. Therefore, we do not force authentic images into manually defined distortion classes. Instead, each authentic image is treated as a pseudo class in an instance-level contrastive objective, where different quality-preserving views of the same distorted image, such as full-scale and half-scale views, are regarded as positive pairs. This design encourages the encoder to learn degradation-related representations that are stable under quality-preserving transformations, without assuming explicit distortion-semantic labels for authentic images. Through this hybrid contrastive strategy, the synthetic branch provides explicit distortion-type guidance, while the authentic branch improves

Table 1. Detailed Information of Five Common Datasets

Dataset	Reference Images	Distorted Images	Distortion Type	Distortion Types	Score Type	Score Range
LIVE [34]	29	779	Synthetic	5	DMOS	[0,100]
CSIQ [15]	30	866	Synthetic	6	DMOS	[0,1]
TID2013 [28]	25	3,000	Synthetic	24	MOS	[0,9]
LIVEC [8]	N/A	1,162	Authentic	N/A	MOS	[0,100]
KonIQ [12]	N/A	10,073	Authentic	N/A	MOS	[1,5]

the invariance and discriminability of degradation-related representations for real-world distorted images. The loss function for authentic distorted images is defined as:

$$\mathcal{L}_i^{\text{aut}} = -\log \frac{\exp(\phi(z_i, z_j) / \tau)}{\sum_{k=1}^N \mathbb{I}(k \neq i) \cdot \exp(\phi(z_i, z_k) / \tau)}. \quad (7)$$

In the image data augmentation stage, each image has two transform versions, full-scale and half-scale transformations, resulting in data samples where at least two distorted images belong to the same class. By introducing the half-scale transformation, the richness of the dataset can be increased while maintaining the integrity of the image information, thus improving the generalization ability and robustness of the model. The total pretraining loss is defined as:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N [\mathbb{I}(x_i \notin \text{Authentic}) \cdot \mathcal{L}_i^{\text{syn}} + \mathbb{I}(x_i \in \text{Authentic}) \cdot \mathcal{L}_i^{\text{aut}}], \quad (8)$$

where $\mathbb{I}$ is an indicator function to determine whether the input image is authentic distortion or not. Through this hybrid contrastive strategy, the synthetic distortion branch provides explicit distortion-type guidance, while the authentic-image branch improves the invariance of quality-relevant representations under quality-preserving transformations. The resulting encoder provides auxiliary degradation-aware features for the subsequent BIQA model, which helps improve robustness to both synthetic and authentic distortions. It should be noted that the approach diverges from the one outlined in CONTRIQUE, wherein the original-sized image is opted for as the direct input, bypassing the use of features derived from multiple scales. The utilization of features extracted from an image at its original dimensions following the processing by a pre-trained encoder is considered to result in a comprehensive depiction of the distorted features.

4 Experiments and Analyses

4.1 Experimental Setup

To effectively evaluate the performance of our proposed model, five common IQA datasets were utilized. These include three synthetic distortion datasets (LIVE [34], CSIQ [15], TID2013 [28]) and two authentic distortion datasets (LIVEC [8], KONIQ [12]). Details of the five common datasets are shown in Table 1, where MOS denotes mean opinion score and DMOS denotes difference mean opinion score.

The experiments are conducted on an NVIDIA GeForce RTX 4090, and our proposed method is implemented through the PyTorch framework. For data preprocessing, patches from each image are randomly selected to cover a portion of the original distorted image, and these patches are subjected to random horizontal and vertical flipping for the purpose of data augmentation. Aggressive data augmentation strategies are avoided, and it should be noted that the flipping strategy used does not affect the quality of the images. The standard training strategy [9] for existing IQA algorithms was employed. During the procedure, a total of 50 patches, with each patch being a square of 224×224 pixels in dimension, are randomly chosen from every image that is part of the training

dataset. The quality scores from the original distorted images are assumed to be inherited by these patches. During training, the learning rate of the model was set to an initial value of $2e - 5$ and reduced by a factor of 10 at predefined milestones. The Adam optimizer was used with weight decay of $5e - 4$, and batch size set to 64. During the model's testing phase, 224×224 patches are cropped randomly from each test image. Subsequently, the quality scores predicted by each patch are averaged to determine the final quality score of the image. The common IQA experimental approach was followed [9], and the same setup was adopted for all experiments. In the dataset, 80% of the distorted images were used in the training phase, while the remaining 20% were set aside for the testing phase. To eliminate the effect of sampling bias in the dataset, each database was randomly split 10 times for all experiments.

For the assessment of model performance, two commonly used metrics were employed: the **Spearman Rank-Order Correlation Coefficient (SRCC)** [29] and the **Pearson Linear Correlation Coefficient (PLCC)** [33]. SRCC reflects the relationship between objective quality scores and predicted scores, and PLCC reflects the linear relationship between objective quality scores and predicted scores. The formula for SRCC is:

$$SRCC = 1 - \frac{6 \sum_{i=1}^B (x_i - y_i)^2}{B(B^2 - 1)}, \quad (9)$$

where B denotes the number of samples, and x_i and y_i denote the subjective quality score ranking and objective score ranking of image i , respectively. The formula for PLCC is:

$$PLCC = \frac{\sum_i (s_i - \hat{s})(o_i - \hat{o})}{\sqrt{\sum_i (s_i - \hat{s})^2} \sqrt{\sum_i (o_i - \hat{o})^2}}, \quad (10)$$

where s_i and o_i denote the predicted and objective quality scores, respectively, and $\hat{s}$ and $\hat{o}$ denote the corresponding mean scores, respectively.

4.2 Performance Comparison with the State-of-the-Art (SOTA) Methods

To validate the superior performance of our proposed method, we conducted a comparative analysis using a curated set of 20 representative IQA models, including 4 hand-crafted-based methods, DIIVINE [24], BRISQUE [23], CORNIA [46], and HOSA [41], and 16 deep-learning-based methods, MEON [19], WaDIQaM [2], TS-CNN [43], RAN4IQA [30], TRIQ [47], MetaIQA [54], HyperIQA [35], CaHDC [39], TReS [9], MMMNet [16], VCRNet [25], Re-IQA [32], ARNIQA [1], DepictQA [48], TOPIQ [3], and Q-Align [38]. The experimental results are shown in Table 2, with the best and second-best results highlighted in bold and underline, respectively. The results presented in Table 2 allow for the following conclusions to be derived: First and foremost, it is determined that the proposed method occupies the optimal tier within the spectrum of the majority of methods. Of a total of 10 metrics in the 5 datasets, the proposed method achieved 4 optimal results and 6 competitive results. Specifically, the proposed model achieves 1.66%, 1.93% (PLCC, SRCC) higher than ARNIQA [1] (second-best) in the TID2013 dataset, which indicates that the proposed HPM-Net excels in handling the quality assessment task for images with diverse and complex synthetic distortions. Second, on the LIVE dataset, HPM-Net does not achieve the best SRCC, but it remains highly competitive. Specifically, it obtains the best or tied-best PLCC and its SRCC is only 0.1% lower than that of VCRNet [25]. This indicates that the proposed method maintains competitive performance even on relatively small synthetic datasets. The effectiveness of the proposed method is further demonstrated by its excellent performance on the large-scale TID2013 dataset. Finally, at the dataset level, although the proposed method is not specifically optimized for either synthetic or authentic distortions, it still achieves strong and competitive performance on both types of datasets.

Table 2. Performance Comparison with the SOTA Methods on the Three Synthetic and Two Authentic Datasets

	LIVE		CSIQ		TID2013		LIVEC		KonIQ	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
DIIVINE [24]	0.908	0.892	0.776	0.804	0.567	0.643	0.591	0.588	0.558	0.546
BRISQUE [23]	0.944	0.929	0.748	0.812	0.571	0.626	0.629	0.629	0.685	0.681
CORNIA [46]	0.938	0.937	0.768	0.695	0.734	0.657	0.636	0.617	0.569	0.545
HOSA [41]	0.943	0.942	0.805	0.721	0.835	0.749	0.657	0.621	0.703	0.684
MEON [19]	0.955	0.951	0.864	0.852	0.824	0.808	0.710	0.697	0.628	0.611
WaDIQaM [2]	0.955	0.960	0.844	0.852	0.855	0.835	0.671	0.682	0.807	0.804
TS-CNN [43]	0.965	0.969	0.904	0.892	0.824	0.783	0.687	0.654	0.724	0.713
RAN4IQA [30]	0.962	0.961	0.931	0.914	0.859	0.820	0.612	0.586	0.763	0.752
TRIQ [47]	0.965	0.949	0.838	0.825	0.858	0.846	0.861	0.845	0.903	0.892
MetaIQA [54]	0.959	0.960	0.908	0.899	0.868	0.856	0.802	0.835	0.856	0.887
HyperIQA [35]	0.966	0.962	0.942	0.923	0.858	0.840	0.882	0.859	0.917	0.906
CaHDC [39]	0.964	0.965	0.914	0.903	0.885	0.862	0.744	0.738	/	/
TreS [9]	0.968	0.969	0.942	0.922	0.883	0.863	0.877	0.846	<u>0.928</u>	0.915
MMNet [16]	0.970	0.970	0.937	0.924	0.853	0.832	0.876	0.852	/	/
VCRNet [25]	0.973	0.975	0.932	0.914	0.886	0.865	0.878	<u>0.864</u>	0.919	0.906
Re-IQA [32]	<u>0.971</u>	0.970	0.960	<u>0.947</u>	0.861	0.804	0.854	0.840	0.923	0.914
ARNIQA [1]	0.970	0.966	0.973	0.962	<u>0.901</u>	<u>0.880</u>	/	/	/	/
DepictQA [48]	0.914	0.913	0.954	0.940	0.886	0.853	/	/	/	/
TOPIQ [3]	0.942	0.943	0.925	0.908	0.845	0.813	0.868	0.833	0.925	0.915
Q-Align [38]	0.919	0.913	0.936	0.915	0.829	0.808	0.921	0.931	0.934	0.935
HPM-Net	0.973	<u>0.974</u>	<u>0.967</u>	0.962	0.916	0.897	0.880	<u>0.864</u>	<u>0.928</u>	<u>0.919</u>

SOTA, State-of-the-Art.

Specifically, HPM-Net achieves the best results of 0.916 and 0.897 (PLCC, SRCC) on TID2013. On KonIQ, although the recent MLLM-based Q-Align [38] achieves the best performance, HPM-Net still obtains competitive results of 0.928 and 0.919 (PLCC, SRCC), achieving second-best or comparable performance among the compared methods. In summary, with cross-frequency interactions modeled via OCHFM, distortion-aware representation learning through CLDM, and distortion-aware cues fused with multi-scale features using HFF, the proposed HPM-Net effectively establishes the mapping relationship between image distortion characteristics and quality perception and achieves enhanced accuracy in BIQA.

4.3 Performance Evaluation of Individual Distortion Types

To evaluate and compare performance across various image distortion types, we conducted additional experiments using the LIVE, CSIQ, and TID2013 datasets for comparative analysis. Across all utilized databases, the training process involves exposure to all types of distortions, whereas the testing phase is confined to evaluating against a single type of distortion. The training data are not duplicated with the test data. The methods compared include some traditional IQA methods (e.g., DIIVINE [24], BRISQUE [23]) and recent deep learning-based methods (e.g., TS-CNN [43], MMNet [16], and VCRNet [25]). In Table 3, we evaluate the performance of LIVE dataset for all distortion types: JP2K, JPEG, WN, GBLUR, and FF. From the results in Table 3, it can be observed that our proposed method achieves competitive results in most distortion types. Specifically, the proposed method performs well in JPEG and FF, reaching 0.974, 0.988 (SRCC, PLCC) and 0.947, 0.945 (SRCC, PLCC), respectively. The effectiveness of this result is due to the fact that HF cues inherent in the two distortion types are efficiently harnessed by our proposed OCHFM. For Gaussian blur, HF responses are usually attenuated, which may reduce the informativeness of HF cues and limit

Table 3. Performance of Single Distortion Types in the LIVE Dataset

Method	SRCC					PLCC				
	JP2K	JPEG	WN	GBLUR	FF	JP2K	JPEG	WN	GBLUR	FF
DIIVINE [24]	0.925	0.913	0.985	0.789	0.873	0.901	0.887	<u>0.987</u>	0.787	0.879
BRISQUE [23]	0.914	0.965	0.979	0.951	0.877	0.923	0.973	0.985	0.951	0.903
CORNIA [46]	0.936	0.934	0.962	0.926	0.912	0.924	0.906	0.945	0.934	0.913
HOSA [41]	0.928	0.936	0.964	0.964	0.934	0.923	0.924	0.959	0.965	0.923
CNN [14]	0.936	0.965	0.974	0.952	0.906	0.942	0.973	0.976	0.953	0.907
MEON [19]	0.965	0.954	0.974	0.941	0.885	0.965	0.955	0.978	0.942	0.868
WaDIQaM [2]	0.962	0.956	0.969	0.946	0.927	0.965	0.960	0.973	0.957	0.936
TS-CNN [43]	0.959	0.965	<u>0.981</u>	0.970	0.930	0.962	0.970	0.990	0.973	0.943
CaHDC [39]	0.948	<u>0.970</u>	0.978	0.951	0.898	0.953	0.973	0.982	0.955	0.913
HyperIQA [35]	<u>0.965</u>	0.956	0.974	0.963	0.923	0.973	0.962	0.975	<u>0.971</u>	0.935
MMNet [16]	0.968	0.974	0.985	0.935	<u>0.936</u>	/	/	/	/	/
VCRNet [25]	0.964	<u>0.970</u>	0.972	0.955	0.935	<u>0.972</u>	<u>0.984</u>	0.981	0.968	0.951
HPM-Net	0.957	0.974	0.980	<u>0.965</u>	0.947	0.970	0.988	<u>0.987</u>	0.960	<u>0.945</u>

The best and second-best results highlighted in bold and underline, respectively

the marginal benefit of HF-guided modulation. It is worth noting that this behavior is not unique to HPM-Net. As shown in Table 3, several competing BIQA methods also exhibit reduced or less stable performance on Gaussian blur compared with their overall performance across distortion types, indicating that HF-suppressed distortions are generally challenging for quality prediction. Nevertheless, since HPM-Net also integrates QEM-based global quality representations and CLDM-based distortion-aware features, the overall framework does not rely solely on HF cues. HPM-Net still achieves competitive performance on Gaussian blur.

In Table 4, the evaluation of a single distortion type through experiments was carried out using the CSIQ dataset. The comparison was conducted using the same methodological approach, as detailed in Table 3. Based on the findings from the experimental results, it is evident that optimal performance is achieved by the proposed method when applied to five specific types of distortion, Gaussian white noise, Jpeg compression, Jpeg2000 compression, additive Gaussian pink noise, and Gaussian blur. For **Contrast-Curtailed (CC)** distortion, the distinction between HF and low-frequency responses becomes less evident, which may reduce the reliability of HF cues for modulation. Similar trends can also be observed for other BIQA methods, whose performance on CC distortion is generally lower or less consistent than on distortions with clearer HF degradation patterns. Although HPM-Net does not achieve a top-performing SRCC on CC distortion, its PLCC ranks second among the compared methods, indicating that the proposed framework still maintains competitive linear prediction accuracy under CC distortion. This suggests that although HF-guided modulation provides limited gains when HF cues are weakly separable, the complementary QEM and CLDM representations help maintain robust overall prediction performance.

Table 5 presents the performance metrics of the proposed method as evaluated on the TID2013 dataset. As the most challenging comprehensive distortion IQA dataset, TID2013 contains 24 distortion types. The comparison includes traditional methods with recent SOTA methods (e.g., MMNet [16], VCRNet [25]). Upon examination of the table, it is evident that the method in question has demonstrated exceptional performance across 20 distinct categories out of a total of 24 distortion types considered. Among the challenging distortion types, which include impulse noise, quantization noise, and comfort noise, the proposed method is noted to have achieved the highest level of performance when compared to other methods within these specific categories. Specifically, the proposed method outperforms the second best performance in these distortions by 5.1%, 5.9%, and 8.5%, respectively.

Table 4. Performance of Single Distortion Types in the CSIQ Dataset

Method	SRCC						PLCC					
	WN	JPEG	JP2K	PN	BLUR	CC	WN	JPEG	JP2K	PN	BLUR	CC
DIIVINE [24]	0.756	0.732	0.803	0.432	0.785	0.789	0.783	0.758	0.844	0.613	0.836	0.806
BRISQUE [23]	0.723	0.806	0.840	0.378	0.820	0.804	0.742	0.828	0.887	0.496	0.891	0.835
CORNIA [46]	0.664	0.513	0.831	0.493	0.836	0.462	0.687	0.563	0.883	0.632	0.904	0.543
HOSA [41]	0.604	0.733	0.818	0.500	0.841	0.716	0.656	0.759	0.899	0.601	0.912	0.744
CNN [14]	0.919	0.915	0.930	0.900	0.918	0.786	0.945	0.944	0.945	0.834	<u>0.939</u>	0.784
MEON [19]	0.947	<u>0.946</u>	0.896	0.895	0.908	0.807	0.952	0.967	0.921	0.895	<u>0.939</u>	0.814
WaDIQaM [2]	0.947	0.926	0.946	0.879	0.909	0.869	<u>0.956</u>	0.934	0.957	0.886	0.916	0.873
TS-CNN [43]	<u>0.957</u>	0.940	0.941	0.931	0.906	0.818	<u>0.956</u>	0.971	0.947	0.933	0.923	0.807
CaHDC [39]	0.896	0.900	0.936	0.874	0.912	0.872	0.912	0.924	0.943	0.896	0.923	0.879
HyperIQA [35]	0.936	0.945	<u>0.956</u>	0.935	<u>0.922</u>	0.892	0.942	0.946	<u>0.959</u>	0.946	0.924	0.897
MMNet [16]	0.879	0.912	0.932	0.941	0.894	0.942	/	/	/	/	/	/
VCRNet [25]	0.883	0.919	0.931	<u>0.958</u>	0.912	<u>0.938</u>	0.897	<u>0.979</u>	<u>0.959</u>	<u>0.952</u>	0.925	0.935
HPM-Net	0.972	0.967	0.968	0.961	0.960	0.877	0.973	0.993	0.967	0.966	0.957	<u>0.924</u>

The best and second-best results highlighted in bold and underline, respectively

Table 5. Performance of Single Distortion Types in the TID2013 Dataset

SRCC	DIIVINE	BRISQUE	CORNIA	HOSA	RAN4IQA	SGDNet	MMNet	VCRNet	Proposed
Additive Gaussian noise	0.785	0.879	0.655	0.817	0.866	0.668	0.833	0.800	<u>0.874</u>
Additive noise in color components	0.493	0.536	0.164	0.553	<u>0.753</u>	0.579	0.614	0.749	0.784
Spatially correlated noise	0.833	0.782	0.709	<u>0.881</u>	0.842	0.704	0.869	0.875	0.902
Masked noise	0.406	0.260	0.463	0.556	0.462	0.473	0.541	<u>0.627</u>	0.872
HF noise	0.836	0.905	0.851	0.865	<u>0.908</u>	0.818	0.791	0.894	0.930
Impulse noise	0.723	0.839	0.596	0.859	0.855	0.719	0.687	<u>0.912</u>	0.963
Quantization noise	0.652	0.681	0.703	0.679	0.849	0.644	0.767	<u>0.901</u>	0.960
Gaussian blur	0.833	0.842	0.866	0.870	0.833	0.892	<u>0.926</u>	0.903	0.945
Image denoising	0.711	0.565	0.736	0.845	0.839	<u>0.874</u>	0.732	0.921	0.838
Jpeg compression	0.818	0.831	0.805	0.896	<u>0.939</u>	0.885	0.889	0.916	0.960
Jpeg2000 compression	0.816	0.756	0.855	0.901	0.912	0.897	0.889	<u>0.907</u>	0.891
Jpeg transmission errors	0.346	0.278	0.663	0.655	0.566	0.712	0.776	0.795	<u>0.780</u>
Jpeg2000 compression errors	0.817	0.703	0.677	0.683	0.772	0.722	<u>0.916</u>	0.848	0.946
Non-eccentricity pattern noise	0.174	0.264	0.199	0.176	0.234	0.283	0.630	<u>0.668</u>	0.785
Local block-wise distortions	0.366	0.272	0.021	0.269	0.339	0.612	0.546	0.716	0.640
Mean shift	0.106	0.239	0.203	0.128	0.135	0.342	0.392	<u>0.528</u>	0.544
Contrast change	0.132	0.090	0.254	0.138	0.578	<u>0.830</u>	0.887	<u>0.785</u>	0.828
Change of color saturation	0.209	0.214	0.156	0.048	0.484	0.752	<u>0.759</u>	0.728	0.863
Multiplicative Gaussian noise	0.688	0.777	0.583	0.731	0.787	0.515	<u>0.835</u>	0.754	0.842
Comfort noise	0.299	0.258	0.573	0.566	<u>0.819</u>	0.773	0.812	0.816	0.904
Lossy compression of noisy images	0.677	0.573	0.717	0.727	0.895	0.877	0.949	0.830	<u>0.933</u>
Color quantization with dither	0.775	0.782	0.699	0.802	0.822	0.848	0.917	0.702	<u>0.856</u>
Chromatic aberrations	0.732	0.768	0.686	0.721	0.762	0.744	0.798	0.932	<u>0.816</u>
Sparse sampling and reconstruction	0.848	0.857	0.888	0.856	0.917	0.882	<u>0.913</u>	0.769	0.839

The best and second-best results highlighted in bold and underline, respectively

4.4 Cross-Dataset Experiments

To demonstrate the generalizability of the proposed method, a comprehensive set of cross-dataset experiments was conducted, including the LIVE, CSIQ, TID2013, and LIVEC datasets. As depicted in Table 6, the training phase is conducted on a single dataset, followed by testing on the three remaining datasets. Results obtained from the prediction models associated with databases that utilize differing scales of subjective image quality are subjected to a scaling process. Post this scaling, the results are aligned and compared on a uniform scale to facilitate a meaningful comparison. For the comparative analysis, a selection of methods was made, consisting of four traditional BIQA approaches, namely DIIVINE [24], BRISQUE [23], CORNIA [46], and HOSA [41], in addition to seven methods that are predicated on deep learning, including CNN, WaDIQaM [2], TS-CNN [43], RAN4IQA [30], Hall-IQA [17], MMNet [16], and VCRNet [25]. Upon reviewing the table, it is

Table 6. Cross-Dataset Performance Comparisons

Train on	LIVE			TID2013			CSIQ			LIVEC		
Test on	TID2013	CSIQ	LIVEC	LIVE	CSIQ	LIVEC	LIVE	TID2013	LIVEC	LIVE	TID2013	CSIQ
DIIVINE [24]	0.342	0.602	0.296	0.714	0.583	0.235	0.817	0.417	0.366	0.354	0.327	0.419
BRISQUE [23]	0.354	0.573	0.326	0.724	0.568	0.109	0.823	0.433	0.106	0.244	0.275	0.236
CORNIA [46]	0.414	0.619	0.429	0.809	0.659	<u>0.263</u>	0.836	0.332	<u>0.406</u>	0.553	0.385	0.433
HOSA [41]	0.47	0.596	0.455	0.844	0.609	0.244	0.786	0.341	0.277	0.522	0.372	0.334
CNN [14]	0.407	0.616	0.103	0.532	0.598	0.104	0.713	0.315	0.013	0.103	0.021	0.095
WaDIQaM [2]	0.396	0.601	0.151	0.805	0.683	0.009	0.813	0.506	0.106	0.323	0.141	0.323
TS-CNN [43]	0.431	0.621	0.273	0.566	0.613	0.014	0.848	0.484	0.169	0.081	0.049	0.200
RAN4QA [30]	0.466	0.64	0.134	0.811	0.694	0.036	0.806	0.471	0.116	0.296	0.157	0.276
Hall-IQA [17]	0.486	0.668	0.126	0.786	0.683	0.116	0.833	0.491	0.107	/	/	/
MMNet [16]	0.546	0.793	<u>0.502</u>	<u>0.853</u>	<u>0.702</u>	0.348	0.890	0.522	<u>0.406</u>	<u>0.528</u>	0.398	0.518
VCRNet [25]	<u>0.521</u>	0.693	0.566	0.843	0.660	<u>0.263</u>	<u>0.916</u>	<u>0.528</u>	0.422	0.428	<u>0.309</u>	<u>0.553</u>
HPM-Net	0.489	<u>0.712</u>	0.472	0.863	0.731	0.240	0.931	0.576	0.323	0.740	0.401	0.602

The best and second-best results highlighted in bold and underline, respectively

evident that out of the entire set of 12 items subjected to comparison, the proposed method has demonstrated a high level of performance in 7 items. In particular, training on CSIQ and testing on LIVE and TID2013 achieved strong results, outperforming the second best by 1.5% and 4.8%, respectively. These results indicate that the transfer performance of HPM-Net benefits from the complementary design of QEM, CLDM, OCHFM, and HFF.

To further examine whether HF-conditioned modulation contributes to cross-dataset generalization, we conduct an additional diagnostic ablation under representative transfer settings. Specifically, we compare the full HPM-Net with two controlled variants, i.e., “w/o OCHFM” and “Late HF Fusion.” The former removes the HF cue extraction branch, while the latter retains OCHFM but fuses HF features only before the final MLP prediction head, without stage-wise injection into intermediate QEM representations. Across four representative transfer settings, HPM-Net consistently outperforms both variants. The average SRCC increases from 0.713 for “w/o OCHFM” and 0.727 for “Late HF Fusion” to 0.738 for the full HPM-Net. These results suggest that HF cues contribute to transferable quality representation learning and that stage-wise HF-conditioned modulation is more effective than simple late-stage fusion. Nevertheless, the cross-dataset performance should not be attributed solely to HF modulation, but to the joint effect of the QEM, CLDM, OCHFM, and HFF modules.

4.5 Ablation Experiment

Table 7 presents the ablation results used to verify the individual and complementary contributions of the proposed components. Due to space constraints, we report the ablation study on CSIQ and TID2013, which are two representative synthetic distortion datasets with different distortion scales and complexity. Since the full benchmark comparison has already been provided in previous experiments, these two datasets are sufficient to demonstrate the effectiveness of each component while avoiding redundant results.

We further evaluate a variant denoted as “w/o CLDM & OCHFM,” which obtains the lowest performance among all ablation settings. This result indicates that CLDM and OCHFM provide complementary information for quality prediction: CLDM contributes distortion-aware representations, while OCHFM provides perceptually important HF cues for hierarchical modulation. When CLDM is removed, the performance drops by 2.2%/3.1% on CSIQ and 1.3%/1.4% on TID2013. This suggests that the contrastive learning-based distortion representation helps the model better distinguish distortion-related quality variations, especially when different distortion types produce similar visual appearances. When OCHFM is removed, the performance shows average drops of 1.5%/2.7%, demonstrating that HF cues are useful for guiding the QEM to focus on perceptually sensitive

Table 7. Performance Comparison of Different Components

Dataset	CSIQ		TID2013	
Metrics	PLCC	SRCC	PLCC	SRCC
w/o CLDM & OCHFM	0.934	0.910	0.892	0.865
w/o CLDM	0.945	0.931	0.903	0.883
w/o OCHFM	0.951	0.931	0.902	0.875
Late HF Fusion	0.964	0.956	0.911	0.890
Addition	0.957	0.943	0.899	0.882
Vanilla Convolution	0.957	0.943	0.898	0.877
HPM-Net	0.967	0.962	0.916	0.897

The best results highlighted in bold.

local details, such as edges, textures, and distortion artifacts. To further distinguish hierarchical HF modulation from standard late-stage feature fusion, we introduce a controlled variant named “Late HF Fusion.” In this variant, OCHFM is retained to extract HF features, but the HF features are not injected into the intermediate QEM representations through HFF. Instead, they are fused only once before the final MLP prediction head. As shown in Table 7, Late HF Fusion outperforms the “w/o OCHFM” variant but remains lower than the full HPM-Net. This comparison indicates that HF cues themselves are beneficial for BIQA, while simply introducing them as late-stage auxiliary information is insufficient to achieve the best performance. In contrast, injecting HF cues into multiple hierarchical stages enables them to progressively reweight and refine quality-aware representations, leading to more effective quality prediction. The term “Addition” denotes the variant where the proposed AFF is replaced by a simple addition operation, while “Vanilla Convolution” refers to replacing octave convolution with standard convolution in the HF extraction stream. Compared with these two variants, HPM-Net achieves consistent improvements on both CSIQ and TID2013, with average absolute gains of 1.4%/1.7% and 1.4%/2.0% in terms of PLCC/SRCC. These results indicate that attention-based fusion is more effective than treating different feature streams equally and that octave convolution better separates and models frequency-aware representations than standard convolution. Overall, these results confirm that CLDM, OCHFM, octave convolution, and AFF all contribute to the final performance of HPM-Net.

4.6 Group MAXimum Differentiation (gMAD) Experiment

To further evaluate the performance of the proposed method, the gMAD competition was conducted [20]. gMAD is based on the idea that an attempt is made to disprove a model, and a model that is more difficult to disprove is considered to be a relatively better model. gMAD utilizes an attacker model to search for counterexamples in the database. These counterexamples are optimal for the attacker model, and if the attack is successful, the defender model is disproved. If, on the contrary, the defender survives such an attack, this indicates that it may be a robust and reliable model. Figure 4 illustrates the results of our experiments. A comparative analysis was performed, positioning it against the current SOTA model known as VCRNet [25]. In Figure 4, panels (a) and (b) denote the image pairs selected by our proposed model as the defender and VCRNet as the attacker in two sets of images at high and low levels, respectively. Panels (c) and (d) show the cases where our proposed model acts as the attacker and VCRNet acts as the defender. As depicted in Figure 4, it can be observed that the image selected by the attacker undergoes minimal perceptual change when the proposed method acts as the defender. On the contrary, when the proposed method functions in the capacity of an attacker, image pairs with significantly larger differences are efficiently recognized and selected. This reflects the robustness of our method.

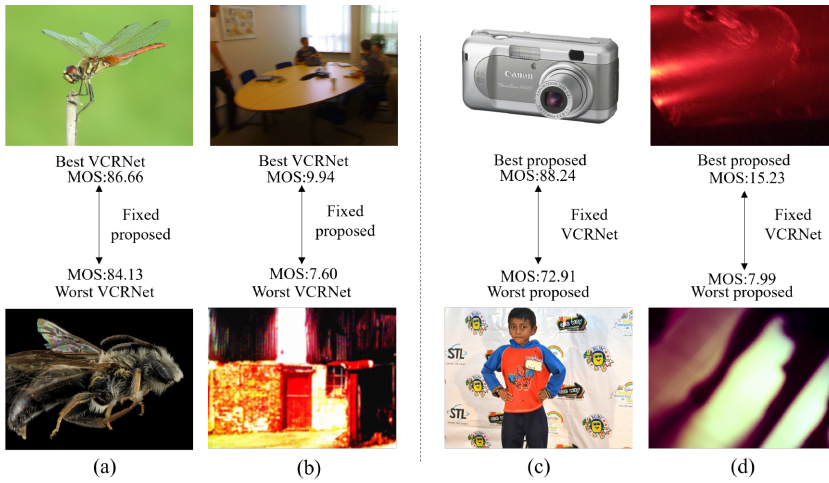


Fig. 4. The gMAD competition results between VCRNet and our proposed method. (a) Fixed proposed at the high-quality level. (b) Fixed proposed at the low-quality level. (c) Fixed VCRNet at the high-quality level. (d) Fixed VCRNet at the low-quality level. To better compare the differences between the two images, the ground truth (MOS) of the images has been shown. gMAD, Group MAXimum Differentiation.

Table 8. Computational Complexity Comparison with Representative BIQA Methods

Method	Params (M)	FLOPs (G)	FPS
WaDIQaM [2]	5.24	2.43	122.40
MetaIQA [54]	13.24	1.82	/
HyperIQA [35]	27.38	4.31	33.36
TReS [9]	152.45	20.03	3.98
TOPIQ [3]	36.04	/	45.17
Q-Align [38]	8.2×10^3	/	22.94
HPM-Net	51.33	9.57	27.00

The best results highlighted in bold.

4.7 Computational Complexity Analysis

To further evaluate the efficiency of the proposed method, we compare HPM-Net with representative BIQA methods in terms of the number of parameters, computational cost, and inference speed. The results are reported in Table 8. Although the proposed model introduces the OCHFM, CLDM, and HFF modules to model hierarchical perceptual modulation, its computational cost remains acceptable compared with existing BIQA models. Compared with TReS [9], HPM-Net is much more compact in terms of both parameters and computational cost and also achieves a faster inference speed. Compared with lightweight methods such as WaDIQaM [2], MetaIQA [54], and HyperIQA [35], HPM-Net has a higher model complexity, but it provides stronger quality prediction performance across multiple benchmarks. Compared with TOPIQ [3], HPM-Net has more parameters and a lower FPS, but achieves competitive or superior performance on the evaluated datasets. In addition, compared with the large multimodal model-based Q-Align [38], HPM-Net is significantly more compact in terms of parameter size and achieves a higher FPS, while still maintaining competitive performance, especially on synthetic distortion datasets. These results demonstrate that HPM-Net achieves a favorable tradeoff between prediction accuracy and computational efficiency.

5 Conclusion

In this article, we identified a key limitation in current BIQA: the lack of a framework that uses HF cues as perceptually motivated conditioning signals for hierarchical quality representation learning. To address this, we introduced HPM-Net, a bio-inspired framework that treats HF cues as modulators guiding representation learning across layers. Concretely, OCHFM extracts and hierarchically injects HF cues into the backbone quality estimation stream; CLDM learns distortion-aware representations and is integrated with the modulated features via AFF to improve generalization; and HFF performs modulation-aware integration, adaptively combining HF cues with globally contextualized quality representations to yield coherent quality estimates. Extensive experiments on five BIQA datasets show consistent gains over SOTA methods, demonstrating both effectiveness and robustness across synthetic and authentic distortions. Despite its effectiveness, HPM-Net still has two limitations. First, HF cues may be less informative for distortion types where HF responses are weakened or less distinguishable, such as blur-related and contrast-related distortions. In such cases, the benefit of HF-guided modulation may become less pronounced. Second, the introduction of OCHFM, CLDM, and HFF increases computational overhead compared with lightweight BIQA models. In future work, we will explore adaptive frequency modeling to better handle distortions with weak HF responses and lightweight network designs to reduce computational cost.

References

- [1] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. 2024. Arnika: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 189–198.
- [2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing* 27, 1 (2017), 206–219.
- [3] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2024. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing* 33 (2024), 2404–2418.
- [4] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. 2019. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3435–3444.
- [5] Russell L. DeValois and Karen K. DeValois. 1990. *Spatial Vision*. Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780195066579.001.0001>
- [6] Yixuan Gao, Xiongkuo Min, Yuqin Cao, Weisi Lin, Bu Sung Lee, and Guangtao Zhai. 2025. Blind image quality assessment by Gaussian mixture distribution. *IEEE Transactions on Image Processing* 34 (2025), 4660–4675.
- [7] Yixuan Gao, Xiongkuo Min, Yucheng Zhu, Xiao-Ping Zhang, and Guangtao Zhai. 2024. Blind image quality assessment: A fuzzy neural network for opinion score distribution prediction. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 3 (2024), 1641–1655.
- [8] Deepti Ghadiyaram and Alan C. Bovik. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing* 25, 1 (2015), 372–387.
- [9] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1220–1230.
- [10] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. 2023. Visual attention network. *Computational Visual Media* 9, 4 (2023), 733–752.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034.
- [12] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* 29 (2020), 4041–4056.
- [13] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*. PMLR, 448–456.

- [14] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1733–1740.
- [15] Eric C. Larson and Damon M. Chandler. 2010. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging* 19, 1 (2010), 011006–011006.
- [16] Fan Li, Yangfan Zhang, and Pamela C. Cosman. 2021. MMMNet: An end-to-end multi-task deep convolution neural network with multi-scale and multi-hierarchy fusion for blind image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 12 (2021), 4798–4811.
- [17] Kwan-Yee Lin and Guanxiang Wang. 2018. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 732–741.
- [18] Xialei Liu, Joost Van De Weijer, and Andrew D. Bagdanov. 2017. Rankiq: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, 1040–1049.
- [19] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. 2017. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing* 27, 3 (2017), 1202–1213.
- [20] Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. 2016. Group mad competition—A new methodology to compare objective image quality models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1664–1673.
- [21] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. 2022. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing* 31 (2022), 4149–4161.
- [22] Xiongkuo Min, Yixuan Gao, Yuqin Cao, Guangtao Zhai, Wenjun Zhang, Huifang Sun, and Chang Wen Chen. 2025. Exploring rich subjective quality information for image quality assessment in the wild. *IEEE Transactions on Circuits and Systems for Video Technology* 35, 8 (2025), 7778–7791.
- [23] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial doma. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708.
- [24] Anush Krishna Moorthy and Alan Conrad Bovik. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing* 20, 12 (2011), 3350–3364.
- [25] Zhaoqing Pan, Feng Yuan, Jianjun Lei, Yuming Fang, Xiao Shao, and Sam Kwong. 2022. VCRNet: Visual compensation restoration network for no-reference image quality assessment. *IEEE Transactions on Image Processing* 31 (2022), 1613–1627.
- [26] Zhaoqing Pan, Feng Yuan, Xu Wang, Long Xu, Xiao Shao, and Sam Kwong. 2022. No-reference image quality assessment via multibranch convolutional neural networks. *IEEE Transactions on Artificial Intelligence* 4, 1 (2022), 148–160.
- [27] Zhaoqing Pan, Hao Zhang, Jianjun Lei, Yuming Fang, Xiao Shao, Nam Ling, and Sam Kwong. 2022. DACNN: Blind image quality assessment via a distortion-aware convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 11 (2022), 7518–7531.
- [28] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* 30 (2015), 57–77.
- [29] Marie-Therese Puth, Markus Neuhäuser, and Graeme D. Ruxton. 2015. Effective use of Spearman’s and Kendall’s correlation coefficients for association between two measured traits. *Animal Behaviour* 102 (2015), 77–84.
- [30] Hongyu Ren, Diqi Chen, and Yizhou Wang. 2018. RAN4IQA: Restorative adversarial nets for no-reference image quality assessment. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.
- [31] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT doma. *IEEE Transactions on Image Processing* 21, 8 (2012), 3339–3352.
- [32] Avinab Saha, Sandeep Mishra, and Alan C. Bovik. 2023. Re-IQA: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5846–5855.
- [33] Philip Sedgwick. 2012. Pearson’s correlation coefficient. *BMJ* 345 (2012), e4483.
- [34] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing* 15, 11 (2006), 3440–3451.
- [35] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3667–3676.
- [36] Li Chien Tan, Haniza Yazid, and Yen Fook Chong. 2019. Image quality assessment (iqa) using high-frequency and image variance (hfiv) for colour image. *Journal of Physics: Conference Series* 1372 (2019), 012034.
- [37] Zheyin Wang, Liqian Shen, Zhengyong Wang, Yufei Lin, and Yanliang Jin. 2022. Generation-based joint luminance-chrominance learning for underwater image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 3 (2022), 1123–1139.

- [38] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. 2024. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research)*. PMLR.
- [39] Jinjian Wu, Jupo Ma, Fuhu Liang, Weisheng Dong, Guangming Shi, and Weisi Lin. 2020. End-to-end blind image quality prediction with cascaded deep neural network. *IEEE Transactions on Image Processing* 29 (2020), 7414–7426.
- [40] Xu Wu, Zhihui Lai, Jie Zhou, Xianxu Hou, Witold Pedrycz, and Linlin Shen. 2024. Light-aware contrastive learning for low-light image enhancement. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 9 (2024), 1–24.
- [41] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. 2016. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing* 25, 9 (2016), 4444–4457.
- [42] Chenggang Yan, Tong Teng, Yutao Liu, Yongbing Zhang, Haoqian Wang, and Xiangyang Ji. 2021. Precise no-reference image quality evaluation based on distortion identification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 3s (2021), 1–21.
- [43] Qingsen Yan, Dong Gong, and Yanning Zhang. 2018. Two-stream convolutional networks for blind image quality assessment. *IEEE Transactions on Image Processing* 28, 5 (2018), 2200–2211.
- [44] Jifan Yang, Zhongyuan Wang, Baojin Huang, Jiaxin Ai, Yuhong Yang, and Zixiang Xiong. 2024. Joint distortion restoration and quality feature learning for no-reference image quality assessment. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 7 (2024), 1–20.
- [45] Jifan Yang, Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Yuhong Yang, and Weiping Tu. 2024. Auxiliary information guided self-attention for image quality assessment. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 4 (2024), 1–23.
- [46] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1098–1105.
- [47] Junyong You and Jari Korhonen. 2021. Transformer for image quality assessment. In *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*, 1389–1393.
- [48] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. 2024. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *Proceedings of the European Conference on Computer Vision*. Springer, 259–276.
- [49] Hui Zeng, Lei Zhang, and Alan C. Bovik. 2018. Blind image quality assessment with a probabilistic quality representation. In *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, 609–613.
- [50] Chaofan Zhang, Ziqing Huang, Shiguang Liu, and Jian Xiao. 2022. Dual-channel multi-task CNN for no-reference screen content image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 8 (2022), 5011–5025.
- [51] Youzhi Zhang, Junming Jiang, Deyang Liu, Xiaofei Zhou, and Caifeng Shan. 2025. Multi-dimension attention network for full-reference light field image quality assessment. *IEEE Transactions on Instrumentation and Measurement* 74 (2025), 1–14.
- [52] Youzhi Zhang, Lifei Wan, Deyang Liu, Xiaofei Zhou, Ping An, and Caifeng Shan. 2024. Saliency-guided no-reference omnidirectional image quality assessment via scene content perceiving. *IEEE Transactions on Instrumentation and Measurement* 73 (2024), 1–15.
- [53] Zicheng Zhang, Wei Sun, Yingjie Zhou, Jun Jia, Zhichao Zhang, Jing Liu, Xiongkuo Min, and Guangtao Zhai. 2023. Subjective and objective quality assessment for in-the-wild computer graphics images. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 4 (2023), 1–22.
- [54] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. 2020. MetaQA: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14143–14152.
- [55] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. 2021. Generalizable no-reference image quality assessment via deep meta-learning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2021), 1048–1060.

Received 9 December 2025; revised 7 June 2026; accepted 10 August 2026