Multitasks-based Deep Evidential Fusion Network for Blind Image Quality Assessment

Yiwei Lou[1], Yuanpeng He[1], Rongchao Zhang[1], Yongzhi Cao[1], Hanpin Wang[1], Yu Huang[2]*

[1]Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education; School of Computer Science, Peking University, Beijing, China
[2]National Engineering Research Center for Software Engineering, Peking University, Beijing, China
hy@pku.edu.cn

Abstract

Blind image quality assessment (BIQA) methods often incorporate auxiliary tasks to improve performance. However, existing approaches face limitations due to insufficient integration and a lack of flexible uncertainty estimation, leading to suboptimal performance. To address these challenges, we propose a multitasks-based **D**eep **E**vidential **F**usion **Net**work (DEFNet) for BIQA, which performs multitask optimization with the assistance of scene and distortion type classification tasks. To achieve a more robust and reliable representation, we design a novel trustworthy information fusion strategy. It first combines diverse features and patterns across subregions to enhance information richness, and then performs local-global information fusion by balancing fine-grained details with coarse-grained context. Moreover, DEFNet exploits advanced uncertainty estimation technique inspired by evidential learning with the help of normal-inverse gamma distribution mixture. Extensive experiments on both synthetic and authentic distortion datasets demonstrate the effectiveness and robustness of the proposed framework. Additional evaluation and analysis are carried out to highlight its strong generalization capability and adaptability to previously unseen scenarios.

Code — https://github.com/cyfqylyw/DEFNet
Extended version — https://arxiv.org/abs/2507.19418

Introduction

Blind image quality assessment (BIQA) is a pivotal area in the field of image processing. Its primary goal is to objectively and consistently assess the quality of images without relying on reference images for comparison. The pursuit of more accurate and efficient methods for BIQA helps to improve the overall quality of experience for end-users (Zhao et al. 2023; Lou et al. 2023). This technique is of great importance in a wide range of application areas, such as real-time multimedia processing (Luo et al. 2024; Guo et al. 2025), medical image analysis (Lou et al. 2024a,b; Zhang et al. 2024) and other scientific fields (Zhang et al. 2025a,b).

Over time, BIQA approaches have undergone a significant evolution from early techniques based on hand-crafted feature extraction and manual characterization (Mittal, Soundararajan, and Bovik 2013; Mittal, Moorthy, and Bovik 2012) to more sophisticated data-driven and deep learning-based approaches (Saha, Mishra, and Bovik 2023; Zhao et al. 2023; Zheng et al. 2024; Chen et al. 2024; Zhou et al. 2025; Shen et al. 2025). Nonetheless, these methods primarily focus on the assessment of image quality, which limits the assistance of auxiliary tasks and information. Inspired by this, efforts have been made to incorporate auxiliary tasks and information in a multitask learning manner, as shown in Figure 1. For instance, scene statistics (Yan, Bare, and Tan 2019) and image content (Li et al. 2022; Zhou et al. 2025) offer valuable contextual information that can significantly influence the quality perception. Besides, distortion type classification (Madhusudana et al. 2022; Zhou et al. 2024) and spatial angular estimation (Qu et al. 2021) provide inspiration as auxiliary tasks that enable more accurate assessment of image quality. These methods typically utilize image content (scene information) and artifact categories (distortion information) to provide complementary insights and knowledge.

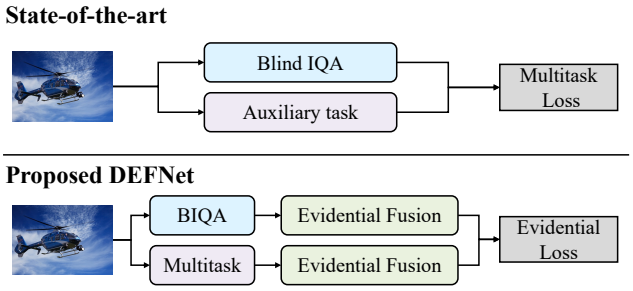


Figure 1: Comparison between the proposed DEFNet and state-of-the-art methods that utilize auxiliary tasks to assist in BIQA. We propose to include evidential fusion within each task for higher performance and lower uncertainty.

Despite these advances, existing methods still face challenges in two major aspects. **(i) In-depth information fusion.** On one hand, this requires ❶ *inter-task information integration*. Some existing approaches treat auxiliary tasks as independent modules, leading to information fragmentation

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

and a lack of in-depth mining of potential inter-task correlations. On the other hand, it necessitates ❷ *multilevel and cross-region feature fusion*, which involves full considering of the complex interactions between features and exploring diverse sub-regions that may contain different distortion patterns and visual characteristics. **(ii) Comprehensive uncertainty estimation.** Though significant progress (Zhang et al. 2021; Gao et al. 2023) has been made in uncertainty estimation for BIQA, it is still difficult to provide a ❸ *flexible and robust uncertainty representation*. A key limitation is the inability to simultaneously model both aleatoric and epistemic uncertainty, which often results in overconfident predictions even when the predictions are not correct.

To address these challenges, we propose a multitasks-based **D**eep **E**vidential **F**usion **Net**work (DEFNet) in this paper. Our framework integrates three core tasks: BIQA, scene classification, and distortion type classification. It starts by utilizing contrastive language-image pre-training (Radford et al. 2021) to extract both local and global image features across the three different tasks, followed by a simultaneous multitask optimization to tackle challenge ❶. To further enhance feature fusion, we introduce a trustworthy information fusion strategy operating at two levels: cross sub-region and local-global. The cross sub-region fusion aggregates diverse features and patterns from different image sub-regions, thereby enhancing the information richness and ensuring accurate capture of regional quality. Meanwhile, the local-global fusion combines insights from both fine-grained details and coarse-grained context, providing a holistic understanding of image quality. This multilevel strategy facilitates in-depth information fusion and cross-region interactions, which serves as a solution to challenge ❷. Furthermore, to address challenge ❸, DEFNet incorporates a robust uncertainty estimation mechanism inspired by evidence theory (Amini et al. 2020). By utilizing the four dimensions of the data distribution and the mixture of normal-inverse gamma distribution, this approach simultaneously captures both aleatoric and epistemic uncertainty, enabling the model to identify the predictive fluctuations. As a result, the proposed DEFNet achieves high adaptability and generalization capabilities in various experimental settings.

The main contributions of this paper are summarized as follows:

- We propose a novel multitask-based deep evidential fusion network for BIQA, which integrates scene classification and distortion type classification to enhance inter-task information fusion.
- We propose a two-level trustworthy information fusion strategy, including cross sub-region and local-global information fusion, which integrate cross-region and cross-grained features, respectively.
- We develop a robust uncertainty estimation mechanism based on evidential learning and normal-inverse gamma distribution mixture, thereby improving the model's performance and adaptability.
- Extensive experiments on both synthetic and authentic distortion datasets are carried out to demonstrate that DEFNet achieves state-of-the-art performance, as well as strong generalization ability.

Problem Statement and Preliminaries

To formalize the problem of blind image quality assessment, denote $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$ as a pristine or distorted image, where C, H, W are the channel number, height, and width, respectively. The goal of BIQA is to train a function $f : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}$ and estimate a quality score $q \in \mathbb{R}$ for the image $\boldsymbol{x}$ that reflects its perceptual quality, ideally aligning with human subjective evaluations.

Viewing the field of BIQA from the perspective of evidential learning, assume that the quality score q of each image is subject to a normal distribution $q \sim \mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are the unknown mean and variance. The posterior distribution $p(\mu, \sigma|q(\boldsymbol{x}_{1\ldots N}))$ is assumed to follow a normal-inverse gamma (NIG) distribution $(\mu, \sigma) \sim \mathrm{NIG}(\delta, v, \alpha, \beta)$, that is $\mu \sim \mathcal{N}(\delta, \sigma^2 v^{-1})$ and $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$, where $\Gamma(\cdot)$ is gamma function, $\mathbf{m} = (\delta, v, \alpha, \beta)$ are distribution parameters with constraints $\delta \in \mathbb{R}, v > 0, \alpha > 1, \beta > 0$. To increase the model evidence, denote $\Omega = 2\beta(1 + v)$, the negative logarithm of model evidence is denoted as:

$$\begin{aligned}\ell^{NLL}(\boldsymbol{x}, \boldsymbol{y}, \theta) = \frac{1}{2}\log\left(\frac{\pi}{v}\right) + \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}\right) \\ -\alpha_t \log(\Omega) + (\alpha + \frac{1}{2})\log\left((\boldsymbol{y}-\delta)^2 v + \Omega\right),\end{aligned} \tag{1}$$

where $\boldsymbol{x}, \boldsymbol{y}$ are the input data and the ground-truth label. To realign confidence in the predictions by reducing the evidence weight for predictions that deviate from expected values, the regression loss is defined as:

$$\ell^R(\boldsymbol{x}, \boldsymbol{y}, \theta) = |\boldsymbol{y} - \mathbb{E}(\mu)| \cdot \phi, \tag{2}$$

where $\phi = 2v + \alpha$ is the total evidence (Amini et al. 2020). This realignment helps to improve the predictive acumen of the model, creating a more rigorous and robust framework for estimating the reasonableness of regression. The total evidential loss aims to combine the term maximizing the model fit and the term minimizing evidence on errors:

$$\ell^U(\boldsymbol{x}, \boldsymbol{y}, \theta) = \ell^{NLL}(\boldsymbol{x}, \boldsymbol{y}, \theta) + \tau\ell^R(\boldsymbol{x}, \boldsymbol{y}, \theta), \tag{3}$$

where τ is the weights keeping the balance between model fitting and uncertainty inflation.

Methodology

This section introduces the proposed DEFNet framework based on multitasks for BIQA. As shown in Figure 2, the proposed framework initiates by extracting feature embeddings and probability scores from both local and global image context, and then performs single task optimization, as well as two levels (cross sub-region and local-global) of evidential fusion across all the three tasks.

Local and Global Probability Scores

In the proposed DEFNet framework, we employ contrastive language-image pre-training (CLIP) (Radford et al. 2021) to extract the feature embeddings and compute both local

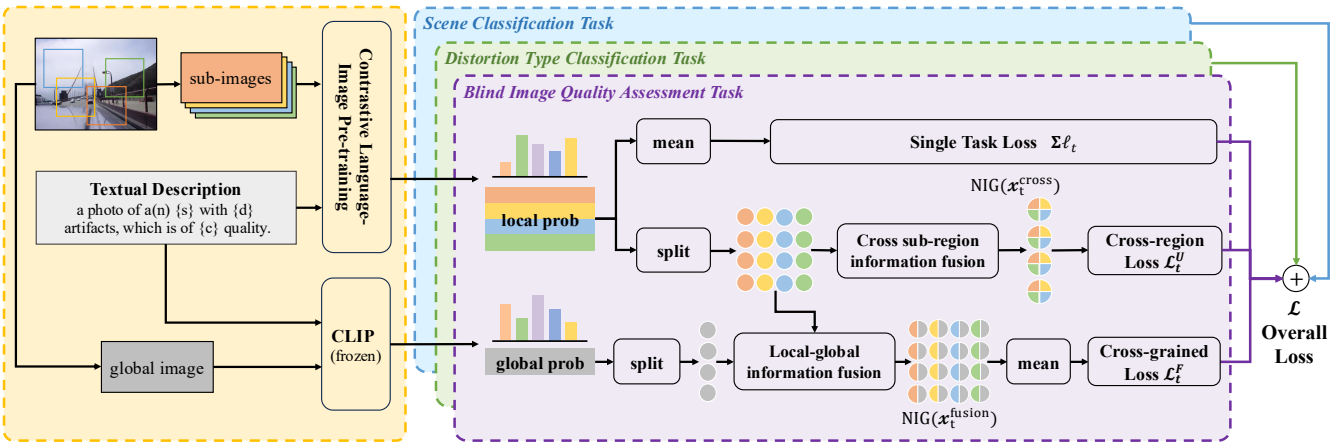


Figure 2: Overview of the proposed DEFNet framework.

and global probability scores. Specifically, the CLIP architecture consists of separate image and text encoders, which are trained by feeding multiple images and corresponding textual description (*"a photo of a(n) {s} with {d} artifacts, which is of {c} quality."*), respectively.

Considering the prerequisite of the image encoder for inputs of a consistent size, local sub-images are obtained through cropping operation, while the global image is acquired after downsampling operation. This image segmentation approach allows to balance detail-oriented local analysis with a broader global perspective. It is worth mentioning that the training process of CLIP in the proposed DEFNet framework only consists of vision-language information pairs for local sub-images. The global feature embeddings are derived using the CLIP model pre-trained on these sub-images, with its parameters frozen to ensure stability and consistency in feature representation. From this, we have the correspondence score $\text{logit}(c, s, d|\boldsymbol{x})$. Subsequently, DEFNet performs softmax activation to derive the joint probability

$$\hat{p}(c, s, d)(\boldsymbol{x}) = \frac{\exp(\text{logit}(c, s, d|\boldsymbol{x})/\kappa)}{\sum_{c,s,d} \exp(\text{logit}(c, s, d|\boldsymbol{x})/\kappa)}, \quad (4)$$

where κ is a learnable parameter, c, s, d indicate the quality class, scene and distortion type, respectively. After that, the local probability scores $\hat{p}(c, s, d|\boldsymbol{x}^{\text{local}})$ and global scores $\hat{p}(c, s, d|\boldsymbol{x}^{\text{global}})$ are derived.

With the assistant of the local probability scores, the quality score of an image is further estimated as:

$$\hat{q}(\boldsymbol{x}) = \sum_{c=1}^{C} \hat{p}(c|\boldsymbol{x}) \times c, \quad (5)$$

where $C = 5$ and $c \in \mathcal{C} = \{1, 2, 3, 4, 5\}$ indicates the quality level from bad to perfect, and the estimated probability of the quality level is calculated by aggregating all the local scores

$$\hat{p}(c|\boldsymbol{x}) = \text{AVG}_{i=1}^{N} \left(\sum_{s\in\mathcal{S}, d\in\mathcal{D}} \hat{p}(c, s, d)(\boldsymbol{x}_i^{\text{local}}) \right), \quad (6)$$

where N is the number of sub-images, $\text{AVG}(\cdot)$ is averaging operation for the local scores.

Multitask Optimization

In the multitask optimization framework, BIQA is the primary task represented by the loss component ℓ_q, while components ℓ_s and ℓ_d correspond to the auxiliary tasks of scene and distortion type classification, respectively. Each loss component, with specific definition as follows, contributes uniquely to the overall multitask loss.

By adopting the fidelity loss (Tsai et al. 2007), the BIQA loss for image pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is defined as:

$$\begin{aligned} \ell_q(\boldsymbol{x}_1, \boldsymbol{x}_2; \theta) =& 1 - \sqrt{p(\boldsymbol{x}_1, \boldsymbol{x}_2)\hat{p}(\boldsymbol{x}_1, \boldsymbol{x}_2)} \\ & - \sqrt{(1 - p(\boldsymbol{x}_1, \boldsymbol{x}_2))(1 - \hat{p}(\boldsymbol{x}_1, \boldsymbol{x}_2))}, \end{aligned} \quad (7)$$

where

$$\hat{p}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \Phi\left(\frac{\hat{q}(\boldsymbol{x}_1) - \hat{q}(\boldsymbol{x}_2)}{\sqrt{2}} \right) \quad (8)$$

quantifies the likelihood that $\boldsymbol{x}_1$ is of higher predicted quality than $\boldsymbol{x}_2$ using standard Normal cumulative distribution function $\Phi(\cdot)$ under the Thurstone's model (Thurstone 2017), and $p(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is a binary label indicating whether the ground-truth MOS $q(\boldsymbol{x}_1) \geq q(\boldsymbol{x}_2)$.

In the settings of DEFNet, an image can be associated with multiple scene categories. Given an image $\boldsymbol{x}$, the estimated probability of a scene s is calculated by aggregating the joint probabilities across all possible quality and distortion combinations:

$$\hat{p}(s|\boldsymbol{x}) = \sum_{c,d} \hat{p}(c, s, d|\boldsymbol{x}), \quad (9)$$

where $\hat{p}(c, s, d|\boldsymbol{x})$ is the joint probability derive in Equation (4). Based on this, the scene classification loss component is defined as:

$$\begin{aligned} \ell_s(\boldsymbol{x}; \theta) = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \Big(& 1 - \sqrt{p(s|\boldsymbol{x})\hat{p}(s|\boldsymbol{x})} \\ & - \sqrt{(1 - p(s|\boldsymbol{x}))(1 - \hat{p}(s|\boldsymbol{x}))} \Big), \end{aligned} \quad (10)$$

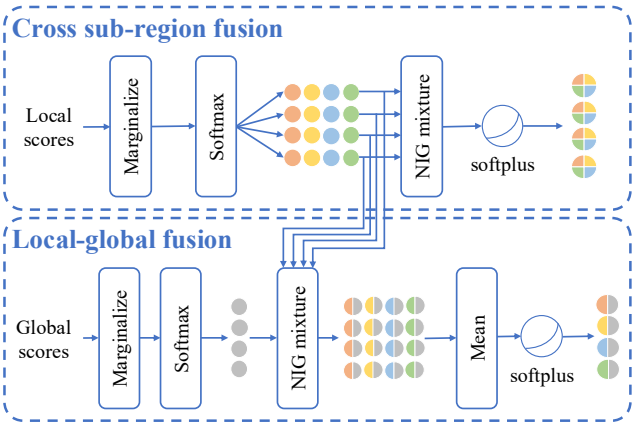


Figure 3: Overview of the cross sub-region information fusion (top) and local-global information fusion (bottom).

where $\mathcal{S}$ is the set of all possible scene categories, $p(s|\boldsymbol{x})$ is a binary label indicating whether the image $\boldsymbol{x}$ falls in ground-truth scene category s.

Similar but different, we assume each image only belongs to one dominant distortion type, and we have the predicted probability for specific distortion type d as:

$$\hat{p}(d|\boldsymbol{x}) = \sum_{c,s} \hat{p}(c, s, d|\boldsymbol{x}), \tag{11}$$

and further define the distortion type classification loss as:

$$\ell_d(\boldsymbol{x}; \theta) = 1 - \sum_{d \in \mathcal{D}} \sqrt{p(d|\boldsymbol{x})\hat{p}(d|\boldsymbol{x})}, \tag{12}$$

where $\mathcal{D}$ is the set of all possible distortion types, $p(d|\boldsymbol{x})$ is the binary ground-truth label, and $\hat{p}(d|\boldsymbol{x})$ is the predicted probability for image $\boldsymbol{x}$ belongs to type d.

Utilizing auxiliary tasks, DEFNet optimizes losses of the three separate tasks, integrating them into the multitask loss (Zhang et al. 2023), which in a mini-batch $\mathcal{B}$ is defined as

$$\begin{aligned}\mathcal{L}^M(\theta) = &\frac{1}{|\mathcal{P}|} \sum_{(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{P}} \lambda_q \ell_q(\boldsymbol{x}_1, \boldsymbol{x}_2; \theta) \\ &+ \frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x} \in \mathcal{B}} \left[\lambda_s \ell_s(\boldsymbol{x}) + \lambda_d \ell_d(\boldsymbol{x})\right],\end{aligned} \tag{13}$$

where θ is the model parameter, $\mathcal{P}$ denotes the set of all possible image pairs with ground-truth quality label, $\lambda_q, \lambda_s, \lambda_d$ are weights updated with the relative descending rate (Liu, Johns, and Davison 2019).

Cross Sub-region Information Fusion

In this section, we introduce the technique of cross sub-region evidential information fusion. As shown in Figure 3, it integrates fragmented information from different sub-regions, allowing the model to make predictions in a more comprehensive way and decrease aleatoric and epistemic uncertainty. Through fusion across sub-regions, the DEFNet framework integrates diverse features and patterns from different regions effectively, which is critical in dealing with complex images and diverse content. This helps to capture the differences in quality across sub-regions in an image more accurately, thus improving the accuracy of the assessment at a detailed level.

To estimate the parameters of NIG distribution, we randomly sample probability scores $\hat{p}(c, s, d|\boldsymbol{x}_i^{\text{local}})$ of four sub-images $i \in \{1, 2, 3, 4\}$. Subsequently, we marginalize the probability scores for the three specific tasks (q for BIQA task, s for scene classification task, and d for distortion type classification task):

$$\boldsymbol{x}_{q,i}^{\text{local}} = \text{softplus}\left(\sum_{c=1}^{C}\left[\sum_{s,d} \hat{p}(c, s, d|\boldsymbol{x}_i^{\text{local}}) \times c\right]\right), \tag{14}$$

$$\boldsymbol{x}_{s,i}^{\text{local}} = \text{softplus}\left(\sum_{q,d} \hat{p}(c, s, d|\boldsymbol{x}_i^{\text{local}})\right), \tag{15}$$

$$\boldsymbol{x}_{d,i}^{\text{local}} = \text{softplus}\left(\sum_{q,s} \hat{p}(c, s, d|\boldsymbol{x}_i^{\text{local}})\right), \tag{16}$$

where softplus is the activation function to satisfy parameter constraints, $C = 5$ is the number of quality levels. The distribution parameters can be computed as follows:

$$\begin{aligned}\boldsymbol{m}_{t,i}^{\text{local}} &= (\boldsymbol{x}_{t,i}^{\text{local}})_\delta, (\boldsymbol{x}_{t,i}^{\text{local}})_v, (\boldsymbol{x}_{t,i}^{\text{local}})_\alpha, (\boldsymbol{x}_{t,i}^{\text{local}})_\beta \\ &= \text{split}(\boldsymbol{x}_{t,i}^{\text{local}}),\end{aligned} \tag{17}$$

where $t \in \{q, s, d\}$ denote the task domain. Then, we adopt the fusion strategy to fuse multiple NIG distribution and to integrate the inter sub-region information extracted from the four sub-images:

$$\begin{aligned}\text{NIG}(\boldsymbol{x}_t^{\text{cross}}) = &\text{NIG}(\boldsymbol{m}_{t,1}^{\text{local}}) \oplus \text{NIG}(\boldsymbol{m}_{t,2}^{\text{local}}) \\ &\oplus \text{NIG}(\boldsymbol{m}_{t,3}^{\text{local}}) \oplus \text{NIG}(\boldsymbol{m}_{t,4}^{\text{local}}),\end{aligned} \tag{18}$$

where $\boldsymbol{x}_t^{\text{cross}}$ is the mixture NIG distribution parameters, $\oplus$ is the summation operation for two NIG distributions (Ma et al. 2021), which is defined as:

$$\text{NIG}(\delta, v, \alpha, \beta) \triangleq \text{NIG}(\delta_1, v_1, \alpha_1, \beta_1) \oplus \text{NIG}(\delta_2, v_2, \alpha_2, \beta_2) \tag{19}$$

where

$$\begin{aligned}&\delta = (v_1 + v_2)^{-1}(v_1\delta_1 + v_2\delta_2), \\ &v = v_1 + v_2, \quad \alpha = \alpha_1 + \alpha_2 + \frac{1}{2}, \\ &\beta = \beta_1 + \beta_2 + \frac{1}{2}v_1(\delta_1 - \delta)^2 + \frac{1}{2}v_2(\delta_2 - \delta)^2.\end{aligned} \tag{20}$$

Then, we compute the evidential loss on local outputs for single task t:

$$\mathcal{L}_t^U(\theta) = \frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x} \in \mathcal{B}} \ell^U\left(\text{softplus}(\boldsymbol{x}_t^{\text{cross}}), \boldsymbol{y}_t, \theta\right), \tag{21}$$

where $\boldsymbol{y}_q = q(\boldsymbol{x})$ is the ground-truth MOS, $\boldsymbol{y}_s = p(s|\boldsymbol{x})$ and $\boldsymbol{y}_d = p(d|\boldsymbol{x})$ are binary labels indicating whether the image $\boldsymbol{x}$ falls in ground-truth scene category s and distortion type category d, respectively. Then, the overall cross-region loss is defined as the sum of evidential loss for the three tasks:

$$\mathcal{L}^U(\theta) = \mathcal{L}_q^U(\theta) + \mathcal{L}_s^U(\theta) + \mathcal{L}_d^U(\theta). \tag{22}$$

Local-global Information Fusion

In this section, we describe the evidential fusion between local and global information, the overall framework is shown in Figure 3. Local information focuses on fine-grained details within sub-images, while global information provides a coarse-grained perspective of the entire image. The local-global fusion allows them to complement each other effectively and enables DEFNet to combine the local view at a detailed level with a broader global view, providing a comprehensive assessment of image quality. This fusion strategy balances the fine-grained details with the coarse-grained whole, ensuring that DEFNet is neither overly focused on micro-details nor ignoring global perspectives.

To combine information from local sub-images and global downsampled image, we first marginalize the global probability scores $\boldsymbol{x}_t^{\text{global}}$ for tasks $t \in \{q, s, d\}$, following the same approach as in Eq. (14), (15) and (16). The parameters of the global image distribution are computed as:

$$\begin{aligned}\boldsymbol{m}_t^{\text{global}} &= (\boldsymbol{x}_t^{\text{global}})_\delta, (\boldsymbol{x}_t^{\text{global}})_v, (\boldsymbol{x}_t^{\text{global}})_\alpha, (\boldsymbol{x}_t^{\text{global}})_\beta, \\ &= \text{split}(\boldsymbol{x}_t^{\text{global}}).\end{aligned} \tag{23}$$

Then, we employ the fusion strategy to merge local NIG distributions derived from each sub-images with global one:

$$\text{NIG}(\boldsymbol{m}_{t,i}^{\text{fusion}}) = \text{NIG}(\boldsymbol{m}_{t,i}^{\text{local}}) \oplus \text{NIG}(\boldsymbol{m}_t^{\text{global}}), \tag{24}$$

where $\boldsymbol{m}_{t,i}^{\text{fusion}}$ represents the local-global parameters of the mixture NIG distribution between the i-th local sub-image and the global image in task t. The local-global fusion information is aggregated through averaging:

$$\boldsymbol{x}_t^{\text{fusion}} = \frac{1}{4} \sum_i \boldsymbol{m}_{t,i}^{\text{fusion}}. \tag{25}$$

Subsequently, we define the evidential loss based on local and global information fusion for single task t as:

$$\mathcal{L}_t^F(\theta) = \frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x} \in \mathcal{B}} \ell^U \left(\text{softplus}(\boldsymbol{x}_t^{\text{fusion}}), \boldsymbol{y}_t, \theta \right). \tag{26}$$

The overall cross-grained loss based on local-global information fusion for multitasks is the sum of these evidential fusion losses for three tasks:

$$\mathcal{L}^F(\theta) = \mathcal{L}_q^F(\theta) + \mathcal{L}_s^F(\theta) + \mathcal{L}_d^F(\theta). \tag{27}$$

Overall Loss

In the proposed DEFNet framework, the overall loss function is composed of multiple components, each targeting a specific aspect of the model performance. The overall loss is denoted as $\mathcal{L}(\theta)$ and contains the multitask loss $\mathcal{L}^M(\theta)$, the cross-region loss $\mathcal{L}^U(\theta)$ resulting from cross sub-region information fusion, and the cross-grained loss $\mathcal{L}^F(\theta)$ from local-global information fusion. Formally, we have the optimization objective of the proposed DEFNet:

$$\mathcal{L}(\theta) = \mathcal{L}^M(\theta) + \lambda_1 \mathcal{L}^U(\theta) + \lambda_2 \mathcal{L}^F(\theta), \tag{28}$$

where λ_1 and λ_2 are parameters that control the relative contribution of each loss component to the overall loss.

Experiments

Experimental Setups

We conduct evaluation on both synthetic and authentic distorted datasets. The former includes LIVE (Sheikh, Sabir, and Bovik 2006), CSIQ (Larson and Chandler 2010) and KADID-10k (Lin, Hosu, and Saupe 2019), while the latter consists of BID (Ciancio et al. 2011), LIVE-C (Ghadiyaram and Bovik 2016) and KonIQ-10k (Hosu et al. 2020). Additional experiments are conducted in the TID2013 (Ponomarenko et al. 2015), SPAQ (Fang et al. 2020), PIPAL (Gu et al. 2020), and Waterloo exploration database (WED) (Ma et al. 2017a). Each dataset is randomly divided into training, validation and test sets in the ratio of 70%, 10%, 20% across ten sessions. The performance is evaluated using Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC).

Implementation Details

Within the CLIP, we employ ViT-B/32 (Radford et al. 2021) as the visual encoder and GPT-2 base model (Radford et al. 2019) as the text encoder. We train the uncertainty-based evidential loss in Eq. (3) with weights $\tau = 0.05$. For the training phase, we initialize the learning rate to $5e-6$ and train the model for a total of 80 epochs. The mini-batch size is set to 48, with 4 samples from each of the LIVE, CSIQ, BID, and LIVE-C datasets, and 16 samples from both the KADID-10k and KonIQ-10k datasets. Throughout the training and inference processes, we perform random cropping to obtain 4 and 15 sub-images from the raw input images, respectively. Each sub-image is with a fixed size of $3 \times 224 \times 224$. All experiments are conducted with one NVIDIA RTX 4090 GPU.

Model Performance

To evaluate the effectiveness of the proposed DEFNet framework, we compare it to four knowledge-driven MOS-free BIQA models like Ma19 (Ma et al. 2019), as well as a number of neural network-based methods. The experimental results in terms of SRCC and PLCC are shown in Table 1, where we draw several conclusions. The experimental results in terms of SRCC and PLCC are shown in Table 1, where we draw several conclusions. First, DEFNet exhibits outstanding performance on both synthetic and authentic distortion datasets compared to existing methods. The superior performance can be attributed to the multilevel information fusion strategy, which effectively integrates quality features and patterns across sub-regions while maintaining a balance between detailed and global perspectives. Second, assessing the image quality of authentic distorted scenarios is more difficult than for synthetic distorted scenarios. This holds for most methods and is the general agreement in the field of BIQA.

Cross-Dataset Evaluation

To evaluate the generalization capability, we conduct cross-dataset evaluation in a zero-shot setting, following the approach outlined in previous works (Zhang et al. 2021, 2023).

Method	Synthetic distortion						Authentic distortion					
	LIVE		CSIQ		KADID-10k		BID		LIVE-C		KonIQ-10k	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIQE (Mittal, Soundararajan, and Bovik 2013)	0.908	0.904	0.631	0.719	0.389	0.442	0.573	0.618	0.446	0.507	0.415	0.438
ILNIQE (Zhang, Zhang, and Bovik 2015)	0.887	0.894	0.808	0.851	0.565	0.611	0.548	0.494	0.469	0.518	0.509	0.534
dipIQ (Ma et al. 2017b)	0.940	0.933	0.511	0.778	0.304	0.402	0.009	0.346	0.187	0.290	0.228	0.437
Ma19 (Ma et al. 2019)	0.922	0.923	0.926	0.929	0.465	0.501	0.373	0.399	0.336	0.405	0.360	0.398
DBCNN (Zhang et al. 2020)	0.963	0.966	0.940	0.954	0.878	0.878	0.864	0.883	0.835	0.854	0.864	0.868
HyperIQA (Su et al. 2020)	0.966	0.968	0.934	0.946	0.872	0.869	0.848	0.868	0.855	0.878	0.900	0.915
UNIQUE (Zhang et al. 2021)	0.961	0.952	0.902	0.921	0.884	0.885	0.852	0.875	0.854	0.884	0.895	0.900
TreS (Golestaneh, Dadsetan, and Kitani 2022)	0.965	0.963	0.902	0.923	0.881	0.879	0.853	0.871	0.846	0.877	0.907	0.924
LIQE (Zhang et al. 2023)	0.970	0.951	0.936	0.939	0.930	0.931	0.875	0.900	0.904	**0.910**	0.919	0.908
CONTRIQUE (Madhusudana et al. 2022)	0.960	0.961	0.942	0.955	0.934	0.937	-	-	0.845	0.857	0.894	0.906
VCRNet (Pan et al. 2022)	0.973	0.974	0.943	0.955	0.853	0.849	-	-	0.856	0.865	0.894	0.909
Re-IQA (Saha, Mishra, and Bovik 2023)	0.970	0.971	0.947	0.960	0.872	0.885	-	-	0.840	0.854	0.914	0.923
DPNet (Wang et al. 2023)	0.971	0.971	0.942	0.952	0.923	0.924	-	-	0.849	0.864	-	-
QAL-IQA (Zhou et al. 2025)	0.971	0.973	0.963	**0.970**	0.908	0.910	-	-	0.859	0.875	0.917	**0.928**
CDINet (Zheng et al. 2024)	0.977	**0.975**	0.952	0.960	0.920	0.919	0.874	0.899	0.865	0.880	0.916	**0.928**
TOPIQ-FR (Chen et al. 2024)	0.887	0.882	0.894	0.894	0.895	0.896	-	-	-	-	-	-
KGANet (Zhou et al. 2024)	0.963	0.966	0.954	0.963	0.940	0.943	-	-	-	-	-	-
CausalQuality-VGG (Shen et al. 2025)	0.932	0.929	0.952	0.949	0.899	0.898	-	-	-	-	-	-
CausalQuality-EffNet (Shen et al. 2025)	0.932	0.927	0.938	0.933	0.907	0.905	-	-	-	-	-	-
DEFNet	**0.978**	0.960	**0.967**	0.964	**0.942**	**0.944**	**0.910**	**0.909**	**0.918**	0.897	**0.920**	0.901

Table 1: Performance comparison of the proposed approach and state-of-the-art methods on datasets with synthetic and authentic distortion. Best and second-best scores are highlighted in bold and underlined, respectively.

Training	TID2013	SPAQ	PIPAL
NIQE (Mittal, Soundararajan, and Bovik 2013)	0.314	0.578	0.153
DBCNN$_d$ (Zhang et al. 2020)	0.471	0.801	0.413
DBCNN$_q$ (Zhang et al. 2020)	0.686	0.412	0.321
PaQ2PiQ (Ying et al. 2020)	0.423	0.823	0.400
MUSIQ (Ke et al. 2021)	0.584	0.853	0.450
UNIQUE (Zhang et al. 2021)	0.768	0.838	0.444
DEFNet	**0.828**	**0.868**	**0.464**

Table 2: SRCC performance in cross-dataset evaluation. The subscripts "d" and "q" represent that the model is trained on KADID-10k and KonIQ-10k, respectively.

The experiments are performed on the TID2013 (Ponomarenko et al. 2015), SPAQ (Fang et al. 2020) and PIPAL training set (Gu et al. 2020) datasets. As shown in Table 2, DEFNet demonstrates high robustness in TID2013 and SPAQ, achieving SRCC values of 0.828 and 0.868, respectively. These results highlight the model's strong ability to generalize effectively to unseen datasets with both synthetic and authentic distortions, outperforming existing methods. However, the model's performance on PIPAL, while competitive, is comparatively lower with an SRCC of 0.464. This indicates that while DEFNet performs well in scenarios where distortions are similar to those encountered during training, its generalization to highly diverse and novel distortions remains a challenge. Overall, the results affirm DEFNet's strong generalization ability but also emphasize opportunities for improvement in handling datasets with unique distortion characteristics like PIPAL.

Ablation Study

In this section, we conduct ablation study to validate the contribution of each task assistance and each loss components to the overall performance. A total of four different task combinations are explored, specifically, ablation experiments are conducted on the presence or absence of scene classification and distortion type categorization. Within each task combination, ablation of each component in the overall loss is also performed. All the results are averaged across the six datasets and listed in Table 3.

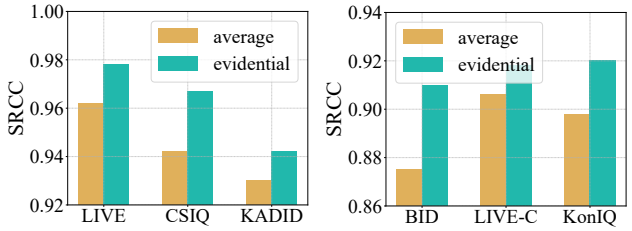
Task	Loss component $\mathcal{L}^M$	$\mathcal{L}^U$	$\mathcal{L}^F$	SRCC	PLCC	ACC$_s$	ACC$_d$
q	✓			0.910	0.898	-	-
	✓	✓		0.914	0.905	-	-
	✓		✓	0.916	0.905	-	-
	✓	✓	✓	0.922	0.908	-	-
$q+s$	✓			0.915	0.904	0.873	-
	✓	✓		0.920	0.915	0.878	-
	✓		✓	0.921	0.910	0.878	-
	✓	✓	✓	0.923	0.921	**0.882**	-
$q+d$	✓			0.913	0.906	-	0.837
	✓	✓		0.924	0.915	-	0.840
	✓		✓	0.925	0.913	-	0.838
	✓	✓	✓	0.933	0.921	-	0.838
$q+s+d$	✓			0.916	0.906	0.870	**0.851**
	✓	✓		0.925	0.921	0.864	0.830
	✓		✓	0.926	0.921	0.873	0.838
	✓	✓	✓	**0.939**	**0.929**	0.879	0.847

Table 3: Mean correlation coefficients (SRCC and PLCC) and mean accuracy (ACC) on the six datasets. The subscripts "s" and "d" stand for accuracy of the scene and distortion type classification tasks, respectively.

A couple of observations can be drawn. First, utilizing auxiliary tasks can significantly improve the performance of BIQA. With both two auxiliary tasks aided, DEFNet archieves the best performance across all 16 settings in the ablation study. Second, the proposed information fu-

Parameter		Synthetic Distortion						Authentic Distortion					
		LIVE		CSIQ		KADID-10k		BID		LIVE-C		KonIQ-10k	
λ_1	λ_2	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
0.1	0.1	0.978	**0.960**	0.967	0.964	0.942	0.944	0.910	0.909	0.918	**0.897**	0.920	**0.901**
0.1	0.2	0.979	0.959	0.971	0.964	**0.947**	**0.947**	0.906	0.911	**0.921**	0.890	**0.921**	**0.901**
0.1	0.3	0.976	0.952	0.970	0.961	0.945	0.943	0.903	0.901	0.907	0.867	0.920	0.894
0.2	0.1	**0.981**	0.952	**0.973**	**0.965**	0.945	**0.947**	**0.912**	**0.921**	0.910	0.869	0.919	0.898
0.3	0.1	0.978	0.946	0.967	0.949	0.946	0.943	0.911	0.900	0.904	0.836	0.918	0.893
0.4	0.1	0.979	0.945	0.966	0.949	0.944	0.940	0.903	0.897	0.903	0.821	0.917	0.890

Table 4: SRCC and PLCC across the six IQA datasets under different weighting parameters. Best scores are highlighted in bold.



(a) Synthetic distortion. (b) Authentic distortion

Figure 4: Comparison of fusion strategy.

sion strategy, either across sub-regions or between local and global image context, contribute positively to BIQA. The inclusion of either cross-region loss or cross-grained loss leads to noticeable improvements in model performance. These loss components enable the model to better capture complementary features. With both loss components aided, DEFNet achieves highest performance in most cases. Third, the extent to which evidential fusion positively impacts performance surpasses that offered by the auxiliary tasks alone. This underscores the contribution of the proposed DEFNet, in which the multilevel trustworthy evidential fusion leads to a more accurate quality assessment.

Fusion Strategy

To validate the superiority of the proposed evidential fusion strategy, we conduct a controlled comparison by replacing evidential fusion operations in Eqs. (18) and (24) with simple averaging. As shown in Figure 4, the evidential fusion strategy consistently outperforms averaging across all synthetic and authentic datasets, with an average improvement of 2.0% in SRCC. This demonstrates the critical role of the proposed evidential fusion strategy in effectively modeling prediction uncertainty and enabling more accurate quality assessment compared to simple averaging fusion.

Hyperparameter Analysis

In order to discuss the effect of the weighting parameters in Eq. (28), we adjust different combinations of λ_1, λ_2 and list the experimental results in the IQA six datasets in Table 4. This gives an illustration of the trade-off between the contributions of the cross-region loss and the cross-grained loss

Method	LIQE (Zhang et al. 2023)	DEFNet
CI width (↓)	0.286	0.251

Table 5: Mean confidence interval widths.

to the overall model performance. As the weighting parameters increase from small values, the performance of both BIQA and the auxiliary tasks improves initially, reaching optimal values at moderate weight settings (e.g., $\lambda_1 = 0.2$ and $\lambda_2 = 0.2$). This improvement can be attributed to the gradual integration of evidential learning, which enhances the model's ability to extract and integrate complementary information from the auxiliary tasks. However, when the weights become excessively large (e.g., $\lambda_1 = 0.4$), the performance begins to degrade. This decline is likely due to the model overemphasizing the evidential loss components, which detracts from the focus on the primary BIQA task.

Uncertainty Analysis

By applying evidence learning and utilizing normal-inverse gamma distribution mixture, we reduce the both aleatoric and epistemic uncertainty of the model for BIQA prediction. Specifically, DEFNet exhibits better performance and lower uncertainty compared to LIQE (a representative of methods that utilize auxiliary tasks to aid BIQA). In addition, the mean confidence interval (CI) widths in the scatter plot are shown in Table 5, which quantitatively illustrates that the proposed evidential fusion strategy with advanced uncertainty estimation technique is beneficial for decreasing uncertainty.

Conclusion

This paper introduced a deep evidential fusion network based on multitasks, which addresses the limitations in insufficient information integration and inflexible uncertainty estimation. To this end, a trustworthy information fusion strategy has been proposed to combine cross sub-region diversity with local-global context. By incorporating NIG distribution mixture, our approach has enhanced uncertainty estimation while improving the robustness of quality information capture. To sum up, DEFNet serves as a practical and effective solution for uncertainty-aware BIQA.

Acknowledgments

This paper was supported by National Key R&D Program of China (2024YFC3308305), National Natural Science Foundation of China under Grants (62436006, 62172016, 62572007), Sanya Science and Technology Special Fund (No. 2024KFJX04), Beijing Natural Science Foundation (No. L257018) and Beijing Nova Program.

References

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, 14927–14937.

Chen, C.; Mo, J.; Hou, J.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; and Lin, W. 2024. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33: 2404–2418.

Ciancio, A.; Targino da Costa, A. L. N. T.; da Silva, E. A. B.; Said, A.; Samadani, R.; and Obrador, P. 2011. No-reference blur assessment of digital pictures based on multi-feature classifiers. *IEEE Transactions on Image Processing*, 20(1): 64–75.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3677–3686.

Gao, Y.; Min, X.; Zhu, Y.; Zhang, X.-P.; and Zhai, G. 2023. Blind image quality assessment: A fuzzy neural network for opinion score distribution prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3): 1641–1655.

Ghadiyaram, D.; and Bovik, A. C. 2016. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1): 372–387.

Golestaneh, S. A.; Dadsetan, S.; and Kitani, K. M. 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.

Gu, J.; Cai, H.; Chen, H.; Ye, X.; Jimmy S, R.; and Dong, C. 2020. Pipal: A large-scale image quality assessment dataset for perceptual image restoration. In *16th European Conference on Computer Vision*, 633–651. Springer.

Guo, X.; Luo, S.; Dong, Y.; Liang, Z.; Li, Z.; Zhang, X.; and Chen, X. 2025. An asymmetric calibrated transformer network for underwater image restoration: X. Guo et al. *The Visual Computer*, 1–13.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale image quality Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1): 011006.

Li, A.; Wu, J.; Tian, S.; Li, L.; Dong, W.; and Shi, G. 2022. Blind image quality assessment based on progressive multi-task learning. *Neurocomputing*, 500: 307–318.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *International Conference on Quality of Multimedia Experience*, 1–3.

Liu, S.; Johns, E.; and Davison, A. J. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1871–1880.

Lou, Y.; Chen, Y.; Xu, D.; Zhou, D.; Cao, Y.; Wang, H.; and Huang, Y. 2023. Refining the unseen: self-supervised two-stream feature extraction for image quality assessment. In *IEEE International Conference on Data Mining*, 1193–1198. IEEE.

Lou, Y.; Xu, D.; Zhang, R.; Zhang, J.; Cao, Y.; Wang, H.; and Huang, Y. 2024a. MR Image Quality Assessment via Enhanced Mamba: A Hybrid Spatial-Frequency Approach. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3561–3564. IEEE.

Lou, Y.; Zhang, J.; Xu, D.; Cao, Y.; Wang, H.; and Huang, Y. 2024b. No-Reference MRI quality assessment via contrastive representation: spatial and frequency domain perspectives. In *2024 IEEE International Conference on Multimedia and Expo*, 1–6.

Luo, S.; Chen, X.; Chen, W.; Li, Z.; Wang, S.; and Pun, C.-M. 2024. Devignet: High-resolution vignetting removal via a dual aggregated fusion transformer with adaptive channel expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4000–4008.

Ma, H.; Han, Z.; Zhang, C.; Fu, H.; Zhou, J. T.; and Hu, Q. 2021. Trustworthy multimodal regression with mixture of normal-inverse Gamma distributions. In *Advances in Neural Information Processing Systems*, volume 34, 6881–6893.

Ma, K.; Duanmu, Z.; Wu, Q.; Wang, Z.; Yong, H.; Li, H.; and Zhang, L. 2017a. Waterloo Exploration Database: New Challenges for Image Quality Assessment Models. *IEEE Transactions on Image Processing*, 26(2): 1004–1016.

Ma, K.; Liu, W.; Liu, T.; Wang, Z.; and Tao, D. 2017b. dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs. *IEEE Transactions on Image Processing*, 26(8): 3951–3964.

Ma, K.; Liu, X.; Fang, Y.; and Simoncelli, E. P. 2019. Blind Image Quality Assessment by Learning from Multiple Annotators. In *IEEE International Conference on Image Processing*, 2344–2348.

Madhusudana, P. C.; Birkbeck, N.; Wang, Y.; Adsumilli, B.; and Bovik, A. C. 2022. Image Quality Assessment Using Contrastive Learning. *IEEE Transactions on Image Processing*, 31: 4149–4161.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain.

IEEE Transactions on Image Processing, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3): 209–212.

Pan, Z.; Yuan, F.; Lei, J.; Fang, Y.; Shao, X.; and Kwong, S. 2022. VCRNet: Visual compensation restoration network for no-reference image quality assessment. *IEEE Transactions on Image Processing*, 31: 1613–1627.

Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; and Jay Kuo, C.-C. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30: 57–77.

Qu, Q.; Chen, X.; Chung, V.; and Chen, Z. 2021. Light Field Image Quality Assessment With Auxiliary Learning Based on Depthwise and Anglewise Separable Convolutions. *IEEE Transactions on Broadcasting*, 67(4): 837–850.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *38th International Conference on Machine Learning*, volume 139, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Saha, A.; Mishra, S.; and Bovik, A. C. 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.

Sheikh, H.; Sabir, M.; and Bovik, A. 2006. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 15(11): 3440–3451.

Shen, W.; Zhou, M.; Chen, Y.; Wei, X.; Feng, Y.; Pu, H.; and Jia, W. 2025. Image quality assessment: Investigating causal perceptual effects with abductive counterfactual inference. In *Computer Vision and Pattern Recognition Conference*, 17990–17999.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 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.

Thurstone, L. L. 2017. A law of comparative judgment. In *Scaling*, 81–92. Routledge.

Tsai, M.-F.; Liu, T.-Y.; Qin, T.; Chen, H.-H.; and Ma, W.-Y. 2007. Frank: A ranking method with fidelity loss. In *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 383–390.

Wang, X.; Xiong, J.; Li, B.; Suo, J.; and Gao, H. 2023. Learning hybrid representations of semantics and distortion for blind image quality assessment. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Yan, B.; Bare, B.; and Tan, W. 2019. Naturalness-Aware Deep No-Reference Image Quality Assessment. *IEEE Transactions on Multimedia*, 21(10): 2603–2615.

Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; and Bovik, A. 2020. From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3575–3585.

Zhang, J.; Xu, D.; Chen, Y.; Lou, Y.; and Huang, Y. 2024. Curriculum learning for self-iterative semi-supervised medical image segmentation. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1342–1349. IEEE.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Transactions on Image Processing*, 24(8): 2579–2591.

Zhang, R.; Huang, Y.; Cao, Y.; and Wang, H. 2025a. MoleBridge: Synthetic Space Projecting with Discrete Markov Bridges. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhang, R.; Huang, Y.; Lou, Y.; Ding, W.; Cao, Y.; and Wang, H. 2025b. Synergistic Attention-Guided Cascaded Graph Diffusion Model for Complementarity Determining Region Synthesis. *IEEE Trans. Neural Networks Learn. Syst.*, 36(7): 11875–11886.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2020. Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1): 36–47.

Zhang, W.; Ma, K.; Zhai, G.; and Yang, X. 2021. Uncertainty-Aware Blind Image Quality Assessment in the Laboratory and Wild. *IEEE Transactions on Image Processing*, 30: 3474–3486.

Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; and Ma, K. 2023. 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*, 14071–14081.

Zhao, K.; Yuan, K.; Sun, M.; Li, M.; and Wen, X. 2023. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22302–22313.

Zheng, L.; Luo, Y.; Zhou, Z.; Ling, J.; and Yue, G. 2024. CDINet: Content Distortion Interaction Network for Blind Image Quality Assessment. *IEEE Transactions on Multimedia*, 26: 7089–7100.

Zhou, M.; Shen, W.; Wei, X.; Luo, J.; Jia, F.; Zhuang, X.; and Jia, W. 2025. Blind image quality assessment: Exploring content fidelity perceptibility via quality adversarial learning. *International Journal of Computer Vision*, 1–17.

Zhou, T.; Tan, S.; Zhao, B.; and Yue, G. 2024. Multitask deep neural network with knowledge-guided attention for blind image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 7577–7588.