

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

Yiwei Lou, Yuanpeng He, Rongchao Zhang, Yongzhi Cao, Hanpin Wang, Yu Huang  
 Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education;  
 School of Computer Science, Peking University

## 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 Deep Evidential Fusion Network (DEFNet) for BIQA, which performs multi-task 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 sub-regions 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.*

## 1. 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 helps to improve the overall quality of experience for end-users. This technique is of great importance in a wide range of application areas, such as real-time multimedia processing [2, 5, 13–15, 60] and medical image analysis [6, 30, 31, 43, 63].

Over time, BIQA approaches have undergone a significant evolution from early techniques based on handcrafted feature extraction and manual characterization [39, 40] to

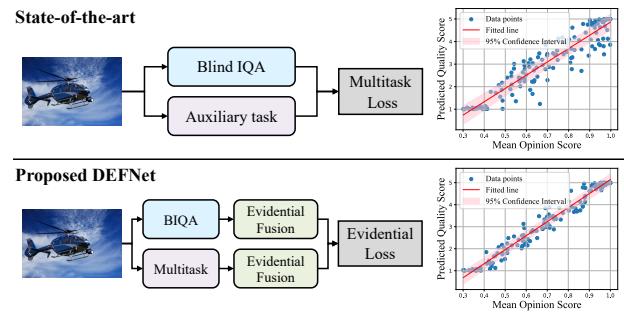


Figure 1. Comparison between the proposed DEFNet and state-of-the-art methods that utilize auxiliary tasks to assist in BIQA. We propose to include evidential fusion for each task for higher performance and lower uncertainty.

more sophisticated data-driven and deep learning-based approaches [11, 29, 37, 47, 51, 68]. 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(a). For instance, scene statistics [58] and image content [26] offer valuable contextual information that can significantly influence the quality perception. Besides, distortion type and degree classification [37] and spatial angular estimation [44] 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.

Despite these advances, existing methods still face challenges in two major aspects. **(i) In-depth information fusion.** On one hand, this requires **1 inter-task information integration**. Some existing approaches treat auxiliary tasks as independent modules, leading to information fragmentation and a lack of in-depth mining of potential inter-task correlations. On the other hand, it necessitates **2 multilevel and cross-region feature fusion**, which involves full considering of the complex interactions between features and ex-

ploring diverse sub-regions that may contain different distortion patterns and visual characteristics. **(ii) Comprehensive uncertainty estimation.** Though significant progress [23, 55, 66] has been made in uncertainty estimation for BIQA, it is still difficult to provide a **③ 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 **Deep Evidential Fusion Network** (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 [46] to extract both local and global image features across the three different tasks, followed by a simultaneous multitask optimization to tackle challenge **①**. 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 **②**. Furthermore, to address challenge **③**, DEFNet incorporates a robust uncertainty estimation mechanism inspired by evidence theory [1]. 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.

## 2. Problem Statement and Preliminaries

To formalize the problem of blind image quality assessment, denote  $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$  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  $\mathbf{x}$  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  $\mu$  and  $\sigma^2$  are the unknown mean and variance. A detailed justification of this assumption is provided in Supplementary B. The posterior distribution  $p(\mu, \sigma|q(\mathbf{x}_1 \dots N))$  is assumed to follow a normal-inverse gamma (NIG) distribution  $(\mu, \sigma) \sim \text{NIG}(\delta, v, \alpha, \beta)$ , 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:

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

where  $\mathbf{x}, \mathbf{y}$  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:

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

where  $\phi = 2v + \alpha$  is the total evidence [1]. 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:

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

where  $\tau$  is the weights keeping the balance between model fitting and uncertainty inflation.

## 3. 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 images contexts, and then performs single task optimization,

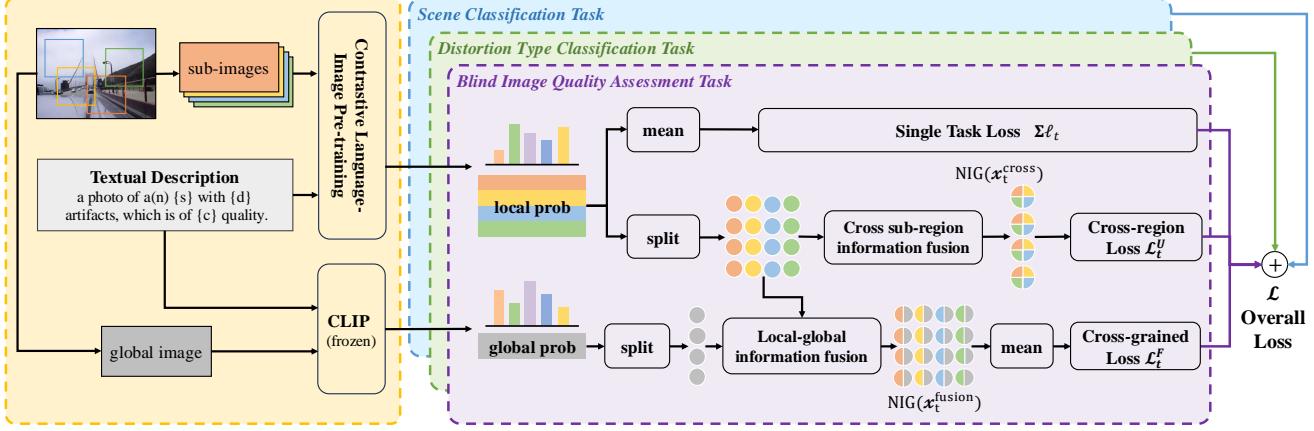


Figure 2. Overview of the proposed DEFNet framework.

as well as two levels (cross sub-region and local-global) of evidential fusion across all the three tasks. A complete algorithm description is shown in Supplementary C.

### 3.1. Local and Global Probability Scores

In the proposed DEFNet framework, we employ contrastive language-image pre-training (CLIP) [46] to extract the feature embeddings and compute both local 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. Detailed information on scenes ( $s \in \mathcal{S}$ ), distortion types ( $d \in \mathcal{D}$ ) and quality levels ( $c \in \mathcal{C}$ ) in the text is given in the Supplementary D.

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  $\text{logit}(c, s, d | \mathbf{x})$ . Subsequently, DEFNet performs softmax activation to derive the joint probability

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

where  $\kappa$  is a learnable parameter,  $c, s, d$  indicate the quality class, scene and distortion type, respectively. After that, the

local probability scores  $\hat{p}(c, s, d | \mathbf{x}^{\text{local}})$  and global scores  $\hat{p}(c, s, d | \mathbf{x}^{\text{global}})$  are derived.

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

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

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

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

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

### 3.2. Multitask Optimization

In the multitask optimization framework, BIQA is the primary task represented by the loss component  $\ell_q$ , while components  $\ell_s$  and  $\ell_d$  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 [53], the BIQA loss for image pair  $(\mathbf{x}_1, \mathbf{x}_2)$  is defined as:

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

where

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

quantifies the likelihood that  $\mathbf{x}_1$  is of higher predicted quality than  $\mathbf{x}_2$  using standard Normal cumulative distribution function  $\Phi(\cdot)$  under the Thurstone's model [52], and  $p(\mathbf{x}_1, \mathbf{x}_2)$  is a binary label indicating whether the ground-truth MOS  $q(\mathbf{x}_1) \geq q(\mathbf{x}_2)$ .

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

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

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

$$\ell_s(\mathbf{x}; \theta) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left( 1 - \sqrt{p(s|\mathbf{x})\hat{p}(s|\mathbf{x})} - \sqrt{(1-p(s|\mathbf{x}))(1-\hat{p}(s|\mathbf{x}))} \right), \quad (10)$$

where  $\mathcal{S}$  is the set of all possible scene categories,  $p(s|\mathbf{x})$  is a binary label indicating whether the image  $\mathbf{x}$  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:

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

and further define the distortion type classification loss as:

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

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

Utilizing auxiliary tasks, DEFNet optimizes losses of the three separate tasks, integrating them into the multitask loss [67], which in a mini-batch  $\mathcal{B}$  is defined as

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

where  $\theta$  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 [28].

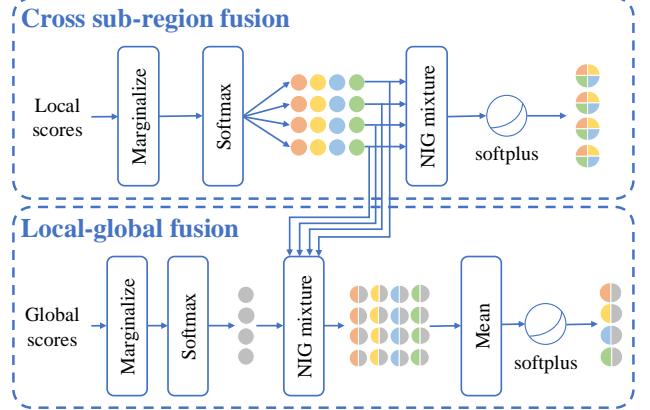


Figure 3. Overview of the cross sub-region information fusion (top) and local-global information fusion (bottom).

### 3.3. 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  $\hat{p}(c, s, d|\mathbf{x}_i^{\text{local}})$  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):

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

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

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

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:

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

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:

$$\text{NIG}(\mathbf{x}_t^{\text{cross}}) = \text{NIG}(\mathbf{m}_{t,1}^{\text{local}}) \oplus \text{NIG}(\mathbf{m}_{t,2}^{\text{local}}) \oplus \text{NIG}(\mathbf{m}_{t,3}^{\text{local}}) \oplus \text{NIG}(\mathbf{m}_{t,4}^{\text{local}}), \quad (18)$$

where  $\mathbf{x}_t^{\text{cross}}$  is the mixture NIG distribution parameters,  $\oplus$  is the summation operation for two NIG distributions [32], which is defined as:

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

where

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

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

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

where  $\mathbf{y}_q = q(\mathbf{x})$  is the ground-truth MOS,  $\mathbf{y}_s = p(s|\mathbf{x})$  and  $\mathbf{y}_d = p(d|\mathbf{x})$  are binary labels indicating whether the image  $\mathbf{x}$  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:

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

### 3.4. 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  $\mathbf{x}_t^{\text{global}}$  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:

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

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

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

where  $\mathbf{m}_{t,i}^{\text{fusion}}$  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:

$$\mathbf{x}_t^{\text{fusion}} = \frac{1}{4} \sum_i \mathbf{m}_{t,i}^{\text{fusion}}. \quad (25)$$

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

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

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

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

### 3.5. 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:

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

where  $\lambda_1$  and  $\lambda_2$  are parameters that control the relative contribution of each loss component to the overall loss.

## 4. Experiments

### 4.1. Experimental Setups

We conduct evaluation on both synthetic and authentic distorted datasets. The former includes LIVE [49], CSIQ [25] and KADID-10k [27], while the latter consists of BID [7], LIVE-C [10] and KonIQ-10k [22]. Additional experiments are conducted in the TID2013 [42], SPAQ [9], PIPAL

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.

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	SRCC	PLCC	SRCC	PLCC
NIQE [40]	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 [64]	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 [34]	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 [35]	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 [65]	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 [51]	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 [66]	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 [11]	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 [67]	0.970	0.951	0.936	0.939	0.930	0.931	0.875	0.900	0.904	<b>0.910</b>	0.919	0.908				
CONTRIQUE [37]	0.960	0.961	0.942	0.955	0.934	0.937	-	-	0.845	0.857	0.894	0.906				
VCRNet [41]	0.973	<u>0.974</u>	0.943	0.955	0.853	0.849	-	-	0.856	0.865	0.894	0.909				
Re-IQA [47]	0.970	0.971	0.947	0.960	0.872	0.885	-	-	0.840	0.854	0.914	0.923				
DPNet [54]	0.971	0.971	0.942	0.952	0.923	0.924	-	-	0.849	0.864	-	-				
QAL-IQA [69]	0.971	0.973	<u>0.963</u>	<b>0.970</b>	0.908	0.910	-	-	0.859	0.875	0.917	<b>0.928</b>				
CDINet [68]	<u>0.977</u>	<b>0.975</b>	0.952	0.960	0.920	0.919	0.874	0.899	0.865	0.880	0.916	<b>0.928</b>				
TOPIQ-FR [4]	0.887	0.882	0.894	0.894	0.895	0.896	/	/	/	/	/	/				
KGANet [70]	0.963	0.966	0.954	0.963	0.940	<u>0.943</u>	/	/	/	/	/	/				
CausalQuality-VGG [50]	0.932	0.929	0.952	0.949	0.899	0.898	/	/	/	/	/	/				
CausalQuality-EffNet [50]	0.932	0.927	0.938	0.933	0.907	0.905	/	/	/	/	/	/				
DPSF [57]	/	/	/	/	/	/	0.872	0.883	0.865	0.882	0.912	<u>0.925</u>				
DEFNet	<b>0.978</b>	0.960	<b>0.967</b>	<u>0.964</u>	<b>0.942</b>	<b>0.944</b>	<b>0.910</b>	<b>0.909</b>	<b>0.918</b>	<u>0.897</u>	<b>0.920</b>	0.901				

[12], and Waterloo exploration database (WED) [33]. 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).

#### 4.2. Implementation Details

Within the CLIP, we employ ViT-B/32 [46] as the visual encoder and GPT-2 base model [45] 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.

#### 4.3. Model Performance

To evaluate the effectiveness of the proposed DEFNet framework, we compare it to four knowledge-driven MOS-free BIQA models including NIQE [40], ILNIQE [64], dipIQ [34] and Ma19 [35], as well as neural network-based methods, including DBCNN [65], HyperIQA [51],

UNIQUE [66], TreS [11], LIQE [67], CONTRIQUE [37], VCRNet [41], Re-IQA [47], DPNet [54], QAL-IQA [69], CDINet [68], TOPIQ-FR [4], KGANet [70], CausalQuality [50] and DPSF [57]. 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, LIQE [67] and CDINet [68] demonstrate promising performance, especially on datasets that are relatively small and with basic distortions. Third, 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.

#### 4.4. gMAD Competition

To illustrate DEFNet’s ability to effectively reintegrate information from different IQA datasets into a shared perceptual scale, gMAD competition [36] are carried out in WED [33], which is representative of synthetic distortion dataset. The gMAD framework presents qualitative differences by fixing one method’s prediction while varying the quality identified by another. This setup highlights scenarios where each model performs well or poorly, providing a deeper

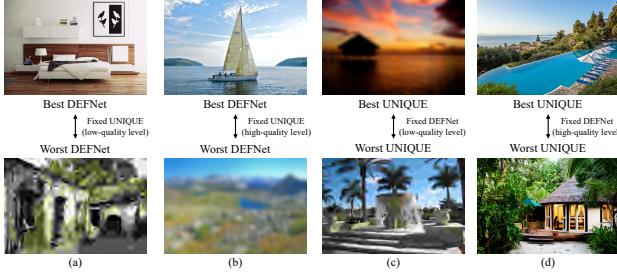


Figure 4. gMAD competition between UNIQUE and DEFNet in WED. (a) Fixed UNIQUE at low-quality level. (b) Fixed UNIQUE at high-quality level. (c) Fixed DEFNet at low-quality level. (d) Fixed DEFNet at high-quality level.

Table 2. SRCC performance in cross-dataset evaluation under zero-shot setting. The subscripts “*d*” and “*q*” represent that the model is trained on KADID-10k and KonIQ-10k, respectively.

Training	TID2013	SPAQ	PIPAL
NIQE [40]	0.314	0.578	0.153
DBCNN <sub>d</sub> [65]	0.471	0.801	0.413
DBCNN <sub>q</sub> [65]	0.686	0.412	0.321
PaQ2PiQ [61]	0.423	0.823	0.400
MUSIQ [24]	0.584	0.853	0.450
UNIQUE [66]	0.768	0.838	0.444
DEFNet	<b>0.828</b>	<b>0.868</b>	<b>0.464</b>

view of their strengths and weaknesses. As shown in Figure 4, the highest predictions of DEFNet align well with high-quality images, while the worst output appropriately reflects severe distortion. In contrast, UNIQUE shows a tendency to misclassify certain low-quality and high-quality images. DEFNet demonstrates superior consistency in ranking high-quality and low-quality images. This indicates that DEFNet not only consistently produces more reliable quality assessment, but also offers strong generalization capability. Further results and analysis in SPAQ (representative of authentic distortion dataset) are given in Supplementary E.1.

#### 4.5. 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 [66, 67]. The experiments are performed on the TID2013 [42], SPAQ [9] and PIPAL training set [12] 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 its generalization to highly di-

Table 3. SRCC performance on diverse distortion types of CSIQ.

Distortion	WN	GB	PN	CD	JPEG	JP2K
BRISQUE [39]	0.682	0.808	0.743	0.396	0.846	0.817
ILNIQE [64]	0.850	0.858	0.874	0.501	0.899	0.906
deepIQA [3]	0.944	0.901	0.867	0.847	0.922	0.934
DBCNN [65]	0.948	0.947	0.941	0.872	0.940	0.953
HyperIQA [51]	0.927	0.915	0.931	0.874	0.934	0.960
DCNet [71]	<b>0.964</b>	<b>0.968</b>	<b>0.958</b>	<b>0.931</b>	<b>0.972</b>	<b>0.966</b>
OLNet [59]	0.945	0.965	0.953	0.925	0.968	0.945
VCRNet [41]	0.939	0.950	0.899	0.919	0.956	0.962
<b>DEFNet</b>	<b>0.969</b>	<b>0.971</b>	<b>0.974</b>	<b>0.941</b>	0.967	<b>0.971</b>

Table 4. Mean correlation coefficients (SRCC and PLCC) and mean accuracy (ACC) on the six datasets. The subscripts “*s*” and “*d*” stands for accuracy of the scene and distortion type classification tasks, respectively. Best scores are highlighted in bold.

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

verse and novel distortions remains a challenge.

#### 4.6. Comparison on Distortion Types

In this section, we present the SRCC performance of the proposed DEFNet compared to several state-of-the-art methods across diverse distortion types in the CSIQ datasets. The distortion types CSIQ include white noise (WN), Gaussian blur (GB), pink Gaussian noise (PN), contrast decrements (CD), JPEG compression (JPEG) and JPEG2000 compression (JP2K). As shown in Table 3, DEFNet outperforms other methods across nearly all types of distortion. This highlights the robustness and adaptability of DEFNet to various types of synthetic image distortion. In addition, this validates the superiority of the multi-level information fusion strategy, which allows to give accurate quality predictions regardless of distortion types. More comparison results on distortion types in other datasets are shown in Supplementary E.2.

Table 5. SRCC and PLCC across the six IQA datasets under different weighting parameters. Best scores are highlighted in bold.

Parameter		Synthetic Distortion						Authentic Distortion					
$\lambda_1$	$\lambda_2$	LIVE		CSIQ		KADID-10k		BID		LIVE-C		KonIQ-10k	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
0.1	0.1	0.978	<b>0.960</b>	0.967	0.964	0.942	0.944	0.910	0.909	0.918	<b>0.897</b>	0.920	<b>0.901</b>
0.1	0.2	0.979	0.959	0.971	0.964	<b>0.947</b>	<b>0.947</b>	0.906	0.911	<b>0.921</b>	0.890	<b>0.921</b>	<b>0.901</b>
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	<b>0.981</b>	0.952	<b>0.973</b>	<b>0.965</b>	0.945	<b>0.947</b>	<b>0.912</b>	<b>0.921</b>	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

#### 4.7. 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. The results for the multitask learning are presented in Supplementary E.3. All the results are averaged across the six datasets and listed in Table 4.

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 achieves the best performance across all 16 settings in the ablation study. Second, the proposed information fusion 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.

#### 4.8. Hyperparameter Analysis

In order to discuss the effect of the weighting parameters in Eq. (28), we adjust different combinations of  $\lambda_1$ ,  $\lambda_2$  and list the experimental results in the IQA six datasets in Table 5. This gives an illustration of the trade-off between the contributions of the cross-region loss and the cross-grained loss 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

Table 6. Mean confidence interval widths.

Method	LIQE [67]	DEFNet
CI width ( $\downarrow$ )	0.286	0.251

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.

#### 4.9. Uncertainty Analysis

By applying evidence theory to BIQA task and utilizing normal-inverse gamma distribution mixture, we reduce the epistemic uncertainty of the model. Specifically, as shown in Figure 1 and Supplementary E.4, 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 6, which quantitatively illustrates that the advanced uncertainty estimation technique is beneficial for decreasing uncertainty.

### 5. Conclusion

This paper introduced evidential fusion into BIQA with the assistance of auxiliary tasks for in-depth information fusion. To this end, we proposed a trustworthy information fusion strategy at two levels. The cross sub-region fusion effectively captures diverse features and patterns from different regions, while the local-global fusion balances fine-grained local details with a broader global view, providing a comprehensive representation of image quality. Our proposed method serves as a practical solution for BIQA uncertainty estimation and in-depth information fusion. **Limitations.** There is still room to improve robustness and generalization for highly diverse and novel distortions and never-before-learned scenarios. In addition, the number of parameters of the model is relatively high and can be further optimized.

## References

- [1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, pages 14927–14937, 2020. [2](#), [12](#)
- [2] Jiesong Bai, Yuhao Yin, Yihang Dong, Xiaofeng Zhang, Chi-Man Pun, and Xuhang Chen. Lensnet: An end-to-end learning framework for empirical point spread function modeling and lensless imaging reconstruction. *arXiv preprint arXiv:2505.01755*, 2025. [1](#)
- [3] Sebastian Bosse, Dominique Maniry, Thomas Wiegand, and Wojciech Samek. A deep neural network for image quality assessment. In *IEEE international conference on image processing*, pages 3773–3777. IEEE, 2016. [7](#), [15](#)
- [4] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. [6](#)
- [5] Xuhang Chen, Zhuo Li, Yanyan Shen, Mufti Mahmud, Hieu Pham, Chi-Man Pun, and Shuqiang Wang. High-fidelity functional ultrasound reconstruction via a visual autoregressive framework. *arXiv preprint arXiv:2505.21530*, 2025. [1](#)
- [6] Yanyuan Chen, Dexuan Xu, Yiwei Lou, Hang Li, Weiping Ding, and Yu Huang. Semi-supervised medical image classification via cross-training and dual-teacher fusion model. *Information Fusion*, page 103389, 2025. [1](#)
- [7] Alexandre Ciancio, André Luiz N Targino Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, 2011. [5](#), [12](#)
- [8] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 57–72. Springer, 2008. [12](#)
- [9] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. [5](#), [7](#)
- [10] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2016. [5](#), [12](#)
- [11] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. 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 (WACV)*, pages 1220–1230, 2022. [1](#), [6](#), [15](#)
- [12] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Ren Jimmy S, and Chao Dong. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 633–651. Springer, 2020. [6](#), [7](#)
- [13] Xiaojiao Guo, Xuhang Chen, Shuqiang Wang, and Chi-Man Pun. Underwater image restoration through a prior guided hybrid sense approach and extensive benchmark analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4784–4800, 2025. [1](#)
- [14] Xiaojiao Guo, Yihang Dong, Xuhang Chen, Weiwen Chen, Zimeng Li, FuChen Zheng, and Chi-Man Pun. Underwater image restoration via polymorphic large kernel cnns. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.
- [15] Xiaojiao Guo, Shenghong Luo, Yihang Dong, Zexiao Liang, Zimeng Li, Xiujun Zhang, and Xuhang Chen. An asymmetric calibrated transformer network for underwater image restoration. *The Visual Computer*, pages 1–13, 2025. [1](#)
- [16] Yuanpeng He. Epl: Evidential prototype learning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2404.06181*, 2024. [12](#)
- [17] Yuanpeng He and Yong Deng. Mmget: A markov model for generalized evidence theory. *Computational and Applied Mathematics*, 41(1):9, 2022. [12](#)
- [18] Yuanpeng He and Fuyuan Xiao. Conflicting management of evidence combination from the point of improvement of basic probability assignment. *International Journal of Intelligent Systems*, 36(5):1914–1942, 2021. [12](#)
- [19] Yuanpeng He, Yali Bi, Lijian Li, Chi-Man Pun, Wenpin Jiao, and Zhi Jin. Mutual evidential deep learning for semi-supervised medical image segmentation. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 2010–2017. IEEE, 2024. [12](#)
- [20] Yuanpeng He, Lijian Li, Tianxiang Zhan, Wenpin Jiao, and Chi-Man Pun. Generalized uncertainty-based evidential fusion with hybrid multi-head attention for weak-supervised temporal action localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3855–3859. IEEE, 2024. [12](#)
- [21] Yuanpeng He, Lijian Li, Tianxiang Zhan, Chi-Man Pun, Wenpin Jiao, and Zhi Jin. Co-evidential fusion with information volume for semi-supervised medical image segmentation. *Pattern Recognition*, 166:111639, 2025. [12](#)
- [22] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. [5](#), [12](#)
- [23] Yuge Huang, Xiang Tian, Rongxin Jiang, and Yaowu Chen. Convolutional neural network with uncertainty estimates for no-reference image quality assessment. In *Tenth International Conference on Graphics and Image Processing*, pages 401–407. SPIE, 2019. [2](#)
- [24] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. [7](#), [12](#)
- [25] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010. [5](#), [12](#)
- [26] Aobo Li, Jinjian Wu, Shiwei Tian, Leida Li, Weisheng Dong, and Guangming Shi. Blind image quality assessment based

- on progressive multi-task learning. *Neurocomputing*, 500: 307–318, 2022. 1
- [27] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience*, pages 1–3, 2019. 5, 12
- [28] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019. 4
- [29] Yiwei Lou, Yanyuan Chen, Dexuan Xu, Doudou Zhou, Yongzhi Cao, Hanpin Wang, and Yu Huang. Refining the unseen: Self-supervised two-stream feature extraction for image quality assessment. In *2023 IEEE International Conference on Data Mining*, pages 1193–1198. IEEE, 2023. 1, 12, 15
- [30] Yiwei Lou, Dexuan Xu, Rongchao Zhang, Jiayu Zhang, Yongzhi Cao, Hanpin Wang, and Yu Huang. Mr image quality assessment via enhanced mamba: A hybrid spatial-frequency approach. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 3561–3564. IEEE, 2024. 1
- [31] Yiwei Lou, Jiayu Zhang, Dexuan Xu, Yongzhi Cao, Hanpin Wang, and Yu Huang. No-reference mri quality assessment via contrastive representation: Spatial and frequency domain perspectives. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2024. 1
- [32] Huan Ma, Zongbo Han, Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. In *Advances in Neural Information Processing Systems*, pages 6881–6893, 2021. 5
- [33] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2017. 6
- [34] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017. 6
- [35] Kede Ma, Xuelin Liu, Yuming Fang, and Eero P. Simoncelli. Blind image quality assessment by learning from multiple annotators. In *IEEE International Conference on Image Processing*, pages 2344–2348, 2019. 6
- [36] Kede Ma, Zhengfang Duanmu, Zhou Wang, Qingbo Wu, Wentao Liu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):851–864, 2020. 6
- [37] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 1, 6, 12
- [38] Fanyong Meng, Dengyu Zhao, Chunqiao Tan, and Zijun Li. Ordinal-cardinal consensus analysis for large-scale group decision making with uncertain self-confidence. *Information Fusion*, 93:344–362, 2023. 12
- [39] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 1, 7, 15
- [40] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 1, 6, 7
- [41] Zhaoqing Pan, Feng Yuan, Jianjun Lei, Yuming Fang, Xiao Shao, and Sam Kwong. Vrnet: Visual compensation restoration network for no-reference image quality assessment. *IEEE Transactions on Image Processing*, 31:1613–1627, 2022. 6, 7, 15
- [42] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30: 57–77, 2015. 5, 7
- [43] Yuanze Qin, Yiwei Lou, Yu Huang, Rigao Chen, and Weihua Yue. An ensemble deep learning approach combining phenotypic data and fmri for adhd diagnosis. *Journal of Signal Processing Systems*, 94(11):1269–1281, 2022. 1
- [44] Qiang Qu, Xiaoming Chen, Vera Chung, and Zhibo Chen. Light field image quality assessment with auxiliary learning based on depthwise and anglewise separable convolutions. *IEEE Transactions on Broadcasting*, 67(4):837–850, 2021. 1
- [45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 6
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6
- [47] Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5846–5855, 2023. 1, 6, 12
- [48] Glenn Shafer. A mathematical theory of evidence turns 40. *International Journal of Approximate Reasoning*, 79:7–25, 2016. 40 years of Research on Dempster-Shafer Theory. 12
- [49] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11): 3440–3451, 2006. 5, 12
- [50] Wenhao Shen, Mingliang Zhou, Yu Chen, Xuekai Wei, Yong Feng, Huayan Pu, and Weijia Jia. Image quality assessment: Investigating causal perceptual effects with abductive counterfactual inference. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17990–17999, 2025. 6
- [51] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blblindly 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*, pages 3667–3676, 2020. 1, 6, 7, 15
- [52] Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017. 4
- [53] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. Frank: a ranking method with fidelity loss. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 383–390, 2007. 3
- [54] Xiaoqi Wang, Jian Xiong, Bo Li, Jinli Suo, and Hao Gao. Learning hybrid representations of semantics and distortion for blind image quality assessment. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6
- [55] Qingbo Wu, Hongliang Li, King N. Ngan, and Kede Ma. Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2078–2089, 2018. 2
- [56] Xingli Wu, Huchang Liao, and Chonghui Zhang. Preference disaggregation analysis for sorting problems in the context of group decision-making with uncertain and inconsistent preferences. *Information Fusion*, 101:102014, 2024. 12
- [57] Jili Xia, Lihuo He, Xinbo Gao, and Bo Hu. Blind image quality assessment for in-the-wild images by integrating distorted patch selection and multi-scale-and-granularity fusion. *Knowledge-Based Systems*, 309:112772, 2025. 6
- [58] Bo Yan, Bahetiyaer Bare, and Weimin Tan. Naturalness-aware deep no-reference image quality assessment. *IEEE Transactions on Multimedia*, 21(10):2603–2615, 2019. 1
- [59] Xiwen Yao, Qinglong Cao, Xiaoxu Feng, Gong Cheng, and Junwei Han. Learning to assess image quality like an observer. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8324–8336, 2023. 7, 12, 15
- [60] Fanghai Yi, Zehong Zheng, Zexiao Liang, Yihang Dong, Xiyang Fang, Wangyu Wu, and Xuhang Chen. Mac-lookup: Multi-axis conditional lookup model for underwater image enhancement. *arXiv preprint arXiv:2507.02270*, 2025. 1
- [61] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepthi Ghadiyaram, and Alan Bovik. 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*, pages 3575–3585, 2020. 7
- [62] Hui Zeng, Lei Zhang, and Alan C Bovik. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint arXiv:1708.08190*, 2017. 15
- [63] Jiayu Zhang, Dexuan Xu, Yanyuan Chen, Yiwei Lou, and Yue Huang. Curriculum learning for self-iterative semi-supervised medical image segmentation. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 1342–1349. IEEE, 2024. 1
- [64] Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 6, 7, 15
- [65] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020. 6, 7, 12, 15
- [66] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 2, 6, 7, 12
- [67] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 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 (CVPR)*, pages 14071–14081, 2023. 4, 6, 7, 8, 12, 15, 17
- [68] Limin Zheng, Yu Luo, Zihan Zhou, Jie Ling, and Guanghui Yue. Cdinet: Content distortion interaction network for blind image quality assessment. *IEEE Transactions on Multimedia*, 26:7089–7100, 2024. 1, 6, 15
- [69] Mingliang Zhou, Wenhao Shen, Xuekai Wei, Jun Luo, Fan Jia, Xu Zhuang, and Weijia Jia. Blind image quality assessment: Exploring content fidelity perceptibility via quality adversarial learning. *International Journal of Computer Vision*, pages 1–17, 2025. 6
- [70] Tianwei Zhou, Songbai Tan, Baoquan Zhao, and Guanghui Yue. 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, 2024. 6
- [71] Zihan Zhou, Yong Xu, Ruotao Xu, and Yuhui Quan. No-reference image quality assessment using dynamic complex-valued neural model. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 1006–1015, New York, NY, USA, 2022. Association for Computing Machinery. 7

## A. Related Work

### A.1. Blind Image Quality Assessment

In the field of blind image quality assessment, state-of-the-art methods have evolved from manual feature extraction to deep learning based approaches. Zhang *et al.* [65] achieved high quality assessment performance by exploring deep bilinear convolutional neural network for synthetic and authentic distorted datasets. Ke *et al.* [24] presented a multi-scale image quality Transformer to process native resolution images with different sizes and aspect ratios. Saha *et al.* [47] proposed an expert hybrid method for automatic perceptual image quality assessment. In addition to improvements at the model level, some research work combined extracted features from different modules [29, 59], some work explored how to perform image quality assessment in a self-supervised manner [37], and some trained a unified model jointly on multiple datasets [66, 67].

### A.2. Evidential Learning

With the rapid development of deep learning technology, although it has achieved good enough results, its reliability and safety have always been concerned [16, 21]. How to make safe and effective decisions is still a very important research focus [17, 18, 38, 56]. To provide a feasible solution to the problem, some researchers choose to transfer the evidence theory into the field of deep learning [19, 20]. The evidence theory is also called Dempster–Shafer theory, which is firstly proposed by Dempster [8] to serve as a general framework for reasoning with uncertainty and further developed by Shafer [48]. The evidnece theory satisfies weaker conditions than Bayesian probability theory and possesses the ability to model uncertainty directly. Recently, a representative work proposed by Amini introduces the concept of uncertainty measures for neural networks for generating more precise predictions [1]. The newly proposed method has two main advantages. First, it can be typically integrated into neural network architectures directly so that the network is able to learn to output parameters of a probability distribution instead of just a single point estimate. Second, when the model makes a prediction, the corresponding uncertainty assessment can be obtained at the same time, and the prediction results can be further modified to achieve better performance.

## B. Assumption Justification

To illustrate the quality score is subject to a normal distribution, theoretical justification and empirical evidence are given as follows.

**Theoretical Justification.** The quality score label (typically mean opinion score) always necessitates averaging the scores of multiple human evaluators. According to the central limit theorem, the sum (or average) of these random

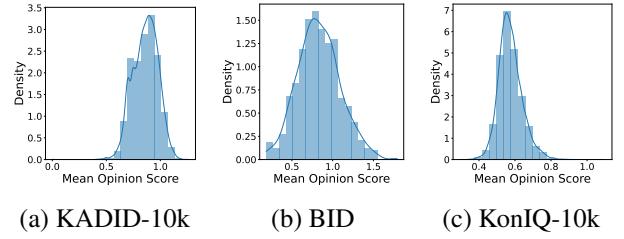


Figure 5. Kernel density estimation plots.

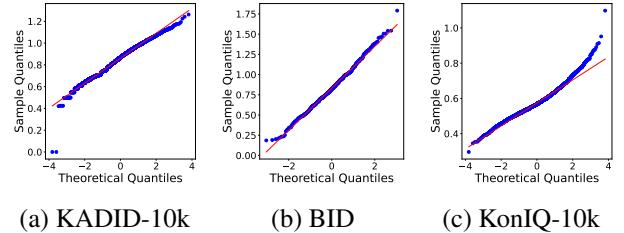


Figure 6. Quantile–quantile plots.

variables, which can be interpreted as the quality scores provided by various evaluators, tends to approximate a normal distribution. This holds true even when the individual variables do not follow a normal distribution. Thus, theoretically, as the number of evaluators increases and the sample size becomes large enough (it holds true for datasets LIVE [49], CSIQ [25], KADID-10k [27], BID [7], LIVE-C [10] and KonIQ-10k [22]), these labels will align with a normal distribution.

**Empirical Evidence.** To empirically validate the assumption, we conducted several statistical analysis. First, we plot kernel density and distribution of the quality scores across datasets (as shown in Figure 5) and observe that the histogram forms a bell-shaped curve, indicating it is subject to normal distribution. Second, we generate Quantile–quantile plots (as shown in Figure 6) comparing the quantiles of the quality score distributions against the quantiles of a standard normal distribution. The alignment of the points along the reference line further confirms the normal distribution.

## C. Algorithm Overview

To give a detailed presentation of variable flow of the method presented in Figure 2 and Section 3, we present Algorithm 1 in association with specific formulas.

## D. More about the Textual Description

To obtain the logits and probability scores for local and global images, CLIP is utilized. The language encoder process the textual template “*a photo of a(n)* {*s*} with

---

**Algorithm 1** Deep Evidential Fusion Network

---

**Input:** Distorted image  $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$

**Output:** Predicted quality scores  $\hat{q}(\mathbf{x}) \in \mathbb{R}$

- 1: # Stage I: Local and Global Probability Scores (Section 3.1)
- 2: Crop to obtain local sub-images  $\{\mathbf{x}_i^{\text{local}}\}_{i=1}^N$ , downsample  $\mathbf{x}$  to generate the global image  $\mathbf{x}^{\text{global}}$
- 3: Use CLIP to obtain  $\text{logit}(c, s, d | \mathbf{x}')$ , and then derive the joint probability with Eq. (4):

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

where  $\mathbf{x}'$  denotes for either the local sub-image  $\mathbf{x}_i^{\text{local}}$  or the global image  $\mathbf{x}^{\text{global}}$

- 4:
- 5: # Stage II: Multitask Optimization (Section 3.2)
- 6: Compute the BIQA loss  $\ell_q$  for each image pair, as defined in Eq. (7):

$$\ell_q(\mathbf{x}_1, \mathbf{x}_2; \theta) = 1 - \sqrt{p(\mathbf{x}_1, \mathbf{x}_2) \hat{p}(\mathbf{x}_1, \mathbf{x}_2)} - \sqrt{(1 - p(\mathbf{x}_1, \mathbf{x}_2))(1 - \hat{p}(\mathbf{x}_1, \mathbf{x}_2))}$$

- 7: Compute the scene classification loss  $\ell_s$  for each image, as defined in Eq. (10):

$$\ell_s(\mathbf{x}; \theta) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left( 1 - \sqrt{p(s | \mathbf{x}) \hat{p}(s | \mathbf{x})} - \sqrt{(1 - p(s | \mathbf{x}))(1 - \hat{p}(s | \mathbf{x}))} \right)$$

- 8: Compute the distortion type classification loss  $\ell_d$  for each image, as defined in Eq. (12):

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

- 9: Combine auxiliary tasks to multitask loss, which is specifically defined in Eq. (13):

$$\mathcal{L}^M(\theta) = \frac{1}{|\mathcal{P}|} \sum_{(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{P}} \lambda_q \ell_q(\mathbf{x}_1, \mathbf{x}_2; \theta) + \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} [\lambda_s \ell_s(\mathbf{x}) + \lambda_d \ell_d(\mathbf{x})],$$

10:

- 11: # Stage III: Cross Sub-regions Information Fusion (Section 3.3)

- 12: For BIQA task ( $t = q$ ), derive the evidential loss on local outputs, as defined in Eq. (21),

$$\mathcal{L}_q^U(\theta) = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} \ell^U(\text{softplus}(\mathbf{x}_q^{\text{cross}}), \mathbf{y}_q, \theta),$$

where  $\mathbf{x}_q^{\text{cross}}$  is obtained with Eq. (18), and  $\ell^U$  is defined in Eq. (3)

- 13: For scene classification task ( $t = s$ ), derive the evidential loss, as defined in Eq. (21):

$$\mathcal{L}_s^U(\theta) = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} \ell^U(\text{softplus}(\mathbf{x}_s^{\text{cross}}), \mathbf{y}_s, \theta)$$

- 14: For distortion type classification task ( $t = d$ ), derive the evidential loss, as defined in Eq. (21)

$$\mathcal{L}_d^U(\theta) = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} \ell^U(\text{softplus}(\mathbf{x}_d^{\text{cross}}), \mathbf{y}_d, \theta)$$

- 15: Integrate sub-region information to the cross-region loss, as defined in Eq. (22):

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


---

- 
- 16: # Stage IV: Local-global Information Fusion (Section 3.4)  
 17: Derive the evidential loss on local-global outputs for BIQA task ( $t = q$ ), as defined in Eq. (26):

$$\mathcal{L}_q^F(\theta) = \sum \ell^U (\text{softplus}(\mathbf{x}_q^{\text{fusion}}), \mathbf{y}_q, \theta),$$

- where  $\mathbf{x}_q^{\text{fusion}}$  is obtained with Eq. (25), and  $\ell^U$  is defined in Eq. (3)  
 18: Derive the evidential loss for scene classification task ( $t = s$ ), as defined in Eq. (26):

$$\mathcal{L}_s^F(\theta) = \sum \ell^U (\text{softplus}(\mathbf{x}_s^{\text{fusion}}), \mathbf{y}_s, \theta)$$

- 19: Derive the evidential loss  $\mathcal{L}_d^F(\theta)$  for distortion type classification task ( $t = d$ ), as defined in Eq. (26):

$$\mathcal{L}_d^F(\theta) = \sum \ell^U (\text{softplus}(\mathbf{x}_d^{\text{fusion}}), \mathbf{y}_d, \theta)$$

- 20: Integrate local-global information to the cross-grained loss, as defined in Eq. (27):

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

- 21:  
 22: # Stage V: Overall Loss (Section 3.5)  
 23: Calculate the overall loss, as defined in Eq. (28):

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

- 24: Update parameters in DEFNet through gradient feedback based on  $\mathcal{L}(\theta)$   
 25:  
 26: # Give predict quality score, as defined in Eq. (5)  
 27: **Return**  $\hat{q}(\mathbf{x}) = \sum_{c=1}^C \hat{p}(c|\mathbf{x}) \times c$
- 

$\{d\}$  artifacts, which is of  $\{c\}$  quality.“, where  $s$  denotes scene type,  $d$  denotes distortion type,  $c$  denotes quality levels. These three variables take arbitrary values in their respective sets. Specifically, the scene type  $s \in \mathcal{S} = \{\text{“animal”}, \text{“cityscape”}, \text{“human”}, \text{“indoor scene”}, \text{“landscape”}, \text{“night scene”}, \text{“plant”}, \text{“still-life”}, \text{and “others”}\}$ , where there are altogether 9 choices. The distortion type  $d \in \mathcal{D} = \{\text{“blur”}, \text{“color-related”}, \text{“contrast”}, \text{“JPEG compression”}, \text{“JPEG2000 compression”}, \text{“noise”}, \text{“over-exposure”}, \text{“quantization”}, \text{“under-exposure”}, \text{“spatially-localized”}, \text{and “others”}\}$ , where there are altogether 11 choices. It is worth noting that the image belongs to the last category “others” if no distortion artifact is performed. The quality level  $c \in \mathcal{C} = \{1, 2, 3, 4, 5\}$ , which corresponds to levels of “bad”, “poor”, “fair”, “good” and “perfect”, respectively.

## E. More Experimental Results

In this section, we present some additional experimental results as supplementary to Section 4:

- Supplementary E.1 presents the results of the gMAD competition in SPAQ, supplementing Section 4.4.
- Supplementary E.2 presents the comparison on distortion

types in LIVE, supplementing Section 4.6.

- Supplementary E.3 presents complete results in ablation study, supplementing Section 4.7.
- Supplementary E.4 presents clear scatter plots and confidence intervals, supplementing Section 4.9.
- Supplementary E.5 presents additional model complexity analysis.

### E.1. gMAD Competition in SPAQ

The results of the gMAD competition in SPAQ are shown in Figure 7. In the experimental settings, SPAQ is a representative of authentic distortion datasets. Similar conclusions can be drawn from the SPAQ’s gMAD comparisons as those in the WED in Section 4.4.

### E.2. Comparison on Distortion Types in LIVE

The performance of DEFNet, as well as several state-of-the-art methods, across diverse distortion types in the LIVE dataset is shown in Table 7. The distortion types in LIVE include white noise (WN), fast fading (FF), Gaussian blur (GB), JPEG compression (JPEG) and JPEG2000 compression (JP2K). In the LIVE dataset, DEFNet exhibits the highest SRCC values across a majority of distortion types.

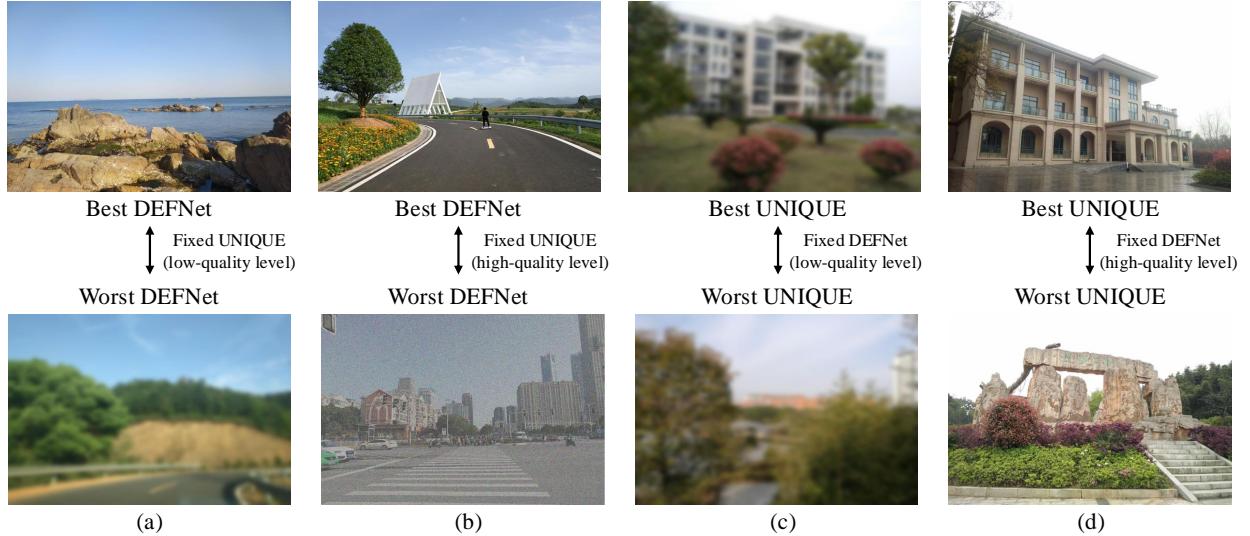


Figure 7. gMAD competition between UNIQUE and DEFNet in SPAQ. (a) Fixed UNIQUE at low-quality level. (b) Fixed UNIQUE at high-quality level. (c) Fixed DEFNet at low-quality level. (d) Fixed DEFNet at high-quality level.

Table 7. SRCC performance on diverse distortion types of LIVE.

Distortion	WN	FF	GB	JPEG	JP2K
BRISQUE [39]	0.977	0.877	0.951	0.965	0.914
ILNIQE [64]	0.975	0.827	0.911	0.931	0.902
deepIQA [3]	0.979	0.897	0.970	0.953	0.968
PQR [62]	0.981	0.921	0.944	0.965	0.953
DBCNN [65]	0.980	0.930	0.935	0.972	0.955
HyperIQA [51]	0.982	0.934	0.926	0.961	0.949
OLNet [59]	0.984	0.930	0.931	0.976	0.970
VCRNet [41]	0.988	<b>0.962</b>	<b>0.978</b>	0.979	0.975
<b>DEFNet</b>	<b>0.989</b>	0.942	0.971	<b>0.985</b>	<b>0.983</b>

### E.3. Ablation Study Results

As shown in Section 4.7 and Table 4, there are altogether 16 ablation settings in terms of the combination of auxiliary tasks and loss components. The SRCC and PLCC performance within different combinations of auxiliary tasks on the six IQA datasets is shown in Figure 8 and Figure 9. The accuracy for specific auxiliary tasks (scene and distortion classification) is shown in Figure 10. In most cases, with the addition of cross region loss ( $+L^U$ ) and cross-grained loss ( $+L^F$ ), the performance of the model improves.

### E.4. Uncertainty Results

Scatter plots between the predicted and ground-truth scores, and their 95% confidence intervals are shown in Figure 11. The experimental results indicate that the introduction of the two-level trustworthy evidential fusion reduces the uncertainty of model predictions. By applying evidence theory to image quality assessment tasks and utilizing the four di-

Table 8. Model complexity comparison.

Methods	# Params (M)
TReS [11]	152.45
CDINet [68]	99.62
TSFE [29]	91.75
LIQE [67]	59.02
<b>DEFNet</b>	84.22

mensions of data distribution, final predictions are made for the three tasks of image quality assessment. By combining aleatoric and epistemic uncertainty with evidential learning, an optimization is carried out for the final prediction, allowing the model to better focus on the parts with significant fluctuations in the prediction results and focus on learning relevant regions. In addition, reallocating uncertainty between cross sub-regions and different granularity fusion can also enable targeted optimization of the model.

### E.5. Model Complexity

In this section, we analyze the model complexity of the proposed DEFNet and several state-of-the-art methods. Table 8 presents the comparison results in terms of the number of parameters (# Params), which provide insights into the computational efficiency and resource demands of each model. It can be observed that DEFNet, with 84.22M parameters, is higher than LIQE in terms of the number and complexity of model parameters. In comparison with TReS [11], CDINet [68] and TSFE [29], DEFNet is still competitive. Despite this, there is still room for improvement in parameter and complexity optimization.

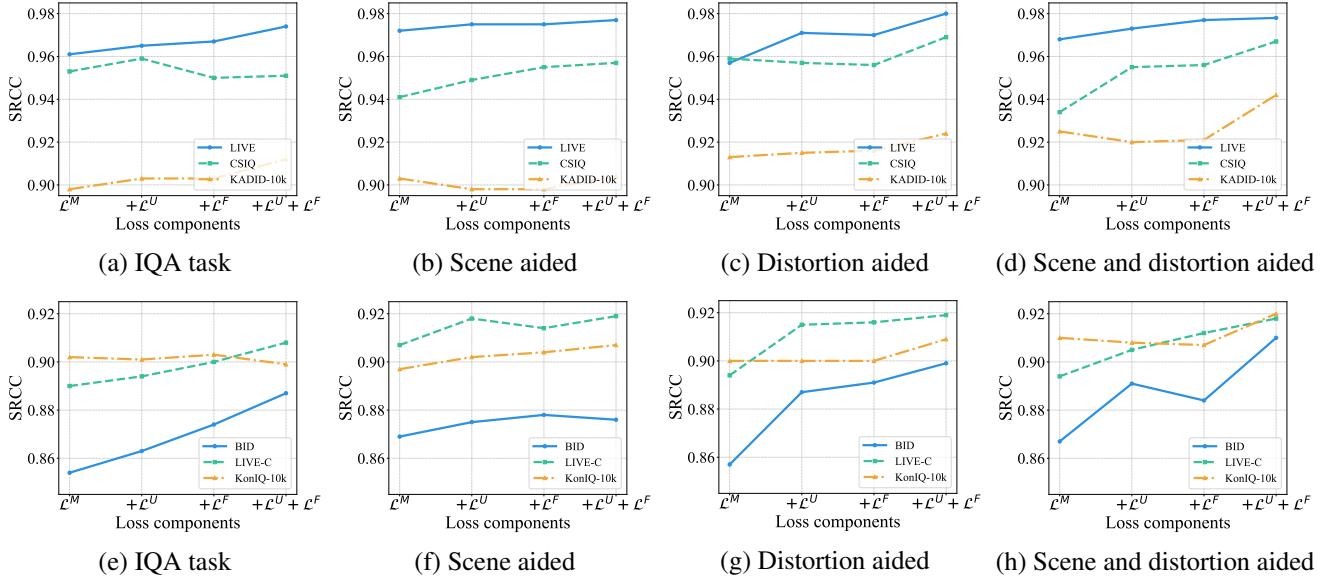


Figure 8. SRCC performance among different combinations of task assistance for synthetic distortion (a-d) and authentic distortion (e-h).

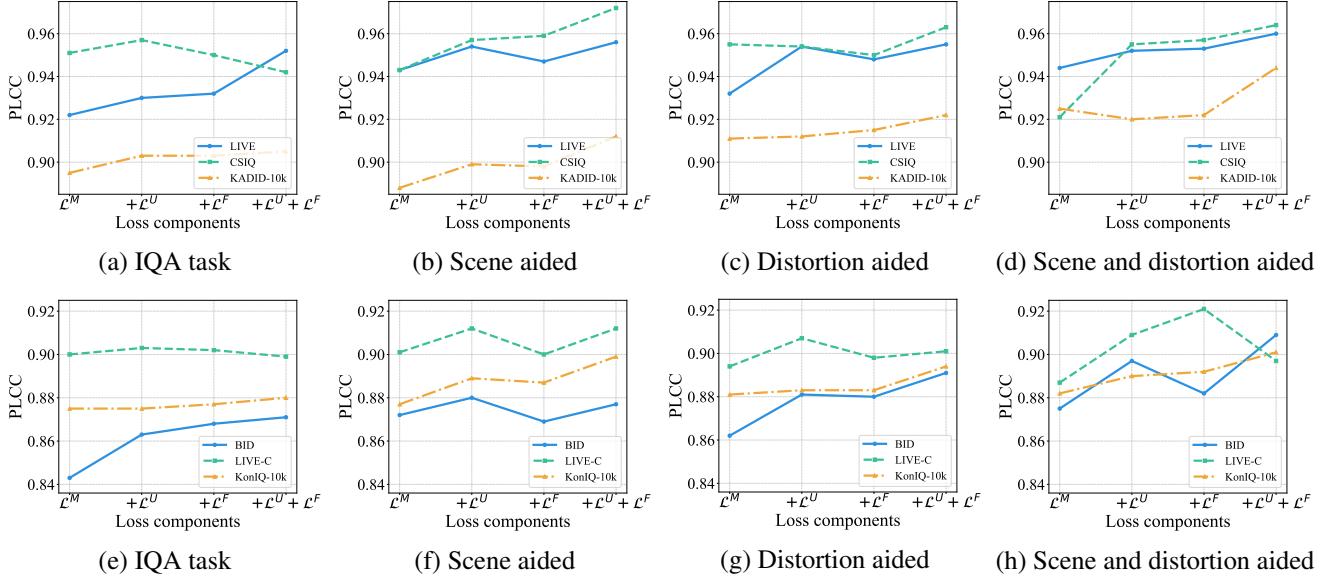


Figure 9. PLCC performance among different combinations of task assistance for synthetic distortion (a-d) and authentic distortion (e-h).

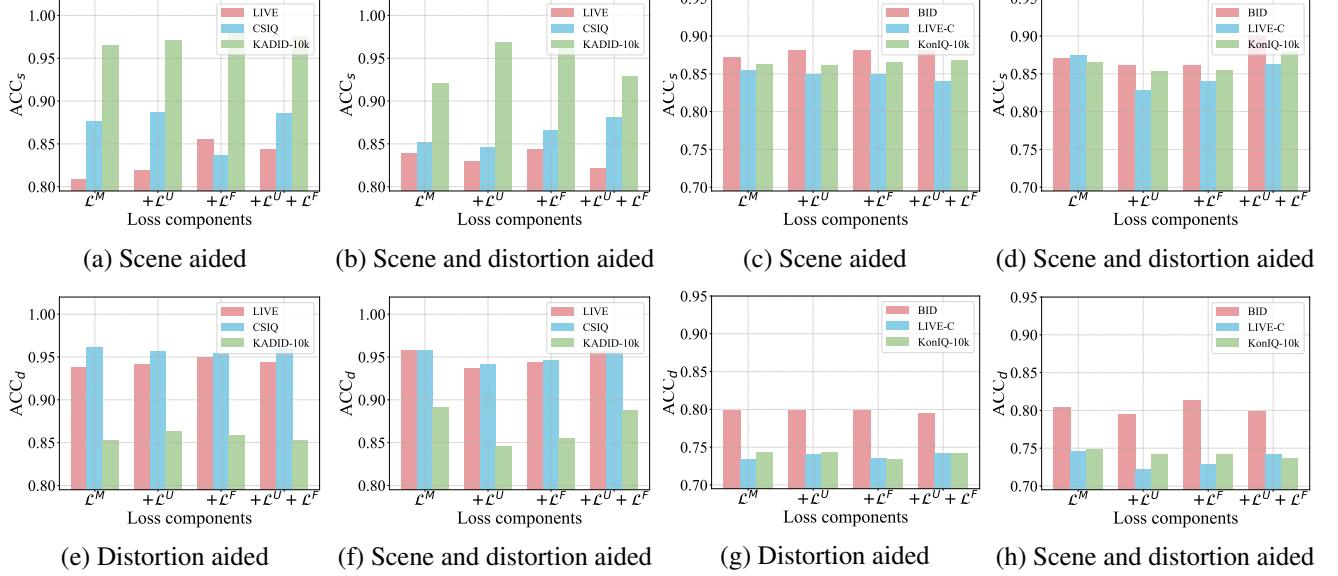


Figure 10. Accuracy performance in scene classification (a-d) and distortion classification (e-h) among different combinations of task assistance for both synthetic and authentic distortion datasets.

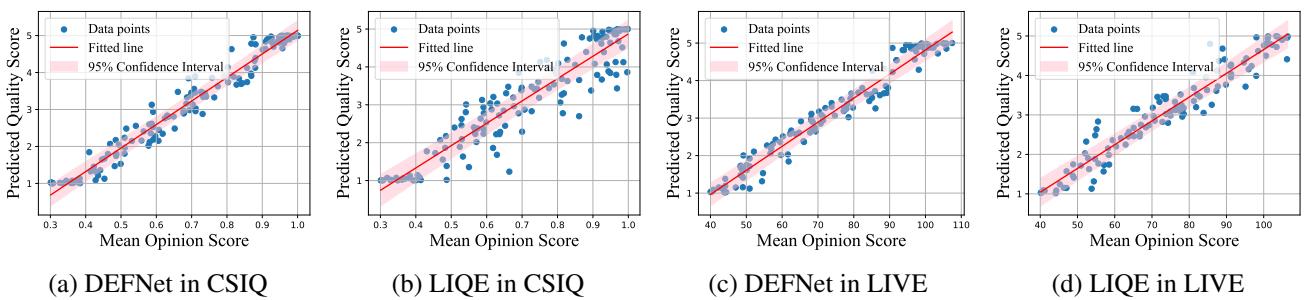


Figure 11. Scatter plots and 95% confidence intervals for DEFNet and LIQE [67] in CSIQ and LIVE datasets.