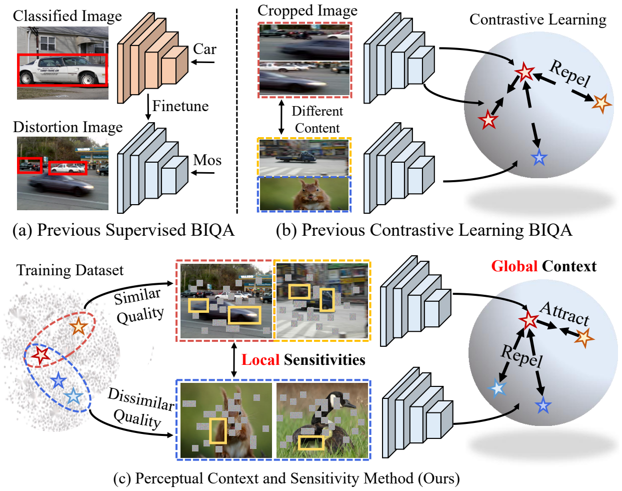
Integrating Global Context Contrast and Local Sensitivity for Blind Image Quality Assessment

Xudong Li [* 1] Runze Hu [* 2] Jingyuan Zheng [3] Yan Zhang [1] Shengchuan Zhang [1] Xiawu Zheng [1] Ke Li [4]
Yunhang Shen [4] Yutao Liu [5] Pingyang Dai [1] Rongrong Ji [1]

Abstract

Blind Image Quality Assessment (BIQA) mirrors subjective made by human observers. Generally, humans favor comparing relative qualities over predicting absolute qualities directly. However, current BIQA models focus on mining the "local" context, i.e., the relationship between information among individual images and the absolute quality of the image, ignoring the "global" context of the relative quality contrast among different images in the training data. In this paper, we present the Perceptual Context and Sensitivity BIQA (CSIQA), a novel contrastive learning paradigm that seamlessly integrates "global" and "local" perspectives into the BIQA. Specifically, the CSIQA comprises two primary components: 1) A Quality Context Contrastive Learning module, which is equipped with different contrastive learning strategies to effectively capture potential quality correlations in the **global context** of the dataset. 2) A Quality-aware Mask Attention Module, which employs the random mask to ensure the consistency with visual **local sensitivity**, thereby improving the model's perception of local distortions. Extensive experiments on eight standard BIQA datasets demonstrate the superior performance to the state-of-the-art BIQA methods.



(a) Previous Supervised BIQA (b) Previous Contrastive Learning BIQA

(c) Perceptual Context and Sensitivity Method (Ours)

Figure 1. The illustration of the difference between previous BIQA methods and ours. (a) Previous supervised BIQA methods predominantly rely on semantic features to directly learn image quality, neglecting the inherent **Global context contrastive**. (b) Previous BIQA with distortion-based contrastive learning focuses on maximizing similarity between different views of the same image but ignores similarity across different content images with similar quality. (c) In CSIQA, we fully explore the above two properties by using quality-based contrastive learning with a fine-grained contrast sample selection mechanism and quality-aware masked attention, which obtain comprehensive quality-aware features.

1. Introduction

Image Quality Assessment (IQA) (Madhusudana et al., 2022; Ding et al., 2020; Gu et al., 2020; Hu et al., 2021) is devoted to estimating the perceptive quality of a digital image consistent with the human visual system (HVS). It has been applied in many computer vision pieces of research (Hu et al., 2022), such as image restoration (Banham & Katsaggelos, 1997) and super-resolution (Dong et al., 2015). IQA methods are generally classified into three types based on the availability of reference images, i.e. full-reference (Sheikh et al., 2006), reduced-reference (Wang et al., 2016), and no-reference or blind IQA (BIQA) (Wu et al., 2020; Liu et al., 2024).

[*] Equal contribution [1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China. [2] School of Information and Electronics, Beijing Institute of Technology, Beijing 100080, China. [3] School of Medicine, Xiamen University. [4] Tencent Youtu Lab. [5] School of Computer Science and Technology, Ocean University of China.. Correspondence to: Yan Zhang <bzhy986@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

In practical settings, reference images are often not available, making reference-free BIQA more relevant and useful. Yet, two main challenges persist in current BIQA methods. **1)Global Context.** Existing methods typically train deep learning models by treating each sample as independent. However, for a more accurate assessment of quality, it's crucial to consider the contrastive relationships between different quality samples (Fig. 1(c)), since humans are better adept at learning image quality representations through the process of comparing the quality of different images (Sheikh & Bovik, 2006; Yin et al., 2022). **2)Local Sensitivity.** Current BIQA methods utilize domain knowledge from large datasets like ImageNet (Deng et al., 2009) to learn about quality representation (Fig. 1(a)). However, the semantic features extracted from pre-trained models are sub-optimal for BIQA (Zhao et al., 2023), since BIQA preferentially depends on distortion-sensitive attributes to perceive the visual quality of pictures with different semantic contexts (Su et al., 2020; Zhang et al., 2023).

Many state-of-the-art (SOTA) BIQA methods have been proposed to address these two challenges. Some approaches (Su et al., 2020; Qin et al., 2023; Ke et al., 2021) attempt to interpret image quality by decoding the abstract semantic knowledge in pre-trained models. However, they may overlook the quality relevance of the dataset context (i.e., Challenge 1). Recent innovations have delved into global quality relations using rank learning (Liu et al., 2017; Zhang et al., 2021; 2023) and contrastive learning (Madhusudana et al., 2022; Zhao et al., 2023; Saha et al., 2023), contributing significantly to the domain. However, these methods are not without their limitations. For example, approaches employing rank learning could theoretically be susceptible to noise in IQA data (Theorem A.2 in the Appendix), potentially impacting the optimal learning of models. Similarly, as shown in Fig. 1(b), existing methods that utilize contrastive learning typically define positive and negative samples based on distortion type and content. However, these definitions are not strict quality labels, potentially leading to inaccurate representation learning. Specifically, these methods treat authentic images with different content as negative samples in contrastive learning, without considering that images with different content can possess similar quality. This oversight can hinder the model's ability to capture the fine-grained relationships between different quality levels. Addressing these gaps, we propose the Perceptual Context and Sensitivity BIQA (CSIQA), a new methodology that maximizes the use of quality ground truths to enhance the model's awareness of global context and local perception.

To address the Challenge **1)**, we introduce the **Quality Context Contrastive Learning** approach to explore the global context relationship. This approach constructs easy positive and negative sets based on quality-level similarities. To mitigate overfitting, we incorporate a **Hard Sampling Strategy** within a knowledge distillation framework. This strategy involves sampling hard positive (negative) samples that have similar (different) quality but different (similar) pseudo-labels generated by pre-trained teacher reasoning. By focusing on these challenging samples, our model becomes more adept at distinguishing subtle differences in quality and effectively utilizing the sample information. We then employ InfoNCE (Oord et al., 2018) to capture the global properties in the embedding space, reflecting the training data's inherent structure, resulting in more precise quality predictions. To address the Challenge **2)**, we propose the **Quality-Aware Mask Attention** method that designs a new quality-aware (QUA) token to help the model enforce models focusing on patch-level localized distortions rather than image-level semantic information. The interaction mode of the QUA token is similar to that of the class (CLS) token during the forwarding process, except that it only targets visible patches that have not been masked. This approach, combined with quality context contrastive learning, enhances quality perception and information exploration. We summarize the contributions of this work as follows:

- We identify the limitations of existing BIQA methods. In contrast to previous approaches that solely concentrate on single-image analysis, we delve into exploring and analyzing the cross-image context quality correlation as a prior, which offers valuable insights for enhancing the accuracy of quality prediction.
- We propose a novel quality context contrastive learning that leverages abundant quality-aware contrastive information. This paradigm offers a well-structured quality embedding space to enhance the model's discrimination of quality features.
- A novel quality-aware mask attention mechanism is introduced to combine contrastive learning for improving quality perception. By randomly masking patches, the designed quality-aware token can effectively emphasize low-level quality features that complement high-level semantic information.

Extensive experiments on eight IQA benchmarks demonstrate the superior performance of CSIQA compared to the existing method and confirm the effectiveness of our proposed method. In addition, our quality context contrastive learning method can be seamlessly incorporated into existing BIQA networks without any changes to the base model and extra inference burden during testing (Table 2).

2. Related Work

2.1. IQA with Local Perspective

Early BIQA methods (Zeng et al., 2018; Liu et al., 2017), leveraged convolutional neural networks (CNNs) for their strong feature representation capabilities. CNN-based BIQA methods, like those by (Zhang et al., 2018) and HyperNet, typically pre-train for object recognition and then fine-tune for IQA. However, CNNs have an intrinsic locality bias, limiting their ability to capture long-range dependencies essential for assessing image quality (Qin et al., 2023). Recently, Vision Transformers (ViTs) have emerged in BIQA with two main architectures: hybrid transformers (Golestaneh et al., 2022) and pure transformers (Ke et al., 2021; You & Korhonen, 2021). The hybrid approach combines CNNs for local features with transformers for long-range features. Pure ViT-based IQA methods often depend on the CLS token, initially designed for image content description, which focuses on higher-level visual abstractions and can limit their effectiveness in BIQA applications.

2.2. IQA with a Global Contrastive Perspective

In the area of IQA, ranking-based learning is crucial for modeling global quality relationships. Studies like (Zhang et al., 2018; 2023; 2021) have utilized discrete ranking data from images with varying distortion levels or sorted label pairs for quality prediction. Continuous ranking data, based on Mean Opinion Scores (MOS) and subjective rating differences, was employed by (Zhang et al., 2021). Binary ranking data, informed by Full-Reference IQA (FR-IQA) methods, was integrated in the research of (Ma et al., 2017b; 2019). While effective, these methods can be constrained in datasets with authentic distortions due to their reliance on reference images. Moreover, rank-based IQA methods like those in (Liu et al., 2017; Ma et al., 2017a; Golestaneh et al., 2022) often necessitate manual hyperparameter tuning, increasing their vulnerability to noisy data. The choice of ranked image pairs is also a critical aspect, as random selection can introduce further noise (Theorem A.2).

Several works, including (Madhusudana et al., 2022; Zhao et al., 2023; Saha et al., 2023), have successfully explored self-supervised contrastive learning in IQA. These studies relied on class labels based on distortion and degradation levels to define positive and negative samples—image pairs with the same distortions were considered positive samples, while those with different distortions were considered negative samples. Although these methods significantly advanced the field, their class definitions are, in fact, distortion labels rather than true-quality labels, potentially failing to capture the fine-grained relationships between different levels of quality. For example, they uniformly considered pairs of authentically distorted images with different content as negative samples, which hindered their ability to model the quality correlation between images with different content but similar quality. Additionally, these methods often incur high pre-training costs. In contrast, our method leverages image quality labels to define positive and negative samples. We classify samples with similar quality as positive samples and those with dissimilar quality as negative samples, and we further designed a more fine-grained sampling method for positive and negative samples to improve the model's ability to distinguish sample pairs with challenging quality relationships. Additionally, we integrated a quality-aware mask self-attention to enhance the model's perception of local features. This combination effectively captures quality differences between real images, resulting in a more comprehensive representation of real-world image features.

3. Methodology

3.1. Overview

In this paper, we introduce the Perceptual Context and Sensitivity IQA (CSIQA) designed to predict image quality from both global and local perspectives. As depicted in Fig. 2, CSIQA seamlessly integrates two main components: Global Quality Context Contrastive Learning and Local Quality-Aware Mask Attention. Initially, CSIQA divides the input image into patches, which are then fed into the mask transformer encoder. After processing through L layers of Quality-Aware Mask Attention (§ 3.2), we obtain the **local** sensitive feature $\mathbf{F}^l$ at the l-th layer. Once $\mathbf{F}^l$ is acquired, it is promptly passed to the Quality Context Contrastive Learning module (§ 3.3). Here, the query embedding $\mathbf{F}^l$ and its corresponding positive $\mathbb{F}_{\mathrm{po}}$ ($\mathbb{F}_{\mathrm{po}} = \mathbb{F}^{\mathrm{e}}_{\mathrm{po}} \cup \mathbb{F}^{\mathrm{h}}_{\mathrm{po}}$) and negative $\mathbb{F}_{\mathrm{ne}}$ ($\mathbb{F}_{\mathrm{ne}} = \mathbb{F}^{\mathrm{e}}_{\mathrm{ne}} \cup \mathbb{F}^{\mathrm{h}}_{\mathrm{ne}}$) samples are used as inputs for the InfoNCE to model the **global** contrastive relation. The selection of easy ($\mathbb{F}^{\mathrm{e}}_{\mathrm{po}}$, $\mathbb{F}^{\mathrm{e}}_{\mathrm{ne}}$) and hard ($\mathbb{F}^{\mathrm{h}}_{\mathrm{po}}$, $\mathbb{F}^{\mathrm{h}}_{\mathrm{ne}}$) positive and negative samples is based on the similarity between ground truths and the alignment between ground truths and pseudo-labels, respectively (§ 3.4). Finally, the output $\mathbf{F}^L$ from the final encoder layer is refined further through a transformer decoder to predict the quality score.

3.2. Quality-Aware Mask Transformer

The self-attention of the transformer (Dosovitskiy et al., 2021) is effective in comprehensively representing perceptual features. The global semantic features extracted by a pre-trained network may not be optimal for capturing quality awareness, as HVS is highly sensitive to local distortions even when the overall image quality is good (Larson & Chandler, 2010; Su et al., 2020). Therefore, we propose the use of random mask (He et al., 2022) interactions with learnable tokens, which enable the learning of local quality-aware features that are independent of global semantic features.

Let $\mathbf{F} = \{\boldsymbol{F}_{qua}; [\boldsymbol{F}_{cls}; \boldsymbol{F}_{1:N-1}]\} \in \mathbb{R}^{(N+1)\times D}$ be the embedding sequence, where N is the number of patches plus one CLS token, D is the embedding dimension, $\boldsymbol{F}_{qua}$ is the learnable quality-aware (QUA) token. Three linear projection layers transform $\boldsymbol{F}$ into matrices $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ for query, key, and value. To regulate the interaction between $\boldsymbol{F}_{qua}$

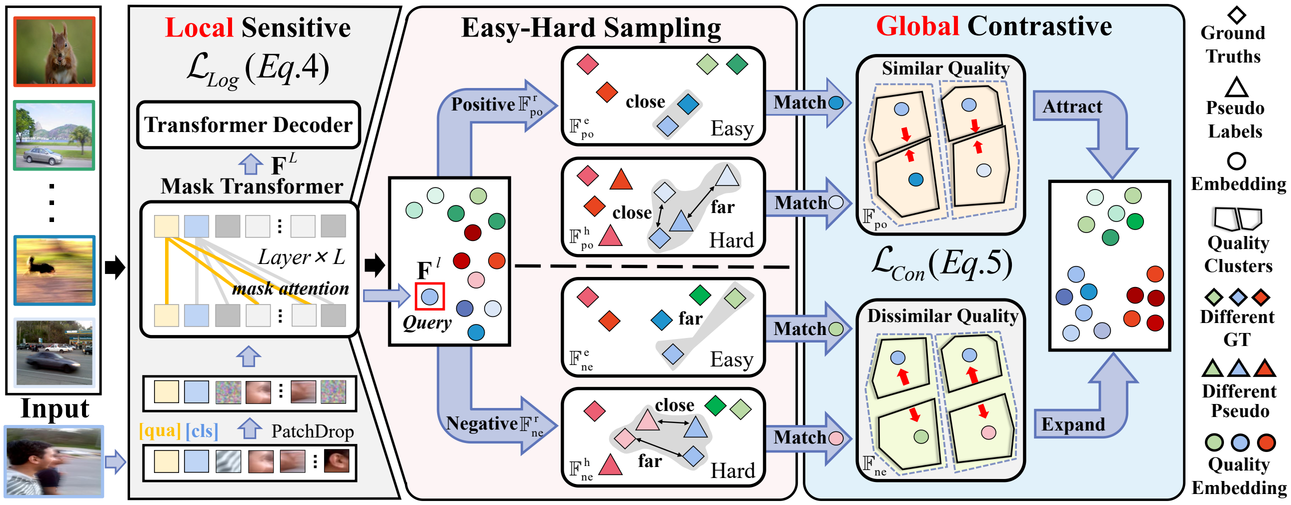


Figure 2. The overview of CSIQA, in which similar colors represent similar quality. For a given query image, CSIQA first obtains each layer's image features $\mathbf{F}^l$ through **Local** Quality-Aware Mask Attention (§ 3.2). Subsequently, we sample easy positive $\mathbb{F}^{\mathrm{e}}_{\mathrm{po}}$ (easy negative $\mathbb{F}^{\mathrm{e}}_{\mathrm{ne}}$) samples, which are very similar (dissimilar) in quality to the query image (§ 3.4). We also sample hard positive $\mathbb{F}^{\mathrm{h}}_{\mathrm{po}}$ (hard negative $\mathbb{F}^{\mathrm{h}}_{\mathrm{ne}}$) samples, which are similar in ground truth but dissimilar in pseudo-labels (or vice versa) to the query image (§ 3.4). These samples form positive $\mathbb{F}_{\mathrm{po}}$ and negative $\mathbb{F}_{\mathrm{ne}}$ sets for computing contrastive loss in **Global** Quality Context Contrastive Learning (§ 3.3).

and the other tokens, we construct a mask $\boldsymbol{m} \in \mathbb{R}^{(N+1)}$.

$$m_i = \begin{cases} 1 - \mathbb{1}\{\text{i-th masked}\}, & i < N \\ 1, & i = N \end{cases}, \tag{1}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, the probability of the random mask α is 0.3. To maintain the original attention, our attention mask $\mathbf{M}$ is constructed as follows:

$$\mathbf{M} = \begin{bmatrix} \mathbf{1}_{N\times N} & \mathbf{0}_{N\times 1} \\ \mathbf{m}_{1\times N} & 1 \end{bmatrix} \in \{0,1\}^{(N+1)\times(N+1)}. \tag{2}$$

Intuitively, Eq. 2 means that $\boldsymbol{F}_{qua}$ only attends to the patch tokens that are not randomly masked. Suppose the Encoder has L attention layers, the Multi-Head Self-Attention (MHSA) is utilized, as follows:

$$\mathbf{F}^l = \begin{cases} \sigma\left(\log \mathbf{M} + \mathbf{Q}^l(\mathbf{K}^l)^\top\right)\mathbf{V}^l + \mathbf{F}^{l-1}; & 0 < l \leq L \\ \mathbf{F}, & l = 0 \end{cases}, \tag{3}$$

where $\boldsymbol{Q}^l$, $\boldsymbol{K}^l$, and $\boldsymbol{V}^l$ are linear transformations of $\mathbf{F}^{l-1}$, σ represents the Softmax function. The encoder produces output features $\mathbf{F}^L$ after L layers. To provide meaningful interpretations for the CLS and QUA tokens, a transformer decoder is employed (Qin et al., 2023). Finally, the output features $\boldsymbol{F}_{qua}$ and $\boldsymbol{F}_{cls}$ from the decoder are concatenated to predict the final quality score. During training, the smooth $\mathcal{L}_1$ loss is minimized. The introduction of quality-aware mask attention enhances the learning capability of quality-aware features in the transformer-based BIQA model, resulting in improved prediction accuracy and generalization ability. Please refer to § B in the appendix for more details.

3.3. Global-Local Quality-Aware Learning

Logits-Based Supervised Loss. In our quality-aware mask attention mechanism, mask positions are generated randomly with a certain probability, inevitably introducing some noise. To mitigate this noise, we employ a knowledge distillation (KD) method, which enhances performance by regularizing the student's logits knowledge using the teacher model (Wang et al., 2019). To leverage this technique to IQA, we supervise our student model using both ground truth and pre-trained teacher-generated pseudo-labels. In order to indicate the effectiveness of this technique, we give detailed proof in Theorem A.1 in the Appendix. This analysis confirms that the expected loss with our method is notably reduced when contrasted with training that relies solely on ground truth. As a result, the reliability of the pseudo-labels is affirmed which facilitates rapid convergence and acts as a safeguard against overfitting. Given the teacher network and our student network produce mappings Y_T and Y_S for the image $\boldsymbol{I}$ to quality scores, respectively. Given $\hat{Y}$ as the ground truth label for image $\boldsymbol{I}$, λ_1 are the hyperparameters, $\|\cdot\|_1$ represents the ℓ_1 regression loss, the logits-based supervised loss defined as follows:

$$\mathcal{L}_{\mathrm{Log}} = \|\, \hat{Y} - Y_S \,\|_1 + \lambda_1 \|\, Y_T - Y_S \,\|_1, \tag{4}$$

Global Quality Context Contrastive Learning. Inspired by the human visual system, humans are better at making accurate quality predictions when presented with contrast (Sheikh & Bovik, 2006). The logits-based supervised loss functions focus solely on local intra-image analysis, neglecting the essential quality contrast global information between different images within the training dataset. In IQA, rank learning can be employed to model the global quality relationships across the dataset. However, rank learning heavily relies on the accuracy of rankings, which becomes challenging to achieve when comparing similar images (Gao et al., 2015). As per Theorem A.2 in the Appendix, the model is significantly influenced by the noise in ranking order. Our analysis suggests that this issue may be exacerbated by the use of randomly sampled samples for comparison without any selective filtering. Consequently, we introduce a quality context contrastive learning that optimally utilizes label information to selectively filter specific samples for comparison, enhancing the quality of learned embeddings. We define quality context Contrastive loss as follows:

$$\mathcal{L}_{\text{Con}} = \frac{-1}{|\mathbb{F}_{\text{po}}|} \sum_{\mathbf{F}^{+}\in\mathbb{F}_{\text{po}}} log\left(\frac{\exp(\frac{\mathbf{F}\cdot\mathbf{F}^{+}}{\tau})}{\exp(\frac{\mathbf{F}\cdot\mathbf{F}^{+}}{\tau}) + \sum\limits_{\mathbf{F}^{-}\in\mathbb{F}_{\text{ne}}} \exp(\frac{\mathbf{F}\cdot\mathbf{F}^{-}}{\tau})}\right), \tag{5}$$

where τ is a temperature hyper-parameter, $\mathbf{F}$ denotes the query image embedding, and $\mathbb{F}_{\text{po}}$ and $\mathbb{F}_{\text{ne}}$ represent the sets of image embeddings for positive and negative samples, respectively. The selection strategy for these samples is detailed in § 3.4. We leverage the distance ranking of quality labels between images to accurately select positive and negative samples. The overarching objective is to minimize the distance between samples of similar quality while maximizing the distance between those of differing quality, thereby refining the embedding space.

Overall Loss. The logits-based supervised loss in Eq. 4 and the contrastive loss in Eq. 5 complement each other. The $\mathcal{L}_{\text{Log}}$ helps the model identify relevant quality features from a "local" perspective, while the $\mathcal{L}_{\text{Con}}$ refines the embedding space from a "global" perspective, promoting compactness for similar-quality images and separability for different-quality ones. This approach aligns with human perception, enabling a global-to-local attentive evaluation. The hyperparameter λ_2 is used to balance $\mathcal{L}_{\text{Log}}$ and $\mathcal{L}_{\text{Con}}$. Details on the selection of λ_2 can be found in § E. Here, L denotes the number of encoder layers. The total loss is defined as:

$$\mathcal{L}_{\text{Scon}} = \sum\nolimits_{I} (\mathcal{L}_{\text{Log}} + \lambda_2 \sum_{l=1}^{L} \mathcal{L}_{\text{Con}}). \tag{6}$$

3.4. Easy-Hard Sampling Strategy

To implement Eq. 5, we first define a coarse-grained set of positive and negative samples. We then refine this set by employing **easy example sampling** and **hard example sampling** to select samples that are either easier or more challenging as the positive and negative sample sets for Eq. 5. Specifically, we define a given query sample's rough positive and negative sample sets by calculating the distance between the query and the quality scores of other samples in the current mini-batch and then ranking these distances.

Roughly Define Positive-Negative Samples. To roughly distinguish between positive and negative samples, we begin by extending the score vector $\boldsymbol{y} \in \mathbb{R}^{B\times 1}$ to a matrix $\boldsymbol{Y} \in \mathbb{R}^{B\times B}$, where B represents the batch size. The score distance matrix $\boldsymbol{Y}_d$ is then calculated using the Manhattan distance $\mathcal{D}$. For a given query image $\boldsymbol{I}_a$ and another image $\boldsymbol{I}_j$, we sort the a-th row of $\boldsymbol{Y}_d$ in ascending order to obtain the quality rank vector $\mathcal{Y}_d^a$. The index of the score distance between $\boldsymbol{I}_a$ and $\boldsymbol{I}_j$ in $\mathcal{Y}_d^a$ is denoted as C. If it is within the top $\gamma_1 \cdot B$ percentile of the $\mathcal{Y}_d^a$ vector, then $\boldsymbol{I}_j$ is considered a positive sample. Conversely, if the index C falls within the last $\gamma_2 \cdot B$ percentile of the $\mathcal{Y}_d^a$ vector, $\boldsymbol{I}_j$ is considered a negative sample. Here, γ_1 and γ_2 are hyperparameters, which we empirically set to 20% and 60%, respectively. The definition of the rough classifier is as follows:

$$\begin{aligned} \boldsymbol{Y}_d &= \mathcal{D}(\boldsymbol{Y} - \boldsymbol{Y}^T), \\ \text{Classifier}(\boldsymbol{I_a}, \boldsymbol{I_j}) &\in \begin{cases} \mathbb{F}_{\text{po}}^{\text{r}}, & \text{if } C \leq \gamma_1 \cdot B, \\ \mathbb{F}_{\text{ne}}^{\text{r}}, & \text{if } C \geq (1-\gamma_2)\cdot B. \end{cases} \end{aligned} \tag{7}$$

Easy-Hard Example Sampling. Previous studies (Schroff et al., 2015; Kalantidis et al., 2020) have highlighted the importance of selecting training samples in metric learning. When considering a query embedding $\mathbf{F}$, the gradient of the contrastive loss (Eq. 5) can be expressed as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{Con}}}{\partial \mathbf{F}} = &- \frac{1}{\tau\,|\mathbb{F}_{\text{po}}|} \sum\nolimits_{\mathbf{F}^{+}\in\mathbb{F}_{\text{po}}} \left((1-p^{+})\cdot \mathbf{F}^{+}\right) \\ &- \frac{1}{\tau\,|\mathbb{F}_{\text{po}}|} \sum\nolimits_{\mathbf{F}^{+}\in\mathbb{F}_{\text{po}}} \sum\nolimits_{\mathbf{F}^{-}\in\mathbb{F}_{\text{ne}}} p^{-}\cdot \mathbf{F}^{-}. \end{aligned} \tag{8}$$

The matching probability between a positive or negative sample $\mathbf{F}^{+/-}$ and the query embedding $\mathbf{F}$ represents as $p^{+/-} = \frac{\exp\left(\mathbf{F}\cdot\mathbf{F}^{+/-}/\tau\right)}{\sum_{\mathbf{F}'\in\mathbb{F}_{\text{po}}\cup\mathbb{F}_{\text{ne}}}\exp\left(\mathbf{F}\cdot\mathbf{F}'/\tau\right)} \in [0,1]$. Hard negative and positive samples (*i.e.*, negatives but similar features and positives but dissimilar features) have a higher contribution to the gradient compared to easy negative and positive samples (Wang et al., 2021). Inspired by this, we aim to explore how learning from hard samples helps the model acquire more quality-aware features. To identify hard samples of query embedding, we employ a distillation framework for sampling, which selects positive (negative) samples that have similar (different) quality but different (similar) pseudo-labels generated by the teacher for contrastive learning. By focusing on these challenging samples, the model can better distinguish fine-grained quality differences. Furthermore, to reduce the risk of the model becoming stuck in local optima, we also incorporate easy samples into the sampling process.

- **Easy Example Sampling.** The sampling process is similar to that described in Eq. 7. For each query embedding $\mathbf{F}$, we sample the top 10% and the last 40% of available samples from the $\boldsymbol{Y}_d$ matrix sorted in ascending order as easy positive samples $\mathbb{F}_{\text{po}}^{\text{e}}$ and negative samples $\mathbb{F}_{\text{ne}}^{\text{e}}$.

Table 1. Performance comparison measured by averages of SRCC and PLCC, where bold entries indicate best results, underlines indicate the second-best.

	LIVE		CSIQ		TID2013		KADID		LIVEC		KonIQ		LIVEFB		SPAQ	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BRISQUE (Mittal et al., 2012)	0.944	0.929	0.748	0.812	0.571	0.626	0.567	0.528	0.629	0.629	0.685	0.681	0.341	0.303	0.817	0.809
ILNIQE (Zhang et al., 2015)	0.906	0.902	0.865	0.822	0.648	0.521	0.558	0.534	0.508	0.508	0.537	0.523	0.332	0.294	0.712	0.713
BIECON (Kim & Lee, 2016)	0.961	0.958	0.823	0.815	0.762	0.717	0.648	0.623	0.613	0.613	0.654	0.651	0.428	0.407	-	-
MEON (Ma et al., 2017b)	0.955	0.951	0.864	0.852	0.824	0.808	0.691	0.604	0.710	0.697	0.628	0.611	0.394	0.365	-	-
WaDIQaM (Bosse et al., 2017)	0.955	0.960	0.844	0.852	0.855	0.835	0.752	0.739	0.671	0.682	0.807	0.804	0.467	0.455	-	-
DBCNN (Zhang et al., 2018)	0.971	0.968	0.959	0.946	0.865	0.816	0.856	0.851	0.869	0.851	0.884	0.875	0.551	0.545	0.915	0.911
TIQA (You & Korhonen, 2021)	0.965	0.949	0.838	0.825	0.858	0.846	0.855	0.85	0.861	0.845	0.903	0.892	0.581	0.541	-	-
MetaIQA (Zhu et al., 2020)	0.959	0.960	0.908	0.899	0.868	0.856	0.775	0.762	0.802	0.835	0.856	0.887	0.507	0.54	-	-
P2P-BM (Ying et al., 2020)	0.958	0.959	0.902	0.899	0.856	0.862	0.849	0.84	0.842	0.844	0.885	0.872	0.598	0.526	-	-
HyperIQA (Su et al., 2020)	0.966	0.962	0.942	0.923	0.858	0.840	0.845	0.852	0.882	0.859	0.917	0.906	0.602	0.544	0.915	0.911
TReS (Golestaneh et al., 2022)	0.968	0.969	0.942	0.922	0.883	0.863	0.858	0.859	0.877	0.846	0.928	0.915	0.625	0.554	-	-
MUSIQ (Ke et al., 2021)	0.911	0.940	0.893	0.871	0.815	0.773	0.872	0.875	0.746	0.702	0.928	0.916	0.661	0.566	0.921	0.918
DACNN (Pan et al., 2022)	0.980	0.978	0.957	0.943	0.889	0.871	0.905	0.905	0.884	0.866	0.912	0.901	-	-	0.921	0.915
CONTRIQUE (Madhusudana et al., 2022)	0.961	0.960	0.955	0.942	0.857	0.843	**0.937**	**0.934**	0.857	0.845	0.906	0.894	0.641	0.580	0.919	0.914
DEIQT (Qin et al., 2023)	0.982	0.980	0.963	0.946	0.908	0.892	0.887	0.889	0.894	0.875	0.934	0.921	0.663	0.571	0.923	0.919
Re-IQA (Saha et al., 2023)	0.971	0.970	0.960	0.947	0.861	0.804	0.885	0.872	0.854	0.840	0.923	0.914	**0.733**	**0.645**	0.925	0.918
CSIQA (Ours)	**0.985**	**0.983**	**0.970**	**0.960**	**0.941**	**0.926**	0.919	0.919	**0.920**	**0.898**	**0.943**	**0.929**	0.688	0.594	**0.935**	**0.930**

- **Hard Example Sampling.** We generate a pseudo-scores vector $\hat{\boldsymbol{y}}$ using the teacher network and expand it into a matrix $\hat{\boldsymbol{Y}}$, which is then used to compute $\hat{\boldsymbol{Y}}_{\boldsymbol{d}}$, similarly to Eq. 7. The final error matrix is $\boldsymbol{Y}_{\text{error}} = \frac{\boldsymbol{Y}_d}{\hat{\boldsymbol{Y}}_d}$. We select the top 10% and last 40% samples from the Y_{error} matrix sorted in ascending order to obtain hard samples $\mathbb{F}^{\text{h}}_{\text{po}}$ and $\mathbb{F}^{\text{h}}_{\text{ne}}$, which is then intersected with the set $\mathbb{F}^{\text{r}}_{\text{po}}$ and $\mathbb{F}^{\text{r}}_{\text{ne}}$ in Eq. 7, respectively, to reduce sampling deviation.

- **Easy-Hard Example Merging.** For contrastive learning, we combine easy positives $\mathbb{F}^{\text{e}}_{\text{po}}$ and hard positives $\mathbb{F}^{\text{h}}_{\text{po}}$ into the set $\mathbb{F}_{\text{po}} = \mathbb{F}^{\text{e}}_{\text{po}} \cup \mathbb{F}^{\text{h}}_{\text{po}}$, and similarly, easy negatives $\mathbb{F}^{\text{e}}_{\text{ne}}$ and hard negatives $\mathbb{F}^{\text{h}}_{\text{ne}}$ into the set $\mathbb{F}_{\text{ne}} = \mathbb{F}^{\text{e}}_{\text{ne}} \cup \mathbb{F}^{\text{h}}_{\text{ne}}$. In Eq. 5, we aim to position similar samples $\mathbb{F}_{\text{po}}$ closer and dissimilar samples $\mathbb{F}_{\text{ne}}$ farther apart. This approach of integrating both easy and difficult samples ensures that the model not only focuses on challenging examples but also reduces the risk of overfitting.

4. Experiments

4.1. Benchmark Datasets and Evaluation Protocols

Our method is evaluated on eight public BIQA datasets, including four synthetic datasets and four authentic datasets. The synthetic datasets are LIVE (Sheikh et al., 2006), CSIQ (Larson & Chandler, 2010), TID2013 (Ponomarenko et al., 2015), and KADID (Lin et al., 2019). The authentic datasets are LIVEC (Ghadiyaram & Bovik, 2015), KONIQ (Hosu et al., 2020), LIVEFB (Ying et al., 2020), and SPAQ (Fang et al., 2020). The synthetic datasets involve original images distorted artificially using methods such as JPEG compression and Gaussian blur. LIVE and CSIQ consist of 779 and 866 synthetically distorted images, respectively, each with five and six distortion types. TID2013 and KADID contain 3,000 and 10,125 synthetically distorted images, respectively, covering 24 and 25 distortion types. For the authentic datasets, LIVEC contains 1,162 images captured by different mobile devices and photographers. SPAQ includes 11,125 photos from 66 smartphones. KonIQ comprises 10,073 images from public sources, while LIVEFB is the largest authentic dataset to date, with 39,810 images. Performance is assessed using Spearman's Rank Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC), both ranging from -1 to 1. Higher values indicate superior accuracy and monotonicity.

4.2. Implementation Details

Following (Qin et al., 2023; Ke et al., 2021), we employ a pre-trained encoder based on ViT-S (Touvron et al., 2022) as the mask encoder of our CSIQA, with a depth of 12, and the Decoder depth is set to one. We train the model for 9 epochs using a learning rate of 2×10^{-4} and a decay factor of 10 every 3 epochs. The size of the batch depends on the size of the dataset, ranging from 16 to 128. We split the datasets into 80% for training and 20% for testing, repeating the process ten times to reduce performance bias. We report the average of SRCC and PLCC to quantify the model's performance in terms of prediction accuracy and monotonicity. For synthetic distortion datasets, training and testing sets are divided by reference images for content independence. More details about the student and teacher network's structure, training, performance, and selection principle are in Appendix C.

4.3. Overall Prediction Performance Comparison

The results of the comparison between CSIQA and 16 classical or state-of-the-art BIQA methods, which include hand-

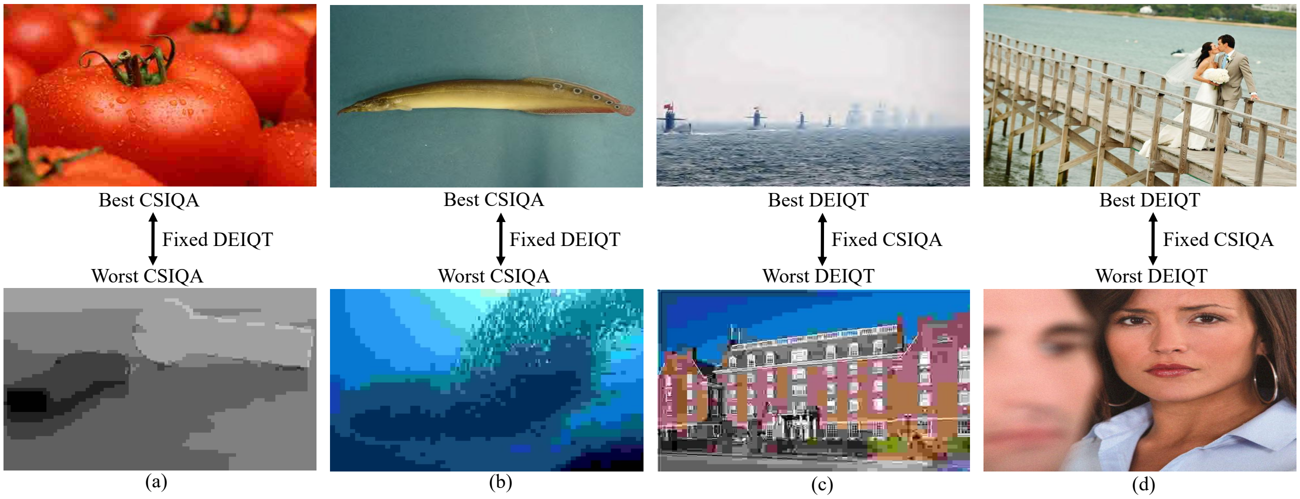


Figure 3. gMAD competition results between DEIQT (Qin et al., 2023) and CSIQA. (a) Fixed DEIQT at the low-quality level. (b) Fixed DEIQT at the high-quality level. (c) Fixed CSIQA at the low-quality level. (d) Fixed CSIQA at the high-quality level.

Table 2. Integration of quality context contrastive learning into IQA networks TIQA and DEIQT, both employing the VIT backbone. The "+" indicates the incorporation of our learning strategy.

	CSIQ		LIVEC		KonIQ	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
TIQA	0.945	0.930	0.881	0.863	0.930	0.914
TIQA +	**0.963**	**0.951**	**0.910**	**0.883**	**0.940**	**0.926**
	(+1.8)	(+2.1)	(+2.9)	(+2.0)	(+1.0)	(+1.2)
DEIQT	0.961	0.950	0.894	0.875	0.934	0.921
DEIQT +	**0.968**	**0.956**	**0.916**	**0.893**	**0.941**	**0.926**
	(+0.7)	(+0.6)	(+2.2)	(+1.8)	(+0.7)	(+0.5)

Table 3. SRCC on the cross datasets validation. The best performances are highlighted in boldface.

Training	LIVEFB		LIVEC	KonIQ	LIVE	CSIQ
Testing	KonIQ	LIVEC	KonIQ	LIVEC	CSIQ	LIVE
DBCNN	0.716	0.724	0.754	0.755	0.758	0.877
P2P-BM	0.755	0.738	0.74	0.77	0.712	-
HyperIQA	0.758	0.735	0.772	0.785	0.744	0.926
TReS	0.713	0.74	0.733	0.786	0.761	-
CONTRIQUE	-	-	0.676	0.731	**0.823**	0.925
DEIQT	0.733	0.781	0.744	0.794	0.781	0.932
CSIQA	**0.760**	**0.787**	**0.777**	**0.814**	0.818	**0.945**

crafted feature-based BIQA methods like ILNIQE (Zhang et al., 2015) and BRISQUE (Mittal et al., 2012), as well as deep learning-based methods such as MUSIQ (Ke et al., 2021) and DEIQT (Qin et al., 2023), are presented in Table 1. It can be observed on six of the eight datasets that CSIQA outperforms all other methods in terms of performance. Achieving leading performance on all of these datasets is a challenging task due to the wide range of image content and distortion types. Therefore, these observations confirm the effectiveness and superiority of CSIQA in accurately characterizing image quality.

4.4. Generalization Capability Validation

To assess the generalization capacity of CSIQA, we conduct cross-dataset validation experiments. In these experiments, the model trains on one dataset and subsequently tests on others, without any fine-tuning or adjustment of parameters. The outcomes of these experiments are displayed in Table 3, which shows the SRCC values achieved across five different datasets. Importantly, CSIQA surpasses state-of-the-art (SOTA) models in all four experiments involving cross-authentic datasets. This includes significant improvements in the LIVEC and KonIQ datasets. Additionally, CSIQA demonstrates substantial competitiveness on synthetic datasets like LIVE and CSIQ.

4.5. Qualitative Results.

Generalization Capability. To further assess our framework's generalization, we train models on the entire LIVE database and then test them using the gMAD competition (Ma et al., 2016b) on the Waterloo Database (Ma et al., 2016a). gMAD efficiently selects image pairs with maximum quality difference predicted by an attacking IQA model to challenge another defending model, which considers them to be of the same quality level. The selected pairs are shown to the observer to determine whether the attacker or the defender is robust. As shown in Fig. 3, in the first two columns, our model attacks the competing method DEIQT, where each column represents images chosen from the poorer and better quality levels predicted by the defender. In the last two columns, we fix our model as the defender, giving image pairs selected from poorer

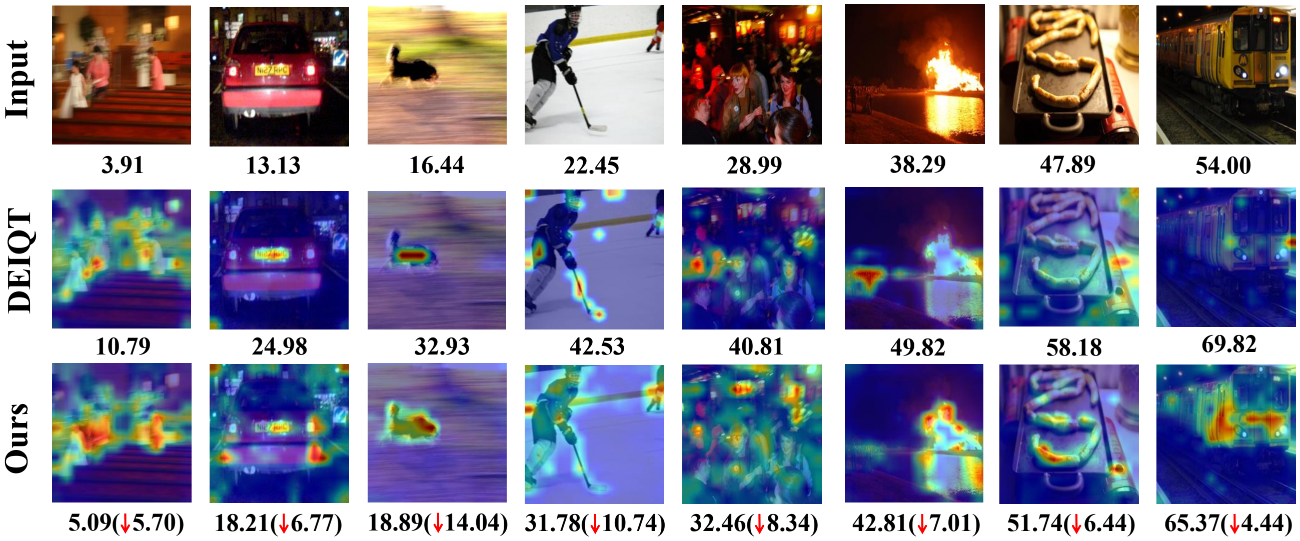


Figure 4. Activation maps of DEIQT (Qin et al., 2023) and CSIQA using Grad-CAM (Selvaraju et al., 2017) on authentic dataset LIVEC. The scores below the first row of images represent *Mean Opinion Scores* (MOS). Our model pays more attention to image distortion regions, and correspondingly our image quality prediction ability is closer to the true value. Rows 1-3 represent input images, CAMs from DEIQT (Qin et al., 2023), and CAMs from CSIQA, respectively.

and better quality levels, respectively. From Fig. 3, it is evident that when our model serves as the defender, the image pairs chosen by the attacker show little perceptual quality change, whereas, as the attacker, our model selects image pairs with more significant quality differences in succession. This indicates that the model has strong defensive and offensive capabilities. Moreover, it is worth mentioning that our model successfully identifies low-quality images with flat content in terms of image quality background (the images of the second column), although they deceive the defending model into considering them as high quality. These results further demonstrate that the proposed model has a strong generalization ability to challenge complex distortions in authentic images.

Framework Portability. To demonstrate the portability of our proposed quality contrastive learning framework (only contrastive learning and distillation are included), which can be applied to any IQA model, we perform an ablation study on previous methods TIQA and DEIQT with the backbone of VIT for fine-tuning, as shown in Table 2. We generate pseudo-labels using the original models (performance shown in rows one and three) and then apply our proposed contrastive framework. The reported results show a significant improvement of 1 to 2 points across multiple datasets, highlighting the universal effectiveness of the proposed method in enhancing existing IQA networks.

Visualization of attention map. Using GradCAM (Selvaraju et al., 2017), we visualize the feature attention map in Fig.4. CSIQA adeptly concentrates on local distortions (e.g., motion blur in column three, underexposure in the last, and overexposure in the third to last column). In contrast, DEIQT often misdirects its focus to clear backgrounds, overlooking distortions like exposure or motion blur. The figure compares CSIQA's and DEIQT's predictions on images with authentic distortions, highlighting CSIQA's superior performance, particularly in moderately distorted images. Additional visualizations are provided in the Appendix F.

4.6. Ablation Study

Effect of Logit Distillation. In Table 4, we compare indices ***a)*** with ***d)***, ***f)*** with ***i)***, and ***g)*** with ***k)*** to elucidate the varying enhancements brought about by logit distillation. By integrating regularization provided by pseudo-label during the training, we bolster the universal inference capabilities for quality estimation, thereby reinforcing generalization. This observation aligns with and validates Theorem A.1.

Effect of Quality-Aware Mask Attention. Next, we compare indices ***a)*** and ***b)***, showing that mask attention (denoted as $\boldsymbol{MA}$) improves quality representation extraction, yielding a performance boost of 0.6% to 1.5% across all datasets. It is worth noting that the random masking inevitably introduces a certain degree of noise to the attention mechanism, which may require further regularization to improve stability. Comparing indices ***c)*** and ***d)***, logits distillation (denoted as $\boldsymbol{LD}$) enhances the $\boldsymbol{MA}$ by providing more effective supervision information. This improvement is attributed to the effective regularization provided by $\boldsymbol{LD}$ and the reduction of error bounds resulting from $\boldsymbol{LD}$. These factors play a crucial role in enhancing the overall stability of the model.

Table 4. Ablation experiments on three datasets. Here, $\boldsymbol{EA}$ and $\boldsymbol{HA}$ refers to contrastive learning with easy and hard sampling; $\boldsymbol{MA}$ refers to mask attention; $\boldsymbol{LD}$ refers to logits distillation. The best performances are highlighted in boldface.

Index	$\boldsymbol{EA}$	$\boldsymbol{HA}$	$\boldsymbol{MA}$	$\boldsymbol{LD}$	TID2013		LIVEC		KonIQ	
					PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
a)					0.910	0.891	0.891	0.867	0.928	0.913
b)			✓		0.918	0.898	0.906	0.879	0.934	0.919
c)				✓	0.922	0.903	0.908	0.885	0.935	0.921
d)			✓	✓	$0.927_{+1.7\%}$	$0.907_{+1.6\%}$	$0.910_{+1.9\%}$	$0.882_{+1.5\%}$	$0.938_{+1.0\%}$	$0.924_{+1.1\%}$
e)	✓				0.927	0.909	0.908	0.882	0.935	0.919
f)	✓	✓			0.939	0.925	0.910	0.889	0.936	0.919
g)	✓	✓	✓		$0.940_{+3.0\%}$	$\mathbf{0.926}_{+3.5\%}$	$0.912_{+2.1\%}$	$0.889_{+2.2\%}$	$0.938_{+1.0\%}$	$0.923_{+1.0\%}$
h)	✓		✓	✓	0.928	0.910	0.918	0.897	0.942	0.927
i)	✓	✓		✓	0.940	0.925	0.914	0.892	0.941	0.926
k)	✓	✓	✓	✓	$\mathbf{0.941}_{+3.1\%}$	$\mathbf{0.926}_{+3.5\%}$	$\mathbf{0.920}_{+2.9\%}$	$\mathbf{0.898}_{+3.1\%}$	$\mathbf{0.943}_{+1.5\%}$	$\mathbf{0.929}_{+1.6\%}$

Effect of Contrast Learning with Easy-hard Sampling. We compare indices ***a)*** and ***e)*** to evaluate the impact of contrastive learning. Notably, employing contrastive learning with easy sampling (denoted as $\boldsymbol{EA}$) yields a significant enhancement, demonstrating an approximate 1.7% performance improvement relative to the baseline on the TID2013 dataset. When combined with hard sampling, denoted as $\boldsymbol{HA}$, we observe an additional performance increase of nearly 1% on the TID dataset, along with slight enhancements on other datasets. These variations in improvement across different datasets can be attributed to the differentiation of positive and negative samples, as well as the batch sizes selected for training. For example, the batch sizes of 16 and 128, used for training on the LIVEC and KonIQ datasets respectively, may result in either a scarcity or a redundancy of positive and negative samples.

Effect of sampling mode in Easy-hard Sampling. In this section, we further investigate the impact of different sampling methods on contrastive learning on the LIVEC dataset, as shown in Table 5. Initially, we set different sampling ratios at 10%/40% and 20%/60%. The results indicate that our contrastive learning approach is relatively insensitive to these hyperparameters. When combining easy and hard sampling, our results outperform those obtained from using either easy or hard sampling alone. This improvement can be attributed to our proposed fine-grained contrast scheme, which models the quality correlations of easily distinguishable and challenging sample pairs. This approach enables us to concentrate on images that pose greater challenges for quality prediction, thereby enhancing the robustness of our method. Additionally, we explored two adaptive selection strategies: 1) Using Average Quality as a Threshold: This strategy employs the average quality within a batch as a threshold to adaptively adjust the ratio of positive and negative samples in different batches. 2) Adaptive Sampling Technique: Inspired by adaptive sampling concepts (Liu et al., 2019), this experiment dynamically adjusts the sampling probability based on the relative difficulty of ranking. More details can be found in the Appendix D. These strategies aim to refine the selection process, further contributing to the effectiveness of our contrastive learning, which inspires us to explore smarter sampling schemes in the future.

Table 5. Ablation experiments of sampling mode in Easy-Hard Sampling. The best performances are highlighted in boldface.

Sampling	Pos. / Neg.	PLCC	SRCC
Easy	10% / 40%	0.908	0.882
Easy	20% / 60%	0.905	0.883
Hard	10% / 40%	0.903	0.884
Hard	20% / 60%	0.904	0.886
Easy-Hard	10% / 40%	0.910	0.889
Easy-Hard	20% / 60%	0.907	0.887
Easy-Hard	40% / 60%	0.908	0.890
Easy-Hard	Adaptive Sampling	**0.911**	**0.892**
Easy-Hard	Adaptive Average	0.909	0.891

5. Conclusion

In this study, we present CSIQA, an innovative quality contrastive learning methodology for BIQA. Contrary to conventional models that predominantly concentrate on local intra-image analysis, CSIQA accentuates both the global context quality contrast and the nuances of local distortions. It incorporates a quality context contrastive module, facilitated by pseudo-labels, to underscore samples with unpredictable quality, thereby capturing global quality contrastive relations. Additionally, a quality-aware mask attention module is employed to refine quality discrimination by focusing on local distortions. Experiments show that CSIQA surpasses current SOTA methods. Moreover, our training approach seamlessly integrates with existing methods, yielding notable enhancements.

Acknowledgements

This work was supported by National Science and Technology Major Project (No. 2022ZD0118202), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

Impact Statement

This paper presents work whose goal is to advance the field of Image Quality Assessment. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Banham, M. R. and Katsaggelos, A. K. Digital image restoration. *IEEE signal processing magazine*, 14(2): 24–41, 1997. 1

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017. 6

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 2

Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 1

Dong, C., Loy, C. C., He, K., and Tang, X. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. 3

Fang, Y., Zhu, H., Zeng, Y., Ma, K., and Wang, Z. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3677–3686, 2020. 6, 16

Gao, F., Tao, D., Gao, X., and Li, X. Learning to rank for blind image quality assessment. *IEEE transactions on neural networks and learning systems*, 26(10):2275–2290, 2015. 5

Ghadiyaram, D. and Bovik, A. C. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 6, 16, 21

Golestaneh, S. A., Dadsetan, S., and Kitani, K. M. 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*, pp. 1220–1230, 2022. 3, 6

Gu, S., Bao, J., Chen, D., and Wen, F. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 369–385. Springer, 2020. 1

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 18

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022. 3

Hosu, V., Lin, H., Sziranyi, T., and Saupe, D. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 6, 16, 21

Hu, R., Yang, R., Liu, Y., and Li, X. Simulation and mitigation of the wrap-around artifact in the mri image. *Frontiers in Computational Neuroscience*, 15:746549, 2021. 1

Hu, R., Monebhurrun, V., Himeno, R., Yokota, H., and Costen, F. An uncertainty analysis on finite difference time-domain computations with artificial neural networks: improving accuracy while maintaining low computational costs. *IEEE Antennas and Propagation Magazine*, 65(1): 60–70, 2022. 1

Kalantidis, Y., Sariyildiz, M. B., Pion, N., Weinzaepfel, P., and Larlus, D. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020. 5

Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021. 2, 3, 6, 7

Kim, J. and Lee, S. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11 (1):206–220, 2016. 6

Larson, E. C. and Chandler, D. M. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010. 3, 6, 16

Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., Zhang, T., and Chen, Q. Involution: Inverting the inherence of convolution for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12321–12330, 2021. 21

Lin, H., Hosu, V., and Saupe, D. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019. 6, 16

Liu, H., Zhu, X., Lei, Z., and Li, S. Z. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11947–11956, 2019. 9

Liu, X., Van De Weijer, J., and Bagdanov, A. D. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pp. 1040–1049, 2017. 2, 3

Liu, Y., Zhang, B., Hu, R., Gu, K., Zhai, G., and Dong, J. Underwater image quality assessment: Benchmark database and objective method. *IEEE Transactions on Multimedia*, 2024. 1

Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., and Zhang, L. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016a. 7

Ma, K., Wu, Q., Wang, Z., Duanmu, Z., Yong, H., Li, H., and Zhang, L. Group mad competition-a new methodology to compare objective image quality models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1664–1673, 2016b. 7

Ma, K., Liu, W., Liu, T., Wang, Z., and Tao, D. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017a. 3

Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., and Zuo, W. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017b. 3, 6

Ma, K., Liu, X., Fang, Y., and Simoncelli, E. P. Blind image quality assessment by learning from multiple annotators. In *2019 IEEE international conference on image processing (ICIP)*, pp. 2344–2348. IEEE, 2019. 3

Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 1, 2, 3, 6, 19

Mittal, A., Moorthy, A. K., and Bovik, A. C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 6, 7

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021. 17

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 15

Pan, Z., Zhang, H., Lei, J., Fang, Y., Shao, X., Ling, N., and Kwong, S. Dacnn: Blind image quality assessment via a distortion-aware convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7518–7531, 2022. 6, 21

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015. 6, 16, 21

Qin, G., Hu, R., Liu, Y., Zheng, X., Liu, H., Li, X., and Zhang, Y. Data-efficient image quality assessment with attention-panel decoder. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023. 2, 3, 4, 6, 7, 8, 16, 18, 21, 22

Saha, A., Mishra, S., and Bovik, A. C. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5846–5855, 2023. 2, 3, 6

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015. 5

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017. 8, 22

Sheikh, H. R. and Bovik, A. C. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006. 2, 4

Sheikh, H. R., Sabir, M. F., and Bovik, A. C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 1, 6, 16, 21

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3667–3676, 2020. 2, 3, 6, 17, 18

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019. 21

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021. 17

Touvron, H., Cord, M., and Jégou, H. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 6, 16

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 17, 21

Wang, S., Gu, K., Zhang, X., Lin, W., Ma, S., and Gao, W. Reduced-reference quality assessment of screen content images. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):1–14, 2016. 1

Wang, T., Yuan, L., Zhang, X., and Feng, J. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4933–4942, 2019. 4

Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., and Van Gool, L. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7303–7313, 2021. 5

Wu, J., Ma, J., Liang, F., Dong, W., Shi, G., and Lin, W. End-to-end blind image quality prediction with cascaded deep neural network. *IEEE Transactions on image processing*, 29:7414–7426, 2020. 1

Yin, G., Wang, W., Yuan, Z., Han, C., Ji, W., Sun, S., and Wang, C. Content-variant reference image quality assessment via knowledge distillation. *arXiv preprint arXiv:2202.13123*, 2022. 2

Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., and Bovik, A. 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*, pp. 3575–3585, 2020. 6, 16

You, J. and Korhonen, J. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1389–1393. IEEE, 2021. 3, 6

Zeng, H., Zhang, L., and Bovik, A. C. Blind image quality assessment with a probabilistic quality representation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 609–613, 2018. doi: 10.1109/ICIP.2018.8451285. 2

Zhang, L., Zhang, L., and Bovik, A. C. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 6, 7

Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 3, 6, 17, 18

Zhang, W., Ma, K., Zhai, G., and Yang, X. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30: 3474–3486, 2021. 2, 3

Zhang, W., Zhai, G., Wei, Y., Yang, X., and Ma, K. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023. 2, 3

Zhao, K., Yuan, K., Sun, M., Li, M., and Wen, X. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22302–22313, 2023. 2, 3

Zhu, H., Li, L., Wu, J., Dong, W., and Shi, G. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14143–14152, 2020. 6

Overview

This Appendix contains six sections. § A.1 offers a theoretical proof underscoring the reliability of incorporating pseudo-labels. § A.2 provides a theoretical demonstration that contrastive learning exhibits greater robustness to noise compared to rank learning. § B presents some details of the CSIQA model architecture § C delves deeper into the training and evaluation specifics. § D extends our discussion with additional ablation studies related to our CSIQA. This includes an exploration of the quality-aware token and an examination of the efficacy of logit distillation in hastening convergence. § E conducts a sensitivity analysis focusing on hyperparameters, specifically addressing mask probability and the weight of the loss function. § F presents a thorough qualitative assessment of CSIQA. This encompasses visual representations of feature distributions across a range of quality scores and a visualization of the activation map.

A. PROOFS

A.1. Qualitative Proof of Theorem 1

The soundness of pseudo-labels is meticulously ensured by both mathematical proofs and experiments. We start with the following three assumptions. 1) The teacher network generates enough pseudo-labels. 2) The pseudo-labels generated by the teacher network are relatively balanced with the samples in the dataset. 3) The teacher randomly samples in the process of generating pseudo-labels.

1. Definition: $T(x)$ represents the teacher's distribution. $R(x)$ represents the true distribution. $S(x)$ represents the student's distribution. $f(x)$ is the probability distribution of x.

2. Minimization of Expression: Let $a, b \in [0,1]$, and $a + b = 1$. The goal is to find the student's distribution $\hat{S}(x)$ that minimizes the expression involving the differences between the student's distribution and the teacher's distribution, as well as the true distribution:

$$\hat{S}(x) = \underset{S}{\operatorname{argmin}} \sum_i a \left|S\left(x_i\right) - T\left(x_i\right)\right| + b \left|S\left(x_i\right) - R\left(x_i\right)\right|$$

3. Inequality Derivation: The optimal solution, denoted as $\hat{S}(x)$, was pursued within the context of this analysis. The derivation is as follows:

$$\int \left(a \left|\hat{S}(x) - T(x)\right| + b \left|\hat{S}(x) - R(x)\right|\right) f(x)dx \leq \int (a|S-T| + b|S-R|) f(x)dx$$

$$\text{(Let: } S(x) = bT + aR\text{)}$$

$$\begin{aligned}
\int \left(a \left|\hat{S}(x) - T(x)\right| + b \left|\hat{S}(x) - R(x)\right|\right) f(x)dx &\leq \int (a|bT + aR - T| + b|bT + aR - R|) f(x)dx \\
&= \int (a^2|R-T| + b^2|T-R|) f(x)dx \\
&= \left(a^2 + b^2\right) \int |R-T| f(x)dx \\
&< (a+b) \int |R-T| f(x)dx
\end{aligned}$$

4. Conclusion: Through deduction, the $\hat{S}(x)$ will be influenced by the teacher and will be superior to the teacher. We conclude that with a sufficiently abundant quantity of pseudo-labels (ensuring the expectation), and a relatively balanced distribution between pseudo-labels and dataset samples (for any values of 'a' and 'b'), our anticipated loss is lower compared to training solely with ground truth. Consequently, under these three assumptions, the reliability of pseudo-labels can be ensured.

A.2. Qualitative Proof of Theorem 2

1. The formalism of contrastive learning (CL) and rank learning (RL): We present the general formalism of CL and RL at the label level: The loss function of CL simplifies to "distance of positive pair-distance of negative pair" (here we disregard the "partition function"). For the purpose of demonstration, we select one sample each from the maximum and minimum values, using them as negative and positive examples respectively. The proof for multiple examples follows the same logic.:

$$\mathcal{L}_{CL} = \min |y_+ - y| - \max |y_- - y|$$

For a label y sample, y_+ and y_- are its positive/negative samples.

The general form of loss in RL is typically the “negative of the distance":

$$\mathcal{L}_{RL} = -|y_1 - y_0|$$

Here we assume without loss of generality that the quality score of y_1 is larger than y_0.

2. Rank Learning (RL) Analysis: Consider a label y_0 for a sample point. We sample around y_0 using a distribution to obtain y_1, where $y_1 = y_0 + \xi$ and ξ follows the sampling distribution. Given noise η in the "Teacher" model, we define the noisy label as $z = y_1 + \eta$ when incorporating noise η. The RL with noise η can be expressed as:

$$\int -|z - y_0| f(\eta) g(\xi) d\eta d\xi = \int -|\eta + \xi| f(\eta) g(\xi) d\eta d\xi$$

In the absence of noise η in RL:

$$\int -|y_1 - y_0| g(\xi) d\xi = \int -|\xi| f(\eta) g(\xi) d\eta d\xi$$

3. Contrastive Learning (CL) Analysis: With $n + 2$ samples $\{\hat{y}_1, ..., \hat{y}_{n+2}\}$, the probability of $|\hat{y}_i - y_0| \geq |\hat{y}_j - y_0|$ is p. Thus, the probability that all samples are farther from y_0 than the positive sample $\hat{y}_j$ is p^{n+1}. The probability density is $(n + 1)p^n h(\eta, \xi)$, where $h(\eta, \xi)$ represents the probability density of p concerning η and ξ. The CL with noise η can be expressed as:

$$\begin{aligned} &\int (n+1) p^n h |\eta + \xi| f g d\eta d\xi - \int (n+1)(1-p)^n (-h) |\eta + \xi| f g d\eta d\xi \\ &= \int (n+1)[p^n + (1-p)^n] |\eta + \xi| f g h d\eta d\xi \end{aligned}$$

In the absence of noise η in CL (due to the absence of noise η, $\hat{p}$ and $\hat{h}$ are used instead to replace p and h from before).

$$\begin{aligned} &\int (n+1) \hat{p}^n \hat{h} |\xi| g d\xi - \int (n+1)(1-\hat{p})^n (-\hat{h}) |\xi| g d\xi \\ &= \int (n+1)[\hat{p}^n + (1-\hat{p})^n] \hat{h} |\xi| f g d\eta d\xi \end{aligned}$$

4. Comparative Analysis of Robustness: We qualitatively prove the robustness of CL and RL:

Evaluating the impact of noise on RL:

$$\left| \int -|\eta + \xi| f g d\eta d\xi - \int -|\xi| f g d\eta d\xi \right| = \int ||\eta + \xi| - |\xi|| f g d\eta d\xi \tag{9}$$

Evaluating the impact of noise on CL:

$$\begin{aligned} &\left| (n+1)\{ \int [p^n + (1-p)^n] \cdot h |\eta + \xi| f g d\eta d\xi - \int [\hat{p}^n + (1-\hat{p})^n] \cdot \hat{h} |\xi| f g d\eta d\xi \} \right| \\ &= \int |[p^n + (1 - p^n)] \cdot |\eta + \xi| \cdot h - [\hat{p}^n + (1-\hat{p})^n] \cdot |\xi| \cdot \hat{h}| \cdot f \cdot g \cdot (n+1) d\eta d\xi \end{aligned} \tag{10}$$

Let $\tilde{p} = \max\{p^n + (1-p)^n, \hat{p}^n + (1-\hat{p})^n\}$ and $\tilde{h} = \underset{\{h,\hat{h}\}}{\text{argmax}}\{h, \hat{h}\}$, *i.e.,*

$$\text{Equ.10} \leq \int ||\eta + \xi| - |\xi|| \cdot [\tilde{p}^n + (1-\tilde{p})^n] \cdot \tilde{h} \cdot f \cdot g \cdot (n+1) d\eta d\xi \leq \text{Equ.9} \tag{11}$$

5. Conclusion: From the aforementioned equation, it can be deduced that when n is sufficiently large and $0 < \tilde{p} < 1$, contrastive learning exhibits enhanced robustness to noise η due to its inherent filtering mechanism. In our empirical validation, both contrastive learning and rank learning techniques were applied across various backbones, with training conducted on noisy labels. The experimental findings underscore that contrastive learning is notably more robust. Further details can be found in Sec. D.

B. Architecture

B.1. Encoder Details

The quality-aware token in the encoder functions in a manner akin to the CLS token, leveraging a self-attention mechanism. Distinctively, it attends solely to visible patches. In the encoder's formulation (Eq. 2 as depicted in the manuscript), a mask operation is designed to regulate the interaction scope of the quality token during forward propagation. The feature values from the Multi-Head Self-Attention (MHSA) block are subsequently refined through a Multi-Layer Perceptron (MLP) and a residual connection. The mathematical representation is as follows:

$$\mathbf{F}^l = \begin{cases} \sigma\left(\log \boldsymbol{M} + \mathbf{Q}^l \left(\mathbf{K}^l\right)^\top\right) \mathbf{V}^l + \mathbf{F}^{l-1}, & 0 < l \leq L \\ \mathbf{F}, & l = 0 \end{cases}$$

$$\boldsymbol{F}^l = \text{MLP}\left(\text{Norm}\left(\boldsymbol{F}^l\right)\right) + \boldsymbol{F}^l$$

$\mathbf{F} = \{\boldsymbol{F}_{qua}; [\boldsymbol{F}_{cls}; \boldsymbol{F}_{1:N-1}]\} \in \mathbb{R}^{(N+1)\times D}$ be the embedding sequence, where N is the number of patches plus one CLS token. Here, F_{qua} and F_{cls} symbolize the QUA token and CLS token. After extracting features through the L layers of the encoder, the final output $\mathbf{F}^L$ is obtained.

B.2. Decoder Details

Within the decoder, both the QUA token and CLS token undergo refinement. The QUA and CLS token is processed inside the MHSA block to discern dependencies among its constituents. The output from the MHSA is then integrated with residual connections, resulting in a query for the transformer decoder. This process is mathematically captured as:

$$\boldsymbol{Q}_{\text{qua}} = \text{MHSA}\left(\text{Norm}\left(\boldsymbol{F}_{\text{qua}}\right)\right) + \boldsymbol{F}_{\text{qua}}$$

$$\boldsymbol{Q}_{\text{cls}} = \text{MHSA}\left(\text{Norm}\left(\boldsymbol{F}_{\text{cls}}\right)\right) + \boldsymbol{F}_{\text{cls}}$$

Subsequently, the patch embeddings, excluding the QUA and CLS tokens, are used as keys and values for the decoder, represented by $\boldsymbol{K}_d$ and $\boldsymbol{V}_d$, respectively. The queries $\boldsymbol{Q}_{\text{qua}}$ and $\boldsymbol{Q}_{\text{cls}}$ are linear transformations of $\mathbf{F}_{\text{qua}}$ and $\mathbf{F}_{\text{cls}}$. The interaction through Multi-Head Cross-Attention (MHCA) helps derive the refined features $\mathbf{F}_{\text{qua}}$ and $\mathbf{F}_{\text{cls}}$. Finally, we employ a Multi-Layer Perceptron (MLP) to regress the final quality score based on the concatenated features of $\mathbf{F}_{\text{qua}}$ and $\mathbf{F}_{\text{cls}}$.

$$F_{\text{qua}} = \text{MLP}\left(\text{MHCA}\left(\text{Norm}\left(\boldsymbol{Q}_{qua}\right), \boldsymbol{K}_d, \boldsymbol{V}_d\right) + \boldsymbol{Q}_{qua}\right)$$

$$F_{\text{cls}} = \text{MLP}\left(\text{MHCA}\left(\text{Norm}\left(\boldsymbol{Q}_{cls}\right), \boldsymbol{K}_d, \boldsymbol{V}_d\right) + \boldsymbol{Q}_{cls}\right)$$

$$Y = \text{MLP}\left(\text{Cat}\left(\boldsymbol{F}_{\text{qua}}, \boldsymbol{F}_{\text{cls}}\right)\right)$$

C. Training and Evaluation Details

Student Network. 1) Training. To train the student network, we adopt a standard approach of randomly cropping input images into ten patches, each with a 224 × 224 resolution. Subsequently, we reshape these patches into a sequence of smaller patches with a patch size of p = 16 and an input token dimension of D = 384. Furthermore, we present additional training preprocessing details for various datasets, which are not included in the main paper, as shown in Table 6. For different benchmarks, we employ different training settings for a fair comparison. In the training phase, we perform an easy-hard sampling of images within each batch to obtain positive and negative samples for contrastive learning. The contrastive loss is then calculated using InfoNCE (Oord et al., 2018).

Table 6. Training preprocessing details of selected BIQA datasets.

Dataset	Resolution	Resize	Batch Size	Label Range
LIVE (Sheikh et al., 2006)	768×512	512×384	12	DMOS [0,100]
CSIQ (Larson & Chandler, 2010)	512×512	512×512	12	DMOS [0,1]
TID2013 (Ponomarenko et al., 2015)	512×384	512×384	48	MOS [0,9]
KADID (Lin et al., 2019)	512×384	512×384	128	MOS [1,5]
LIVEC (Ghadiyaram & Bovik, 2015)	$500P \sim 640P$	$500P \sim 640P$	16	MOS [1,100]
KonIQ (Hosu et al., 2020)	$768P$	512×384	128	MOS [0,5]
LIVEFB (Ying et al., 2020)	$160P \sim 700P$	512×512	128	MOS [0,100]
SPAQ (Fang et al., 2020)	$1080P \sim 4368P$	512×384	128	MOS [0,100]

Table 7. Teacher performance on authentic and synthetic datasets.

	LIVE		CSIQ		TID2013		KADID	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Teacher	0.978	0.978	0.960	0.945	0.910	0.891	0.904	0.902
	LIVEC		KonIQ		LIVEFB		SPAQ	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Teacher	0.900	0.878	0.930	0.914	0.667	0.567	0.922	0.920

Table 8. Experiments on the reliability ablation of teacher-generated pseudo-labels.

	LIVE		LIVEC	
Method	PLCC	SRCC	PLCC	SRCC
baseline baseline	0.970	0.967	0.891	0.867
CSIQA w/ pesudo	0.983(+1.3%)	0.982(+1.5%)	0.916(+2.5%)	0.898(+3.1%)
CSIQA w/ pesudo+label	0.985(+1.5%)	0.983(+2.1%)	0.920(+2.9%)	0.898(+3.1%)

2) Testing. We do not adjust the hyperparameters during training and testing with different train-test dataset splits. We repeat this process 10 times to mitigate performance bias, and we report the average SRCC and PLCC values.

Teacher Network. We will start with structure, training, performance, and selection principles to introduce the teacher:

• **Structure.** The pre-trained teacher transformer consists of VIT-S (Touvron et al., 2022) and one decoder layer (Qin et al., 2023). It is worth noting that, as a general framework (Table 2 in our manuscript), the teacher network is all the pre-trained version of the student network without mask attention.

• **Teacher Training.** The training process of the teacher network is consistent with the training process of the student network, we use a learning rate of 2×10^{-4} and a decay factor of 10 every 3 epochs to train the model for 9 epochs. During training and inference, we adopt a standard approach of randomly cropping input images into ten patches, each with a 224 × 224 resolution. Subsequently, we reshape these patches into a sequence of smaller patches with a patch size of p = 16 and an input token dimension of D = 384.

• **Performance.** We guarantee a high accuracy of the pre-trained teacher model on multiple datasets to provide reliable pseudo-labels, as shown in Table 7. Our study further illustrates that our teacher model produces relatively reliable pseudo-labels: We compare the student model's performance using pseudo-labels with its performance using the true labels. Table 8 indicates that the student model achieves similar performance with both, thus the pseudo-labels are reliable.

Table 9. Experiments on the selection principle of teacher models.

	DBCNN		HyperNet		VIT-D		VIT-DF	
Teacher	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
baseline	0.891	0.867	0.891	0.867	0.891	0.867	0.891	0.867
CSIQA	**0.905**	**0.880**	**0.903**	**0.878**	**0.910**	**0.882**	**0.920**	**0.898**

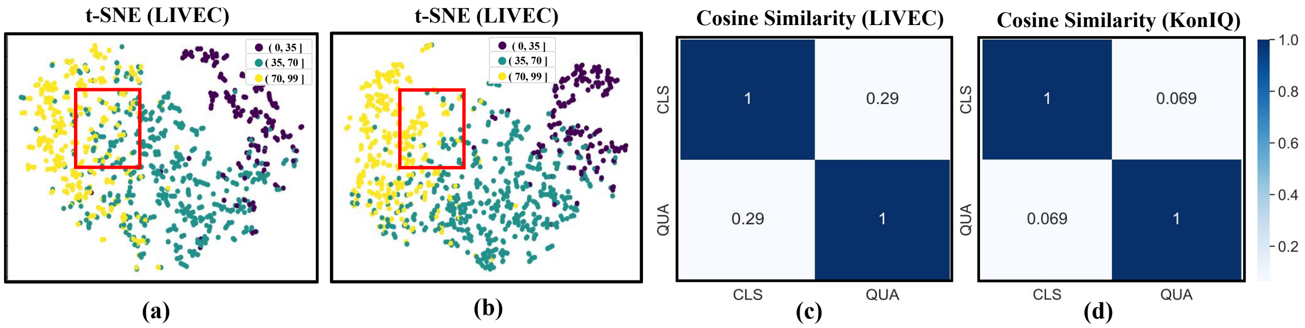


Figure 5. (a) and (b) depict the embedding space without and with the inclusion of a quality token. The comparison indicates that introducing the quality token in contrastive learning optimizes more quality-related features by incorporating semantic information. This results in a well-structured quality-aware embedding space (b). Furthermore, (c) and (d) demonstrate that the minimal similarity between the quality token and cls token highlights their distinct focuses on different regions of distortion. Specifically, the cls token prioritizes global distortion, while the quality token prioritizes local distortion. As a result, the quality perception becomes more comprehensive.

• **Selection Principle.** In our experiments on the LIVEC dataset, we explore two distinct settings to understand the influence of teacher models on the student's performance, as shown in Table 9: 1) Teacher models with identical architecture but varying performances. Specifically, we consider models VIT-D and VIT-DF. Here, VIT-DF represents the fully converged VIT-D, and the performance of VIT-DF is shown in Table 7. 2) Teacher models with differing architectures but consistent performance. For this, we evaluate models DBCNN (Zhang et al., 2018), HyperNet (Su et al., 2020), and VIT-D. The PLCC of these three teacher networks are 0.863, 0.867, and 0.868, respectively. From the results, it is evident that sharing the same architecture between the teacher and student models significantly enhances the student's performance. This is because our contrastive learning method for sampling difficult examples is closely tied to the teacher's predictive capabilities. When the teacher and student share the same architecture, it highlights the potential shortcomings of the student model trained solely under label supervision. Additionally, improvements in the teacher model's performance also aid the student. This implies that the best choice for the teacher is one that has both training convergence and shares the same architecture as the student, thereby maximizing the effectiveness of our framework.

D. More Ablation Studies

Ablation about quality-aware token. To illustrate the effectiveness of introducing the QUA token in generating a well-structured embedding space for contrastive learning, we categorize the quality scores of the LIVEC dataset into three distinct categories: heavily distorted images (score 0-35), moderately distorted images (score 35-70), and lightly distorted images (score 70-100). Each set includes a minimum of 300 images from the LIVEC real-world dataset. We then generate the quality-aware feature representations using t-SNE (Van der Maaten & Hinton, 2008), which are depicted in Fig. 5 (a) and (b). Fig. 5(b) presents the embedding space obtained when quality tokens are utilized and compared with the quality space without quality tokens (Fig. 5(a)). The overlapping part of the red area represents those features that are difficult for the model to distinguish. Our analysis reveals that our CSIQA is more discriminative for the quality features after introducing the quality token. To further demonstrate the unique features modeled by different tokens (for cls token and quality token), we compute the cosine similarity between the class token and quality token (averaged over the LIVEC and KonIQ datasets) in Fig. 5(c) and (d), resulting in low values of 0.29 and 0.069, which is significantly lower than the similarity between class and distillation labels in previous work (Touvron et al., 2021); 0.96 and 0.94 in DeiT-T and Deit-S, respectively (Naseer

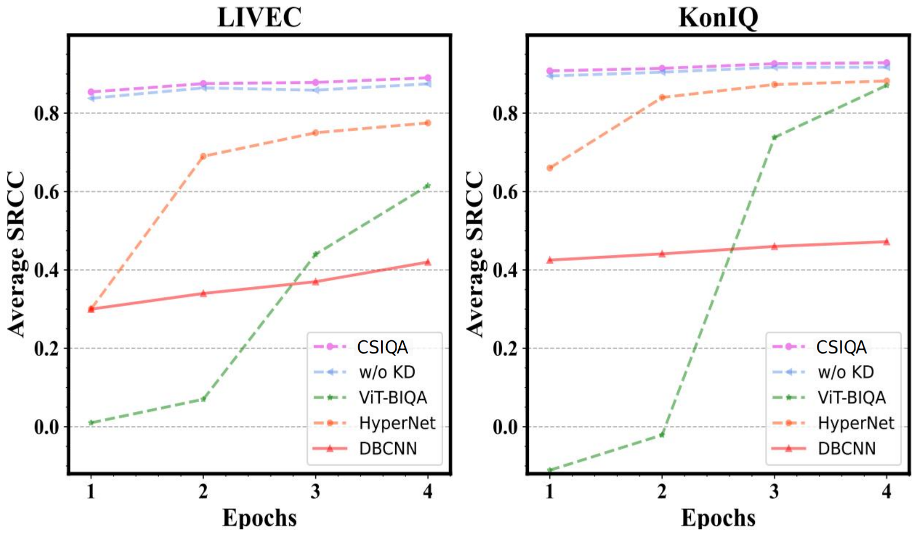


Figure 6. Average SRCC versus Epochs on different datasets ablation on logit distillation.

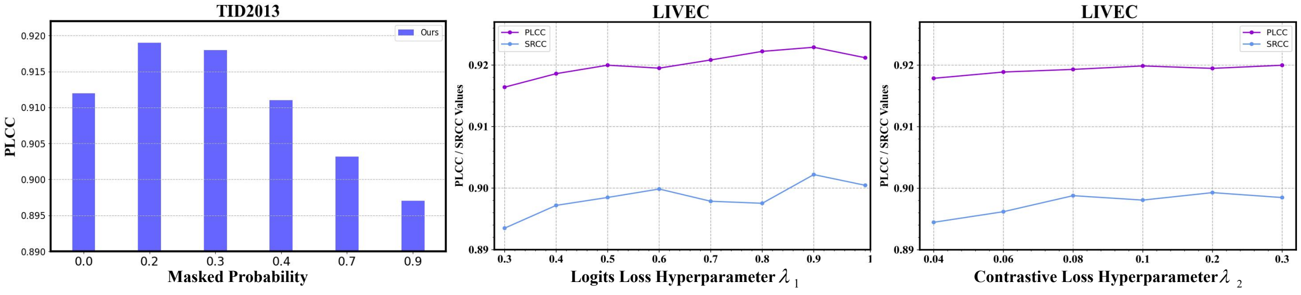


Figure 7. (a) Sensitivity analysis of patch mask probabilities (b) Hyperparameter λ_1 for balanced logit distillation loss (c) Hyperparameter λ_2 for balanced contrastive loss. All these results were obtained on the TID2013 and LIVEC datasets.

et al., 2021). This finding supports our hypothesis that the use of individual tokens for partial patch features in ViT can enable judges of the image quality from a different perspective (i.e., global and local).

Ablation about logit distillation. Fig. 6 presents the average SRCC with respect to the number of epochs on the LIVEC and KonIQ test sets. The results indicate that our CSIQA method exhibits remarkably faster convergence compared to other methods (Qin et al., 2023; Su et al., 2020; Zhang et al., 2018). Additionally, incorporating the logit distillation module in the training process results in an accelerated convergence speed and a further improvement in performance on both the LIVEC and KonIQ datasets (Compare CSIQA and w/o KD in Fig. 6). Notably, it achieves the fastest convergence after only a single epoch of training, outperforming the second-best NR-IQA method mentioned in Table 1 of the main paper.

Ablation about decoder. Inspired by (Qin et al., 2023), we introduce a Transformer decoder for further improved feature refinement. It uses masked self-attention and cross-attention on “CLS" and “QUA" tokens. Masked self-attention minimizes irrelevant features, while cross-attention enhances quality-related aspects, adapting tokens for BIQA. The ablation experiments presented in Table 10 further validate the efficacy of the decoder in refining features.

Ablation about the hard-easy sampling strategy with different backbone. Table11 outlines the fine-tuning outcomes using the ResNet-50(He et al., 2016) backbone, employing both hard and hard-easy samples. When fine-tuning with simple examples, there’s a slight performance drop, possibly due to model overfitting caused by an abundance of simple examples, and fine-tuning with easy-hard samples yields a 1.3% enhancement on the LIVEC dataset. Easy-Hard Sampling helps us circumvent the issue of excessive hard samples leading to local minima, while also addressing overfitting stemming from straightforward examples.

Table 10. Ablation experiments concerning the decoder.

	LIVE		LIVEC	
Method	PLCC	SRCC	PLCC	SRCC
VIT	0.965	0.949	0.861	0.845
CSIQA w/o decoder	0.981	0.980	0.901	0.878
CSIQA	0.985(+0.4%)	0.983(+0.3%)	0.920(+1.9%)	0.898(+2.0%)

Table 11. Ablation study about hard-easy sampling strategy for ResNet-50

		LIVE		LIVEC	
Sampling	Method	PLCC	SRCC	PLCC	SRCC
	ResNet-50	0.942	0.938	0.858	0.851
Easy	ResNet-50	0.949	0.945	0.857	0.848
Hard	ResNet-50	0.952(+1.0%)	0.947(+0.9%)	0.864(+0.6%)	0.858(+0.7%)
Easy-Hard	ResNet-50	0.956(+1.4%)	0.950(+1.2%)	0.871(+1.3%)	0.858(+0.7%)

Ablation about sampling mode in hard-easy sampling. We explore two adaptive selection strategies, as shown in Table 5:

1. **Mean Quality as Threshold:** This strategy utilizes the mean quality within a batch as a threshold to adaptively adjust the ratio of positive and negative samples across different batches.
2. **Adaptive Sampling Technique:** Inspired by the concept of adaptive sampling, this technique dynamically adjusts the sampling probabilities according to the difficulty of relative ranking. Specifically:
   - For a given sample, it is initially ranked according to the distance of its quality label from those of all other samples in the same batch, from the nearest to the furthest. An initial sampling probability p_i is then assigned to each ranked position i.
   - During the training phase, we assign a selection probability p_i to each rank of samples. If the relative rank of a sample within a batch is correctly identified, it is deemed an "easy rank" sample, leading to a reduction in the sampling probability for this rank, recalculated as $p_i = (1 - \delta) \times p_i$. Conversely, a misprediction signifies a "difficult rank" sample, prompting an increase in the sampling probability to $p_i = (1 + \delta) \times p_i$.

This dynamic, adaptive methodology shifts our focus progressively towards ranks that present more significant challenges in accurate ranking, particularly those at intermediate positions within the sorting order. Simultaneously, it reduces our emphasis on ranks with more apparent relative positions, such as those at the extremities of the sorting spectrum. These dynamic adjustments help to improve the accuracy of quality predictions.

More comparisons with CONTRIQUE. To make a fair comparison with CONTRIQUE (Madhusudana et al., 2022), we take the same ResNet-50 as our backbone. For further alignment, we try two training methods. One is to fine-tune only the last encoder layer for contrastive learning and the fully connected layer (aligned with CONTRIQUE), and the other is to fine-tune the entire network. The experimental results are shown in Table 12. In the case of the frozen encoder, the method based on context contrastive learning shows a 0.7% (1.1% with fine-tuning) improvement over CONTRIQUE on the synthetic dataset LIVE. It is worth noting that on the real dataset LIVEC, there is a significant increase of 2.6% (2.5% with fine-tuning). CONTRIQUE treats each distorted image as a unique class and discusses the positive class contrast information that other real images of similar quality can provide. This approach indirectly reduces the model's ability to perceive differences in real distorted images. However, our method compensates for this shortcoming by using accurate and reliable labels, enabling it to explicitly mine the subtle differences between images with different quality scores.

More experimental results on the framework portability. We assess existing IQA methods' generality, denoted by "+," signifying our model with our learning strategy. Pseudo-labels are from original pre-trained IQA models. Reported

Table 12. Comparison of CSIQA with CONTRIQUE using ResNet-50 as the backbone, Co. denotes CONTRIQUE.

		LIVE		LIVEC	
Method	Backbone	PLCC	SRCC	PLCC	SRCC
Baseline	ResNet-50	0.947	0.923	0.852	0.827
Co.(Linear Regression)	ResNet-50	0.961	0.960	0.857	0.845
CSIQA(Linear Regression)	ResNet-50	0.968(+0.7%)	0.966(+0.6%)	0.883(+2.6%)	0.855(+1.0%)
CSIQA(fine-tuning)	ResNet-50	0.972(+1.1%)	0.970(+1.0%)	0.882(+2.5%)	0.859(+1.4%)

Table 13. We assess existing IQA methods' generality, denoted by "+" signifying our model with our learning strategy. Pseudo-labels are from original pre-trained IQA models. Reported results show our method effectively enhances IQA networks.

	LIVE		LIVEC	
Method	PLCC	SRCC	PLCC	SRCC
DBCNN	0.957	0.955	0.859	0.833
DBCNN +	**0.970**$_{+1.3\%}$	0.968$_{+1.3\%}$	**0.878**$_{+1.9\%}$	0.859$_{+2.6\%}$
HyperNet	0.966	0.962	0.880	0.867
HyperNet +	**0.974**$_{+0.8\%}$	**0.971**$_{+0.9\%}$	**0.903**$_{+2.3\%}$	**0.884**$_{+1.7\%}$
TReS	0.968	0.969	0.866	0.848
TReS +	-	-	**0.887**$_{+2.2\%}$	**0.865**$_{+1.7\%}$

results show our method effectively enhances IQA networks. As evidenced by Table 13 and Table 14, our proposed model seamlessly integrates with existing IQA methodologies, resulting in notable performance enhancements.

Experimental support for Theorem A.2. As shown in Table 15 and Table 16, to illustrate contrastive learning's noise-resilience, we conducted experiments using ResNet-50 and VGG-16 backbones in both contrastive and ranking settings across diverse datasets. The choice of comparison samples for both methods was influenced by noisy labels. Results summarized in the table highlight contrastive learning's competitive advantage. With the VGG-16 backbone, we improved by 8.1% on TID2013 and remained competitive on LIVE. Using RankNet-50, contrastive learning boosted performance by 5.3% on Kadid and 2.7% on KonIQ. This is attributed to contrastive learning's robustness in handling extreme sample noise. While ranking learning can be vulnerable to noise, especially when samples are closely ranked, contrastive learning, as demonstrated theoretically, sidesteps this issue. The ample size of 'n' in the theoretical proof allows the largest negative sample to converge toward its expectation and the smaller positive sample to approximate its positive counterpart. Consequently, contrastive learning exhibits heightened noise robustness.

E. Sensitivity study of Hyperparameters

Patch Masked Probability. The probability of applying the random mask is a hyperparameter that adjusts the likelihood of incorporating the random quality-aware mask attention citation. To assess its impact on CSIQA, we conduct ablation experiments with varying random mask probabilities. As illustrated in Fig. 7(a), we find that a lower mask probability decreases the effectiveness of the mask, whereas a higher mask probability also reduces its effectiveness due to the limited information from each patch. Balancing these effects, we set the mask probability to α = 0.3 for our experiments.

Loss Weight Hyperparameter. In this paper, we utilize λ_1 and λ_2 to balance logit supervision and contrastive learning. In this subsection, we conduct a sensitivity study of the hyperparameters to explore their effects. As shown in Fig. 7 (b) and (c), CSIQA is found to be sensitive to the loss weight λ_1, while less sensitive to the loss weight λ_2. Specifically, small values of λ_1 weaken the impact of our logit distillation, while large values of λ_1 may result in learning too many noisy pseudo-labels and lead to performance degradation. Therefore, choosing an appropriate value of λ_1(e.g., 0.9) is critical for achieving optimal performance. On the other hand, our experiments show that CSIQA is less sensitive to λ_2. However, it still plays an essential role in the overall performance of our model. We recommend using a moderate value of λ_2(e.g., 0.1) to achieve a balance between contrastive learning and logit supervision.

Table 14. Context contrastive learning paradigm is applied to the IQA network ADANet (Pan et al., 2022) with the backbone of EfficientNet-b0 (Tan & Le, 2019) and RedNet101 (Li et al., 2021) (referred by “Effi-b0” and “Red101”, respectively)."+" refers to the model combined with our learning strategy.

	CSIQ		LIVEC		KonIQ	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Effi-b0	0.922	0.910	0.851	0.828	0.921	0.905
Effi-b0 +	$\mathbf{0.932}_{+1.0\%}$	$\mathbf{0.920}_{+1.0\%}$	$\mathbf{0.865}_{+1.4\%}$	$\mathbf{0.849}_{+2.1\%}$	$\mathbf{0.928}_{+0.7\%}$	$\mathbf{0.913}_{+0.8\%}$
Red101	0.942	0.931	0.757	0.734	0.881	0.877
Red101 +	$\mathbf{0.949}_{+0.7\%}$	$\mathbf{0.940}_{+0.9\%}$	$\mathbf{0.798}_{+4.2\%}$	$\mathbf{0.782}_{+4.8\%}$	$\mathbf{0.890}_{+0.9\%}$	$\mathbf{0.887}_{+1.0\%}$

Table 15. Comparison between contrastive learning and rank learning using the VGG16 backbone.

		LIVE		TID2013	
Method	Backbone	PLCC	SRCC	PLCC	SRCC
Learning to Rank	VGG16	0.976(+0.3%)	0.975(+0.4%)	0.803	0.791
Contrast Learning	VGG16	0.973	0.971	0.884(+8.1%)	0.863(+7.2%)

Table 16. Comparison between contrastive learning and rank learning using the ResNet backbone.

		KADID		KonIQ	
Method	Backbone	PLCC	SRCC	PLCC	SRCC
Learning to Rank	ResNet-50	0.822	0.820	0.896	0.884
Contrast Learning	ResNet-50	0.875(+5.3%)	0.881(+6.1%)	0.923(+2.7%)	0.910(+2.6%)

F. Qualitative Analysis

Visualization of activation map. In IQA datasets, there tend to be a large number of images with distinct foreground/background or without any specific object of interest (Sheikh et al., 2006; Ghadiyaram & Bovik, 2015). In the main paper, we show visualization results that show that our CSIQA can well focus on the distortion of the foreground part, as this is also more in line with the feature that the Human Visual System is significantly more concerned about (Ponomarenko et al., 2015; Hosu et al., 2020). In this section, we further visualize attention maps in images with distinct backgrounds or without any specific object of interest to embody the comprehensive ability of our CSIQA to capture image distortion. As shown in Fig. 8, compared to DEIQT (the second best method in Table 1), our CSIQA shows the superior ability to accurately focus distinct backgrounds or lacking any specific object image distortion regions which are highlighted in the green box (i.e., blur sun, dim clouds, and overly bright red lights), resulting in improved prediction.

Visualization of feature distributions across different quality scores. In Fig. 9, we further use t-SNE (Van der Maaten & Hinton, 2008)to visualize the feature distribution of CSIQA and DEIQT (Qin et al., 2023), across all images of KonIQ (a total of 10073 authentic images). The purple points represent lower-quality images (scoring less than 50 on a scale of 0-100), while the yellow points represent higher-quality images (scoring more than 50 on a scale of 0-100). As shown in Fig. 9(b), CSIQA demonstrates superior discriminative capability in distinguishing between high-quality and low-quality images. Specifically, the red box area highlights CSIQA’s ability to accurately capture the distinguishing features of images with varying quality levels. Moreover, the intra-class distance between images of the same quality level is noticeably more compact with CSIQA, indicating its effectiveness in producing more distinct feature representations.

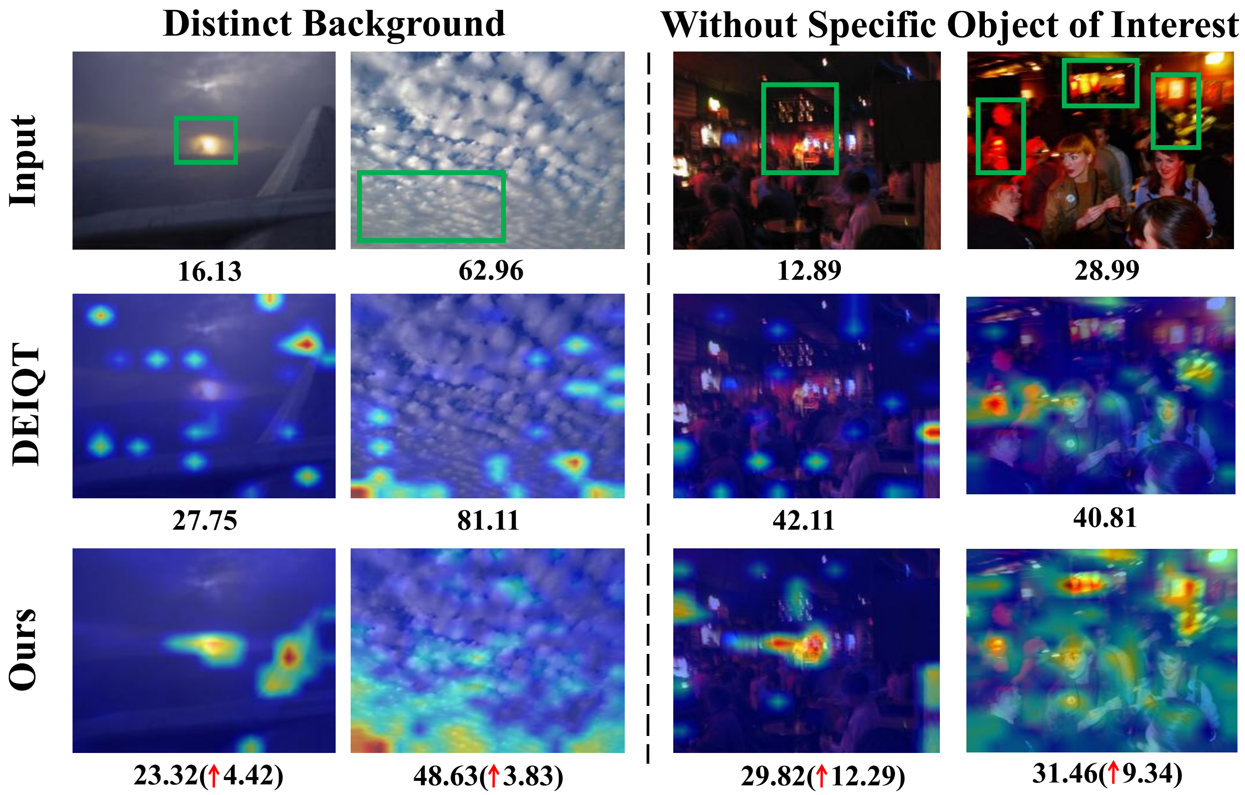
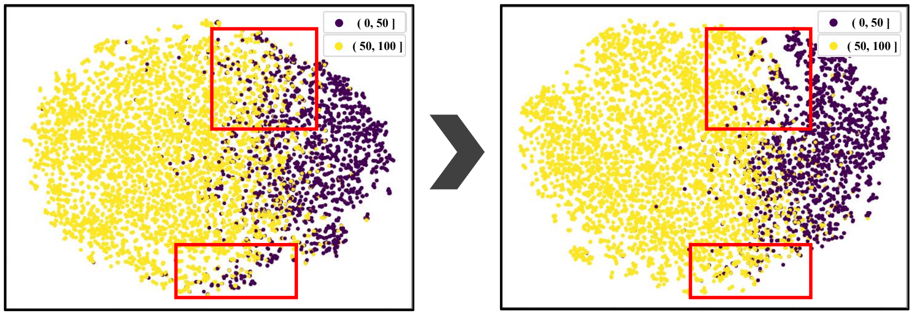


Figure 8. Activation maps of DEIQT (Qin et al., 2023) and CSIQA(ours) using Grad-CAM (Selvaraju et al., 2017). Scores in this figure represent *Mean Opinion Scores* (MOS). CSIQA still well-performs on background or without any specific object of interest, resulting in accurate prediction. Rows 1-3 represent input images, and CAMs from DEIQT and CSIQA, respectively.



(a) DEIQT

(b) CSIQA (Ours)

Figure 9. The t-SNE of the current SOTA method DEIQT and our CSIQA visualizes 10073 (all images) quality-aware representations learned from KonIQ.