

Distilling Spatially-Heterogeneous Distortion Perception for Blind Image Quality Assessment

Xudong Li¹ Wenjie Nie¹ Yan Zhang¹ Runze Hu² Ke Li³ Xiwu Zheng¹ Liujuan Cao^{1*}

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, China

³ Tencent Youtu Lab, Shanghai, China

{lxd761050753, bzhy986, hrzlpk2015}@gmail.com, {zhengxiawu, caoliujuan}@xmu.edu.cn

Abstract

In the Blind Image Quality Assessment (BIQA) field, accurately assessing the quality of authentically distorted images presents a substantial challenge due to the diverse distortion types in natural settings. Existing state-of-the-art IQA methods mix a sequence of distortions into entire images to establish global distortion priors, but are inadequate for authentic images with spatially varied distortions. To address this, we introduce a novel IQA framework that employs knowledge distillation tailored to perceive spatially heterogeneous distortions, enhancing quality-distortion awareness. Specifically, we introduce a novel *Block-wise Degradation Modelling* approach that applies distinct distortions to different spatial blocks of an image, thereby expanding local distortion priors. Following this, we present a *Block-wise Aggregation and Filtering* module that enables fine-grained attention to the quality information within different distortion areas of the image. Furthermore, to effectively capture the complex relationships between distortions across different regions while preserving overall quality perception, we introduce *Contrastive Knowledge Distillation* to enhance the model's ability to discriminate between different types of distortions and *Affinity Knowledge Distillation* to model the correlation among distortions in different regions. Extensive experiments on standard BIQA datasets demonstrate the effectiveness and competitiveness of the proposed method.

1. Introduction

Image Quality Assessment (IQA) methods are designed to evaluate images in a way that aligns with the human judgment of visual distortions [47, 49]. These methods are crucial for applications that rely heavily on image quality

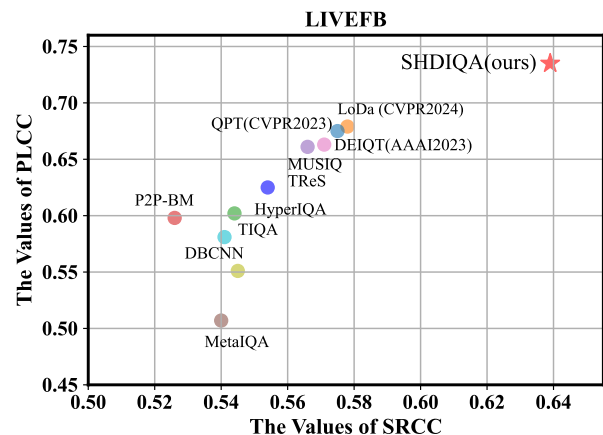


Figure 1. Comparison of state-of-the-art (SOTA) BIQA methods on the largest authentic dataset, LIVEFB. The proposed SHDIQA method outperforms all existing BIQA approaches.

and serve as standards to gauge the performance of image processing techniques. Objective Image Quality Assessment includes Full-Reference IQA (FR-IQA) [36], Reduced-Reference IQA (RR-IQA) [41], and no-reference IQA (NR-IQA) or Blind IQA (BIQA) [60]. As the presence of various types of artifacts in real-world scenes and the unavailability of reference images, BIQA gains attention for tasks like image restoration [1] and super-resolution [6].

Recent advances in data-driven BIQA models, especially those based on deep neural networks, have led to impressive progress [37, 51]. However, these models still rely heavily on Mean Opinion Scores (MOS), which are expensive and time-consuming to collect—often requiring up to 120 human ratings per image [13]. As an alternative, synthetic distortions generated through controlled data augmentation offer a cost-effective way to create annotated datasets [48], enabling the learning of quality representations without extensive human labeling [38]. To leverage such synthetic data, recent studies have introduced multi-task end-to-end

*Corresponding author.

frameworks [26, 57] that jointly optimize distortion-specific objectives and quality prediction, using synthetic distortions to enrich training signals and improve generalization. Another line of work focuses on image degradation modeling. Methods like Re-IQA [31] and QPT [58] simulate realistic distortions by applying diverse types and combinations to entire images, as illustrated in Fig.2. They then leverage unsupervised learning frameworks such as MoCo-v2[3] to extract distortion priors from large-scale synthetic datasets.

However, both categories of methods share a key limitation: they focus on global distortion priors while ignoring fine-grained local distortions, which are essential for accurate BIQA. Research [15, 17] shows that the human visual system (HVS) employs both local detection and global appearance strategies when assessing image quality. For subtle, near-threshold distortions, the HVS prioritizes local details, while for highly visible distortions, it shifts focus back to global content. This highlights the vital role of local distortion features in comprehensive image quality evaluation.

In this paper, we propose a novel Spatial Heterogeneous Distortion-aware Image Quality Assessment (SHDIQA) method, designed to improve the local distortion-awareness of BIQA models based on the Vision Transformer (ViT) architecture. Unlike prior works [10, 14, 50] that focus on modifying ViT architectures to enhance local feature extraction, our method fully leverages rich local distortion priors generated from synthetic distortions, while introducing minimal architectural changes (only 0.2M additional parameters).

Specifically, we introduce a Block-wise Image Degradation (BWID) method tailored for IQA tasks to generate patch-level distortion priors. As shown in Fig. 2, BWID partitions a distorted image into multiple blocks and applies a randomly selected distortion from 24 categories across 4 severity levels to each block. This results in a composite image with spatially varying local distortions. Each image naturally comes with an image-level multi-class label and corresponding patch-level single-class labels, effectively addressing the lack of localized distortion annotations in existing methods. To further utilize the patch-level distortion priors, we propose a Block-wise Aggregation and Filtering (BWAf) mechanism. BWAf uses the local priors from BWID to train multiple distortion (DIS) tokens, which interact with a global classification (CLS) token through a cross-attention mechanism with an adaptive head. This interaction mimics the HVS’s dual strategy of local detection and global perception, thereby improving overall quality assessment, as illustrated in Fig. 1. Moreover, previous studies [16, 40] highlight the importance of modeling the relationships among different distortion types in perceptual quality assessment. To achieve this, we incorporate Contrastive Knowledge Distillation (CKD) and Affinity Knowledge Distillation (AKD). CKD aligns similar distortions and separates dissimilar ones, improving the model’s ability to differentiate between distortion

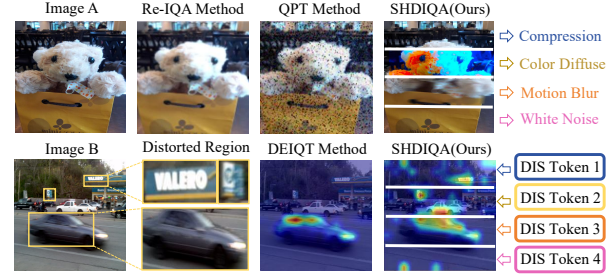


Figure 2. The proposed SHDIQA method creates a heterogeneous degradation space with richer priors (Top) and assigns dedicated distortion-aware (DIS) tokens to each local region (Bottom).

tion types. Meanwhile, AKD transfers knowledge about the relationships between different distortions, enhancing the model’s generalization across various types of distortions.

Our contributions are summarized as follows:

- We present a Block-wise Degradation Modelling approach that enriches the local distortion priors of the training datasets by dividing images into blocks and applying distinct distortions to each, enhancing the model’s ability to assess image quality accurately in complex scenarios.
- We introduce the Block-wise Aggregation and Filtering mechanism to fully capture and leverage localized distortions. It first directs precise attention to distortion details within image blocks, and then utilizes an adaptation head to filter out irrelevant distortion features, focusing the analysis on quality-aware information.
- We propose a novel Distortion-Guided Knowledge Distillation strategy that combines Contrastive Knowledge Distillation (CKD) to enhance distortion discrimination and Affinity Knowledge Distillation (AKD) to enrich relational insights between local distortion areas.

2. Related Work

2.1. BIQA with Degradation Modelling

Recent studies [10, 61] have shown that simple distortion simulations do not fully capture the complexities of real-world degradations. To address this gap, Image Degradation Modeling has been introduced to represent various distortion types [34, 45, 46]. DBCNN [56] decomposes real distortions into multiple synthetic ones, creating large-scale synthetic images for training. DipIQ [25] categorizes common distortions like noise, blur, and compression into severity levels, while Re-IQA [31] introduces 25 distortion types to expand the degradation space. QPT [58] models second-order degradation using skip and shuffle operations across nine types in three groups. In contrast to these global degradation approaches, our method applies localized distortions to different regions of the image, enabling the model to learn fine-grained distortion perception both within and across blocks through knowledge distillation.

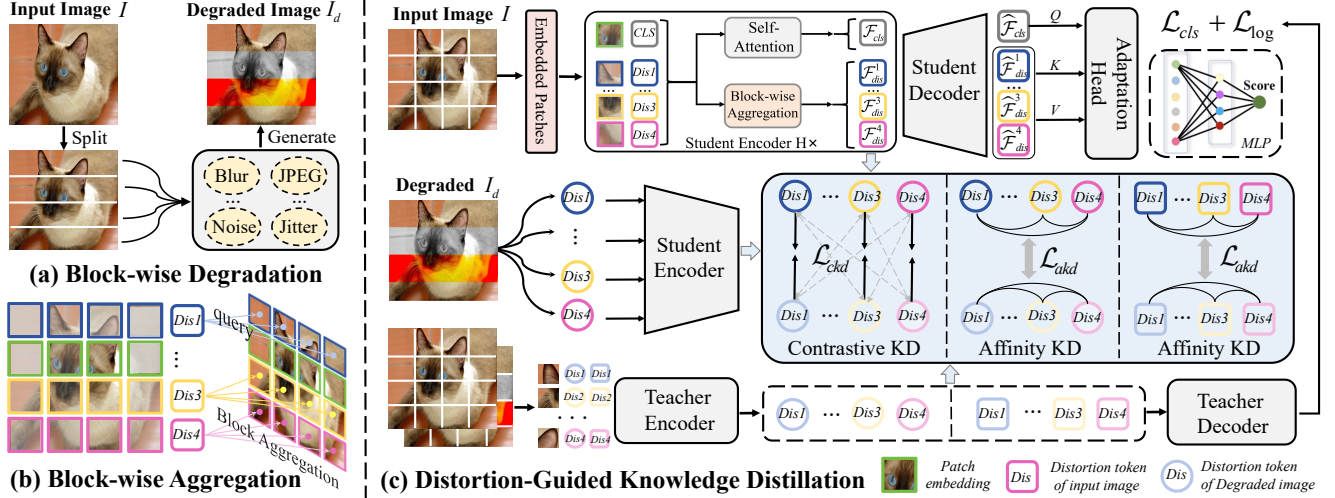


Figure 3. **Framework Overview.** Given an input image I , (a) Block-wise Degradation is first applied to generate a degraded version I_d (Sec. 3.1). Both I and I_d are then fed into a student encoder-decoder network, which is initialized with multiple distortion (DIS) tokens (each attending to a specific local block) and a CLS token for global quality representation. (b) Guided by the Block-wise Aggregation (Sec. 3.2), the DIS tokens and the CLS token are used to extract local distortion features $\mathcal{F}_{dis}$ and a global quality feature $\mathcal{F}_{cls}$, which are further refined by the decoder into $\hat{\mathcal{F}}_{dis}$ and $\hat{\mathcal{F}}_{cls}$. These refined features are processed by a quality-aware adaptation head to suppress irrelevant information and predict the quality score via an MLP. (c) To further enhance distortion perception, Contrastive KD is used to improve inter-block distortion discrimination, while Affinity KD captures relationships between distortions across different regions (Sec. 3.3).

2.2. BIQA with Transformers

Vision Transformer (ViT)-based IQA architectures include hybrid transformers [2, 10, 21, 50, 54] and pure transformers [14, 20, 29]. Hybrid models integrate multi-scale CNN information for better local distortion perception but increase complexity. Pure transformers like DEIQT [29] capture global abstraction using the CLS token but struggle with local structure modeling [63]. To reduce memory overhead, methods using local self-attention and sparse block structures have been developed [23, 30], though at the cost of some long-distance modeling. Sampling-based methods like VT-IQA [17] combine global and local information by resampling distorted images and using global attention. However, they lack explicit supervision for local distortions, which limits their ability to capture localized degradations during training. In contrast, our approach enhances local distortion perception in pure ViT by incorporating patch-level distortion priors, while preserving global context modeling through a mask matrix specifically applied to distortion tokens, leaving the self-attention process unaffected.

2.3. IQA with Knowledge Distillation

Knowledge Distillation (KD), originally proposed for model compression [12], has been adapted for tasks like action recognition [8], re-identification [28], and depth estimation [11]. Recently, KD has been used in methods like Degraded-Reference IQA (DR-IQA) [59] and Non-aligned Reference IQA (NAR-IQA) [52] to transfer reference information, though these still partially rely on reference images.

A related BIQA method is SSLIQA [55], which employs KD to enforce semantic consistency across different triples within a batch. However, our approach differs significantly in that our KD explicitly targets inter-block and intra-block distortion information, allowing for more direct extraction of quality-related distortion priors from the augmented image.

3. Methodology

We introduce Spatial-Heterogeneous Distortion-aware Image Quality Assessment (SHDIQA), a framework designed to enhance the perception of local and non-uniform distortions in authentic images, as illustrated in Fig. 3. Specifically, the proposed SHDIQA first divides the input image I into vertical blocks and applies random distortions using a Block-wise Degradation module (Sec. 3.1), producing a degraded image I_d . Both I and I_d are then processed by a student encoder-decoder network equipped with a Block-wise Aggregation module (Sec. 3.2), which extracts local distortion features $\mathcal{F}_{dis}$ via multiple distortion (DIS) tokens and global quality feature $\mathcal{F}_{cls}$ via classification (CLS) token. Next, the local and global features $\mathcal{F}_{dis}$ and $\mathcal{F}_{cls}$ are fused via a cross-attention mechanism in the adaptation head to suppress irrelevant distortion information (Sec. 3.2). To enhance the model's ability to represent and distinguish distortion patterns, we further introduce Contrastive Knowledge Distillation (CKD) and Affinity Knowledge Distillation (AKD) (Sec. 3.3). These modules jointly improve distortion discriminability and capture inter-block distortion correlations, enabling a more distortion-aware representation.

3.1. Block-wise Degradation Modelling

Recent methods [31, 58] apply multiple distortions uniformly across the entire image, ignoring the finer details of local distortions. To address this, we propose a Block-wise Degradation Modelling strategy, which applies 24 distortion types and 4 degradation levels to different regions of the image, thereby generating a composite local distortion prior. Specifically, given an image of size $W \times H$, we divide it into K vertical blocks of size $W \times \frac{H}{K}$. For each block k ($k = 1, \dots, K$), we randomly select a rectangular region C_k with top-left coordinate α_i and size $w_k \times h_k$, where $w_k = \beta_w \times W$ and $h_k = \beta_h \times \frac{H}{K}$. Empirically, we find that setting $\beta_w = \beta_h = 1$, *i.e.*, degrading the entire block, yields the best performance. Then, the degradation is applied using the function $\mathcal{D}_k(C_k; q_k, p_k)$, where q_k denotes the type of distortion (e.g., noise, darkening), and p_k represents its intensity. The degraded image I_d is generated as follows:

$$I_d = \mathcal{A}(I) = \mathcal{D}_K(\dots \mathcal{D}_1(C_1; q_1, p_1) \dots; q_K, p_K) \quad (1)$$

where $K = 4$ denotes the number of degraded blocks. Unlike previous methods, our approach focuses on localized distortions, providing richer priors for feature aggregation (Sec. 3.2) and knowledge distillation (Sec. 3.3), improving the model's ability to handle complex real-world distortions.

3.2. Block-wise Aggregation and Filtering

In this section, we introduce the Block-wise Aggregation method to extract the local distortion feature $\mathcal{F}_{dis}$ from the degraded image I_d and the input image I . We then present the Quality-aware Adaptation Head, which filters out quality-irrelevant distortion features from distortion feature $\mathcal{F}_{dis}$.

3.2.1. Block-wise Aggregation

We begin by randomly initializing K local distortion-aware (DIS) tokens, denoted as $\mathcal{F}_{dis}$, matching the number of degraded regions. Each DIS token $\mathcal{F}_{dis}^k$ corresponds to a specific region C_k . The resulting embedding sequence is defined as $\mathcal{F} = \{\mathcal{F}_{dis}; [\mathcal{F}_{cls}; \mathcal{F}_{patch}]\} \in \mathbb{R}^{(K+N) \times D}$, where N is the number of image patches plus one CLS token, and D is the embedding dimension. To ensure each DIS token focuses on its associated distorted region C_k , we construct a mask map $\mathbf{m} \in \mathbb{R}^{(K+N)}$. Based on this, we define the attention mask $\mathcal{M}$ to preserve the original global attention structure while enforcing localized attention for DIS tokens as follows:

$$m_j^i = \begin{cases} 1 & \text{the } i\text{-th patch overlaps with } C_i, \quad j < N \\ 0, & j = N \end{cases},$$

$$\mathbb{M} = [m^1; m^2; \dots; m^K]^T, \mathcal{M} = \begin{bmatrix} \mathbf{1}_{N \times N} & \mathbf{0}_{N \times K} \\ \mathbb{M}_{K \times N} & \mathbf{1}_{K \times K} \end{bmatrix}, \quad (2)$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Intuitively, Eq. 2 enforces that each DIS token $\mathcal{F}_{dis}^k$ attends only to the distorted

content in block C_k . Assuming the encoder has H layers, the Multi-Head Self-Attention is applied as follows:

$$\mathcal{F}^h = \begin{cases} \sigma(\log \mathcal{M} + \mathbf{Q}^h (\mathbf{K}^h)^T) \mathbf{V}^h + \mathcal{F}^{h-1}, & 0 < h \leq H \\ \mathcal{F}, & h = 0 \end{cases}, \quad (3)$$

where $\mathbf{Q}^h, \mathbf{K}^h, \mathbf{V}^h$ are the query, key, and value matrices, obtained via linear projections of $\mathcal{F}^{h-1}$, and σ denotes the Softmax function. The encoder outputs features $\mathcal{F}^H$ after H layers. Following [29], we apply a transformer decoder to refine these features, yielding $\{\hat{\mathcal{F}}_{dis}; \hat{\mathcal{F}}_{cls}\} \in \mathbb{R}^{(K+1) \times D}$, which are more relevant for image quality.

3.2.2. Quality-aware Adaptation Head

To effectively integrate global quality and local distortion features while suppressing quality-irrelevant information, we design a quality-aware adaptation head that highlights distortion cues closely related to perceived quality. Specifically, the global quality feature $\hat{\mathcal{F}}_{cls}$ is used as the query to selectively attend to the local distortion features $\hat{\mathcal{F}}_{dis}$, which serve as key and value. This allows the model to emphasize distortion patterns most relevant to the overall quality. Overall, the final predicted quality score $\hat{Y}$ is computed as:

$$\hat{Y} = \text{LN} \left(\hat{\mathcal{F}}_{cls} + \text{CrossAttention}(\hat{\mathcal{F}}_{cls}, \hat{\mathcal{F}}_{dis}) \right), \quad (4)$$

where $\text{CrossAttention}(\cdot)$ denotes a cross-attention module, and LN is a single-layer linear projection. During training, the model is optimized by minimizing the smooth $\mathcal{L}_1$ loss.

3.3. Distortion-Guided Knowledge Distillation

Inspired by [5, 18, 33], we adopt a knowledge distillation method to learn correlations among distortion types by transferring quality-aware distortion knowledge from the teacher to the student. This distillation strategy addresses two key challenges: First, the student's ViT encoders, pre-trained on ImageNet [4], focus on semantics and lack distortion awareness in early training [19]. The teacher, pre-trained on IQA data, provides quality-aware guidance, directing the student to critical distortion patterns. Second, the absence of explicit ground truth for distortion correlations makes it difficult for the student to model these relationships independently. The teacher acts as a reliable target for correlation modeling, thereby improving the student's discriminative ability.

3.3.1. Contrastive Knowledge Distillation

Contrastive Knowledge Distillation (CKD) aims to extract the distortion features $\mathbf{V}_{dis}^k$ of the k -th local block from a teacher encoder $\mathcal{V}(\cdot)$ and transfer them to a student network. Simply maximizing feature similarity is insufficient to capture distinct distortion representations across spatial domains (Table 5). To address this, we reformulate the KD task as a pseudo-classification problem. As shown in Fig. 3, features from corresponding distorted regions of the teacher

Method	LIVE		CSIQ		KADID		LIVEC		KonIQ		LIVEFB		SPAQ	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BRISQUE [27]	0.944	0.929	0.748	0.812	0.567	0.528	0.629	0.629	0.685	0.681	0.341	0.303	0.817	0.809
DBCNN [56]	0.971	0.968	0.959	0.946	0.856	0.851	0.869	0.851	0.884	0.875	0.551	0.545	0.915	0.911
TIQA [54]	0.965	0.949	0.838	0.825	0.855	0.85	0.861	0.845	0.903	0.892	0.581	0.541	-	-
MetalQA [61]	0.959	0.960	0.908	0.899	0.775	0.762	0.802	0.835	0.856	0.887	0.507	0.54	-	-
HyperIQA (27M) [39]	0.966	0.962	0.942	0.923	0.845	0.852	0.882	0.859	0.917	0.906	0.602	0.544	0.915	0.911
TReS (152M) [10]	0.968	0.969	0.942	0.922	0.858	0.859	0.877	0.846	0.928	0.915	0.625	0.554	-	-
MUSIQ (27M) [14]	0.911	0.940	0.893	0.871	0.872	0.875	0.746	0.702	0.928	0.916	0.661	0.566	0.921	0.918
Re-IQA (48M) [31]	0.971	0.970	0.960	0.947	0.885	0.872	0.854	0.840	0.923	0.914	-	-	0.925	0.918
DEIQT (24M) [29]	0.982	0.980	0.963	0.946	0.887	0.889	0.894	0.875	0.934	0.921	0.663	0.571	0.923	0.919
QPT (24M) [58]	-	-	-	-	-	-	<u>0.914</u>	0.895	0.941	0.927	0.675	0.575	<u>0.928</u>	<u>0.925</u>
LIQE (151M) [57]	0.951	0.970	0.939	0.936	0.931	0.930	0.910	<u>0.904</u>	0.908	0.919	-	-	-	-
QFM-IQM (24M) [19]	<u>0.983</u>	<u>0.981</u>	<u>0.965</u>	<u>0.954</u>	0.906	0.906	0.913	0.891	0.936	0.922	0.667	0.567	0.924	0.920
LoDa (118M) [50]	0.979	0.975	-	-	<u>0.936</u>	<u>0.931</u>	0.899	0.876	<u>0.944</u>	<u>0.932</u>	<u>0.679</u>	<u>0.578</u>	<u>0.928</u>	<u>0.925</u>
SHDIQA (24.2M)	0.984	0.982	0.972	0.962	0.940	0.937	0.922	0.909	0.948	0.937	0.735	0.639	0.934	0.930

Table 1. Performance comparison measured by the average of SRCC and PLCC. Best results are in bold, second-best are underlined.

and student are treated as positive pairs, while all other pairs are considered negative. The CKD loss is defined as:

$$\mathcal{L}_{ckd} = -\frac{1}{K} \sum_k \log \left(\sigma \left(\frac{\cos(\mathcal{F}_{dis}^k, \mathbf{V}_{dis}^k)}{\tau_c} \right) \right) - \frac{1}{K} \sum_{\substack{k,j \\ j \neq k}} \log \left(1 - \sigma \left(\frac{\cos(\mathcal{F}_{dis}^k, \mathbf{V}_{dis}^j)}{\tau_c} \right) \right), \quad (5)$$

where $\cos(\cdot)$ is cosine similarity, σ is the sigmoid function, and $\tau_c = 0.3$ is the temperature factor. $\mathcal{F}_{dis}^k$ is the student feature vector of the k -th local block feature extracted by the DIS token. This approach enhances contrast among features of varying distortion types, improving the model’s discrimination and distortion-awareness capabilities.

3.3.2. Affinity Knowledge Distillation

Affinity Knowledge Distillation (AKD) focuses on distilling the interrelations among multiple blocks within an image I or its degraded version I_d . As shown in Fig. 3, given teacher embeddings $\mathbf{V}_{dis}$ and student features $\mathcal{F}_{dis}$, we compute student and teachers’ pairwise affinity matrices:

$$\mathbf{R}_{stu} = \bar{\mathbf{F}} \bar{\mathbf{F}}^T, \quad \mathbf{R}_{tea} = \bar{\mathbf{V}} \bar{\mathbf{V}}^T, \quad (6)$$

where $\bar{\mathbf{F}}$ and $\bar{\mathbf{V}}$ are $\mathcal{L}_2$ -normalized student $\mathcal{F}_{dis}$ and teacher features $\mathbf{V}_{dis}$, respectively. These matrices capture pairwise correlations, and the goal is to align the student affinity matrix $\mathbf{R}_{stu}$ with the teacher matrix $\mathbf{R}_{tea}$:

$$\mathcal{L}_{akd} = -\frac{1}{K} \sum_{k=1}^K \text{KL} \left(\sigma \left(\frac{\mathbf{R}_{stu}^k}{\tau_a} \right) \parallel \sigma \left(\frac{\mathbf{R}_{tea}^k}{\tau_a} \right) \right), \quad (7)$$

where $\tau_a = 2$ and $\text{KL}(\cdot)$ denotes the Kullback-Leibler divergence. AKD complements CKD by focusing on the relational dynamics between image blocks, allowing the model to learn the relationship between distortion. However, relying solely on AKD may overlook fine-grained distortion-discriminative details, as highlighted in Table 2. Therefore, both CKD and AKD are essential for learning comprehensive distortion-aware features.

3.3.3. Overall Training Loss

The overall loss combines four complementary components: $\mathcal{L}_{cls}$ provides category-level supervision for distortion identification but lacks awareness of relationships between distortion types; $\mathcal{L}_{ckd}$ enhances feature discrimination by clustering similar distortions and separating dissimilar ones, thus compensating for this limitation; $\mathcal{L}_{akd}$ further captures global perceptual structures by modeling relative similarities across different distortion. and $\mathcal{L}_{log}$ regularizes training via teacher outputs [44], mitigating overfitting to quality-irrelevant patterns. Finally, the total loss is defined as:

$$\mathcal{L} = \lambda_{cls} \mathcal{L}_{cls} + \lambda_{ckd} \mathcal{L}_{ckd} + \lambda_{akd} \mathcal{L}_{akd} + \lambda_{log} \mathcal{L}_{log} + \|\hat{Y} - Y_g\|_1, \quad (8)$$

where Y_g is the ground truth, $\mathcal{L}_{log}$ is the $\mathcal{L}_1$ loss between student and teacher predictions, and λ_{cls} , λ_{ckd} , and λ_{akd} are set to 1, with λ_{log} set to 0.8.

4. Experiments

4.1. Benchmark Datasets and Evaluation Protocols

Experiments were conducted on 7 typical BIQA datasets, including 3 synthetic datasets LIVE [35], CSIQ [15], KADID [22], and 4 authentic datasets LIVEC [9], KonIQ [13], LIVEFB [53], SPAQ [7]. Specifically, for authentic datasets,

$Loss \mathcal{L}_{log}$	$Loss \mathcal{L}_{cls}$	$Loss \mathcal{L}_{akd}$	$Loss \mathcal{L}_{ckd}$	CSIQ		KADID		LIVE		LIVEC		Avg.
				PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	
$\times$	$\times$	$\times$	$\times$	0.952	0.935	0.878	0.884	0.966	0.964	0.891	0.867	0.917
$\checkmark$	$\times$	$\times$	$\times$	0.955	0.941	0.898	0.904	0.972	0.970	0.898	0.872	0.926
$\checkmark$	$\checkmark$	$\times$	$\times$	0.959	0.947	0.919	0.916	0.974	0.972	0.904	0.890	0.935
$\checkmark$	$\times$	$\checkmark$	$\times$	0.970	0.960	0.927	0.925	0.980	0.978	0.919	0.905	0.945
$\checkmark$	$\times$	$\times$	$\checkmark$	0.966	0.955	0.913	0.911	0.976	0.974	0.914	0.902	0.939
$\checkmark$	$\times$	$\checkmark$	$\checkmark$	0.969	0.958	0.937	0.935	0.983	0.981	0.920	0.904	0.948
$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	0.972	0.962	0.940	0.937	0.984	0.982	0.922	0.909	0.951

Table 2. Ablation experiments about knowledge distillation components on multiple datasets, where bold entries indicate the best result.

Training	LIVEFB	LIVEC	KonIQ	LIVE	CSIQ	
Testing	KonIQ	LIVEC	KonIQ	LIVEC	CSIQ	LIVE
DBCNN	0.716	0.724	<u>0.754</u>	0.755	0.758	0.877
P2P-BM	<u>0.755</u>	0.738	0.740	0.770	0.712	-
TReS	0.713	0.740	0.733	0.786	0.761	-
DEIQT	0.733	0.781	0.744	0.794	<u>0.781</u>	<u>0.932</u>
LoDa	0.763	0.805	0.745	<u>0.811</u>	-	-
SHDIQA	0.723	<u>0.791</u>	0.782	0.819	0.789	0.943

Table 3. SRCC on the cross datasets validation. The best results are highlighted in bold, second-best are underlined.

LIVEC contains 1162 images from different mobile devices and photographers. SPAQ includes 11125 photos from 66 smartphones. KonIQ contains 10073 images from public sources, while LIVEFB is the largest real-world dataset with 39810 images. The synthetic datasets involve artificially distorting original images using methods such as JPEG compression and Gaussian blur. LIVE and CSIQ respectively have 779 and 866 synthetic distortion images, with 5 and 6 types of distortions. KADID contains 10125 synthetic distortion images, spanning 25 types of distortions.

Evaluation Criteria. The experiments use Pearson’s Rank-Order Correlation Coefficient (SRCC) and Pearson’s Linear Correlation Coefficient (PLCC) to assess prediction monotonicity and accuracy, respectively. Both metrics scale from -1 to 1, with values close to 1 indicating better performance.

4.2. Implementation Details

For the student network, we follow the typical training strategy [29], using a Transformer encoder based on ViT-S from DeiT III [43] and a one-layer decoder. Our model is trained for 9 epochs with a learning rate of 2×10^{-4} , decaying by a factor of 10 every 3 epochs. The batch size is 16 for LIVEC and 128 for KonIQ. We use 80% of the images for training and 20% for testing, repeating the process 10 times to reduce performance bias, reporting the average SRCC and PLCC. For synthetic distortion datasets, training and testing sets are divided by reference images to ensure con-

Mode	Methods	LIVEC		KonIQ	
		PLCC	SRCC	PLCC	SRCC
20%	DEIQT	0.822	0.792	0.908	0.888
	LoDa	0.854	0.815	0.923	0.907
	SHDIQA	0.866	0.840	0.928	0.915
40%	DEIQT	0.855	0.838	0.922	0.903
	LoDa	0.879	0.849	0.935	0.922
	SHDIQA	0.889	0.869	0.942	0.929

Table 4. Data-efficient learning validation with the training set containing 20% and 40% of the images.

tent independence. To provide reliable distortion perception early in training, the teacher model adopts a similar encoder-decoder architecture aligned with the student model, and is pre-trained with $\mathcal{L}_1$ loss on the same IQA dataset. Training is performed on four NVIDIA RTX 3090 GPUs.

4.3. Overall Prediction Performance Comparison

Table 1 reports the comparison results between the proposed SHDIQA and other state-of-the-art (SOTA) BIQA methods, including ViT-based IQA methods, *i.e.*, LoDa [50] and DEIQT [29], and Degradation Modelling-based methods, *i.e.*, Re-IQA [31] and QPT [58]. It is observed from these seven datasets that SHDIQA achieves superior performance over all other methods across the 7 datasets. Notably, on the most difficult LIVEFB dataset, SHDIQA achieves an impressive performance improvement. This enhancement can be attributed to the prevalence of local distortions [53] in the LIVEFB dataset, highlighting the importance of effectively utilizing local distortion priors.

4.4. Data-Efficient Learning Validation

We showcase the data efficiency of our approach, as presented in Table 4. Following the efficient validation protocol outlined in LoDa [50] and DEIQT [29], we train models using 20% and 40% of the available images. This process is repeated 10 times, and we report the average SRCC and PLCC scores. By introducing distortion priors through the

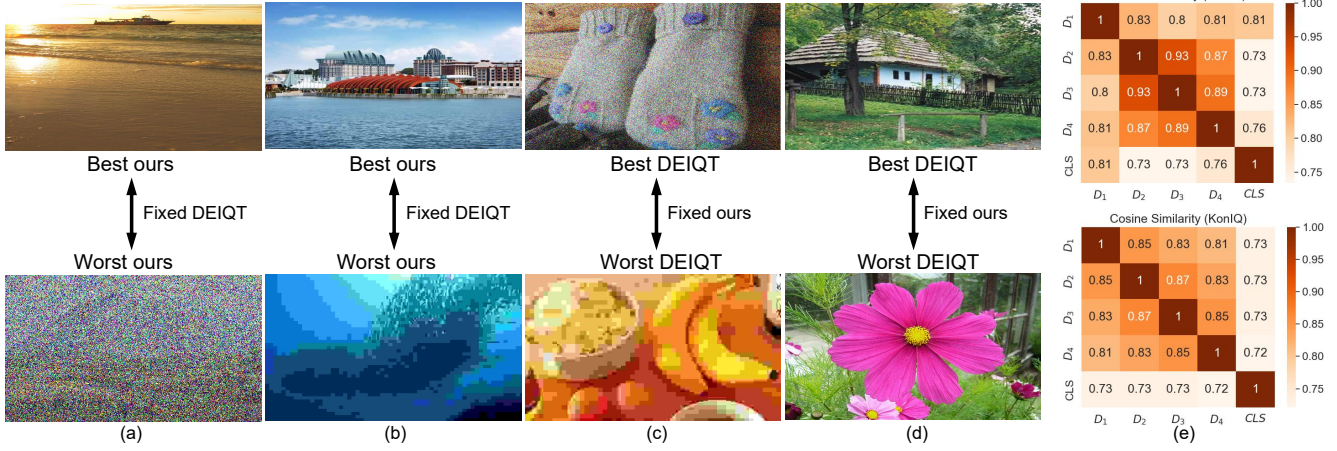


Figure 4. gMAD competition results between DEIQT and our method SHDIQA. (a) Fixed DEIQT at low quality. (b) Fixed DEIQT at high quality. (c) Fixed SHDIQA at low quality. (d) Fixed SHDIQA at high quality. (e) Cosine similarity between different DIS and CLS tokens.

$\mathcal{L}_{\text{ckd}}$	LIVEC		CSIQ		Avg.
	PLCC	SRCC	PLCC	SRCC	
-	0.891	0.867	0.952	0.935	0.911
$\mathcal{L}_1$	0.908	0.895	0.957	0.946	0.926
BCE	0.914	0.902	0.966	0.955	0.934

Table 5. Ablation study on Contrastive Knowledge Distillation.

Block-wise Degradation, our method achieves performance that is nearly with SOTA results, even when trained on only 20% of the dataset. Furthermore, our approach consistently outperforms all other methods on both datasets, demonstrating its superiority over existing efficient BIQA techniques.

4.5. Generalization Capability Validation

We evaluate SHDIQA’s generalization ability using cross-dataset validation, training the BIQA model on one dataset and testing it on others. Table 3 shows the SRCC averages for five datasets, where SHDIQA achieves the best performance in four out of six cross-dataset evaluations. To further test our framework’s generalization, we trained the model on the LIVE database and tested it on the Waterloo database using the gMAD competition [24]. In the gMAD competition, image pairs are selected where the attacker model predicts a large quality difference, while the defender perceives them as similar. Observers then evaluate these pairs to assess the model’s generalization ability. As shown in Fig. 4, our model correctly selects image pairs with significant quality differences when attacking DEIQT, where DEIQT misclassifies pairs like (a) as low quality and (b) as high quality, contrary to human perception and SHDIQA’s predictions. As a defender, SHDIQA identifies pairs with minimal perceptual variation, correctly classifying image pairs like (c) and (d) as both low and high quality, respectively, while DEIQT fails. This demonstrates the strong generalization capability of SHDIQA in handling real-world image distortions.

$\mathcal{L}_{\text{akd}}^I$	$\mathcal{L}_{\text{akd}}^{I_d}$	LIVEC		LIVE		Avg.
		PLCC	SRCC	PLCC	SRCC	
✓		0.917	0.901	0.982	0.978	0.944
	✓	0.915	0.900	0.980	0.980	0.943
✓	✓	0.922	0.909	0.984	0.982	0.949

Table 6. Ablation experiment on Affinity Knowledge Distillation.

4.6. Ablation Study

4.6.1. Effectiveness of Proposed Components

We conduct an ablation study on components of our framework. Table 2 shows that using the teacher’s quality score as a regularization improves model performance. Adding a distortion classification task boosts SRCC by 1.2 and 1.8 points on KADID and LIVEC, respectively, highlighting the effectiveness of distortion-aware supervision. However, identifying distortions without modeling their relationships provides limited benefits. To address this, we introduce CKD and AKD to capture distortion correlations, improving SRCC by up to 3.5 and 4.9 points on KADID, respectively. The limited gain from CKD suggests that aggressively separating regions with different distortions may be problematic, as these regions can have similar quality. AKD mitigates this issue by capturing structural relationships between image blocks. When combined, CKD and AKD improve the baseline by 6.2 points, confirming their complementarity. Finally, Table 8 shows that block-wise aggregation and filtering further enhance the model’s ability to assess local distortions, leading to better quality prediction.

4.6.2. Loss type of CKD and Implementation of AKD

Table 5 shows the effect of different CKD loss functions. Using $\mathcal{L}_1$ loss improves average quality prediction by 1.5 points, and switching to a BCE-based contrastive loss adds another 0.8-point gain. This confirms the effectiveness of our CKD design in promoting feature clustering within the

Number	LIVEC	
	PLCC	SRCC
3	0.920	0.903
4	0.922	0.909
5	0.918	0.905

Table 7. Ablation study on the number of degenerated blocks.

Method	LIVEC		LIVE	
	PLCC	SRCC	PLCC	SRCC
Baseline	0.891	0.867	0.966	0.964
+BWA	0.918	0.902	0.982	0.978
+BWF	0.922	0.909	0.984	0.982

Table 8. Ablation study on the Block-wise Aggregation (BWA) and Filtering (BWF) module.

Method	LIVEC		KonIQ	
	PLCC	SRCC	PLCC	SRCC
Homogenous	0.907	0.893	0.939	0.930
Heterogenous	0.922	0.909	0.948	0.937

Table 9. Ablation experiments on heterogeneous and homogeneous distortion on the whole image.

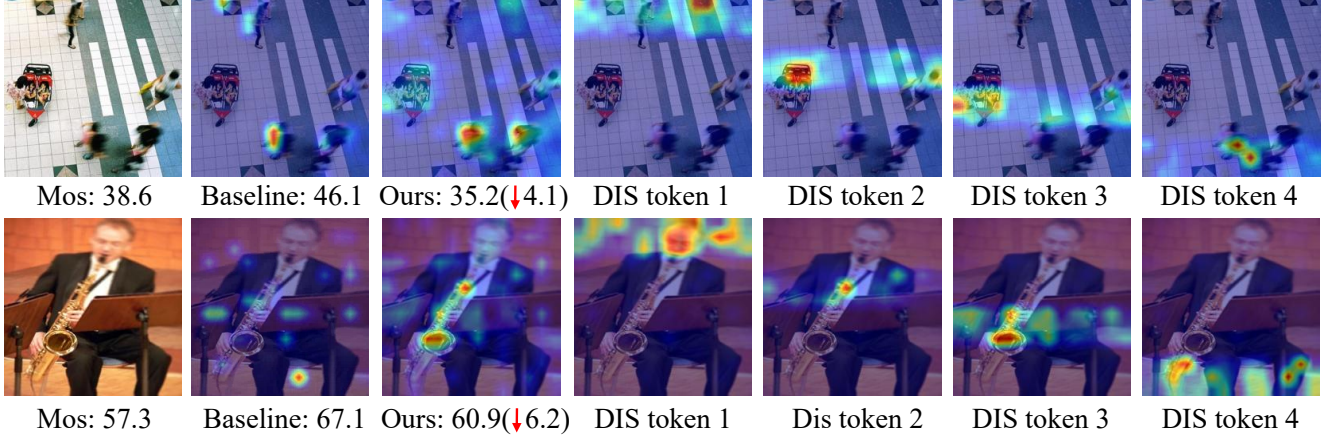


Figure 5. Activation maps of DEIQT [29] and SHDIQA using Grad-CAM [32] on the LIVEFB dataset, with MOS shown below the images. Our SHDIQA effectively captures the interplay between image content, local distortions, and overall quality, enabling accurate predictions.

same distortion type and separation across types. For AKD, we evaluate three strategies: applying it to the input image I , the degraded image I_d , or both. Table 6 shows that the best performance is achieved by using both, as applying AKD to the image I helps model relationships within local content, while applying it to the I_d captures correlations between different types of distortions. Combining both enables complementary learning of content and distortion information.

4.6.3. Ablation on the Degradation Space

We investigate the effectiveness of the Block-wise Degradation Modelling. As shown in Table 9, applying globally uniform heterogeneous distortions significantly underperforms compared to localized distortions, highlighting the importance of local degradation. We also evaluate how varying the number of degraded blocks affects performance. As shown in Table 7, the impact is relatively insensitive across different partitioning strategies, demonstrating the method’s resilience to changes in block quantity. While more blocks increase the scope of degradation, they also lead to greater content loss. Thus, using four blocks strikes the best balance.

4.7. Qualitative Analysis

To show that tokens from different perspectives capture unique features, we analyze the cosine similarity between CLS and DIS tokens on the LIVEC and KonIQ datasets. As shown in Fig. 4, the similarities range from 0.72 to 0.81, lower than the 0.96 and 0.94 found in DeiT-T and DeiT-S [42]. Tokens near the center show higher similarity,

aligning with the HVS’s focus on salient regions [62]. This implies that DIS tokens capture localized distortions in a way that simulates human perception, and when combined with CLS tokens, it improves both global and local distortion awareness. To further explore the DIS tokens, we visualize the attention maps of DEIQT and our SHDIQA, as shown in Fig. 5. By incorporating four additional distortion tokens, the proposed model enhances its ability to detect distortions across different regions of the image. Through the interaction between distortion tokens and CLS tokens in the adaptation head, the model adaptively focuses on the most relevant areas for image quality, while also considering the overall content. This results in more accurate quality predictions.

5. Conclusion

In this paper, we propose SHDIQA, a novel BIQA approach that leverages spatial-heterogeneous distortion-aware strategies. Our Block-Wise Degradation Modelling generates locally distorted images, exposing the model to diverse distortion patterns. The Block-wise Aggregation and Filtering mechanism enhances the ViT model’s sensitivity to local distortions, improving quality assessments through focused attention and feature filtering. Additionally, our Distortion-Guided Knowledge Distillation combines CKD and AKD to improve distortion discrimination and relational learning. Experimental results show SHDIQA’s superior generalization and robustness across various datasets and distortion types, confirming its effectiveness in real-world BIQA tasks.

Acknowledgements. This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. U23A20383, 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. 2021J06003, No.2022J06001).

References

- [1] Mark R Banham and Aggelos K Katsaggelos. Digital image restoration. *IEEE signal processing magazine*, 14(2):24–41, 1997. 1
- [2] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024. 3
- [3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4
- [5] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 4
- [6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1
- [7] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 5
- [8] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 3
- [9] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 5
- [10] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1220–1230, 2022. 2, 3, 5
- [11] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 3
- [12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3
- [13] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 1, 5
- [14] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 2, 3, 5
- [15] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006, 2010. 2, 5
- [16] Aobo Li, Jinjian Wu, Yongxu Liu, and Leida Li. Bridging the synthetic-to-authentic gap: Distortion-guided unsupervised domain adaptation for blind image quality assessment. *arXiv preprint arXiv:2405.04167*, 2024. 2
- [17] Leida Li, Tianshu Song, Jinjian Wu, Weisheng Dong, Jian-sheng Qian, and Guangming Shi. Blind image quality index for authentic distortions with local and global deep feature aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8512–8523, 2021. 2, 3
- [18] Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6501–6510, 2023. 4
- [19] Xudong Li, Timin Gao, Runze Hu, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Jingyuan Zheng, Yunhang Shen, Ke Li, Yutao Liu, et al. Adaptive feature selection for no-reference image quality assessment by mitigating semantic noise sensitivity. In *International Conference on Machine Learning*, pages 27808–27821. PMLR, 2024. 4, 5
- [20] Xudong Li, Runze Hu, Jingyuan Zheng, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Ke Li, Yunhang Shen, Yutao Liu, Pingyang Dai, et al. Integrating global context contrast and local sensitivity for blind image quality assessment. In *International Conference on Machine Learning*, pages 27920–27941. PMLR, 2024. 3
- [21] Xudong Li, Jingyuan Zheng, Runze Hu, Yan Zhang, Ke Li, Yunhang Shen, Xiawu Zheng, Yutao Liu, Shengchuan Zhang, Pingyang Dai, et al. Feature denoising diffusion model for blind image quality assessment. *arXiv preprint arXiv:2401.11949*, 2024. 3
- [22] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 5
- [23] Hao Liu and Pieter Abbeel. Blockwise parallel transformers for large context models. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [24] Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group mad competition-a new methodology to compare objective image quality models. In *Proceedings of the IEEE Conference on*

- Computer Vision and Pattern Recognition*, pages 1664–1673, 2016. 7
- [25] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017. 2
- [26] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017. 2
- [27] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 5
- [28] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 93–110. Springer, 2020. 3
- [29] Guanyi Qin, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. Data-efficient image quality assessment with attention-panel decoder. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023. 3, 4, 5, 6, 8
- [30] Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen-tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. *arXiv preprint arXiv:1911.02972*, 2019. 3
- [31] Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-iqa: Un-supervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5846–5855, 2023. 2, 4, 5, 6
- [32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8
- [33] Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, and Muhammad Zeshan Afzal. Sparse semi-detr: Sparse learnable queries for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5850, 2024. 4
- [34] Hamid R Sheikh, Alan C Bovik, and Lawrence Cormack. No-reference quality assessment using natural scene statistics: Jpeg2000. *IEEE Transactions on image processing*, 14(11):1918–1927, 2005. 2
- [35] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 5
- [36] Chenyang Shi and Yandan Lin. Full reference image quality assessment based on visual salience with color appearance and gradient similarity. *IEEE Access*, 8:97310–97320, 2020. 1
- [37] Tianshu Song, Leida Li, Deqiang Cheng, Pengfei Chen, and Jinjian Wu. Active learning-based sample selection for label-efficient blind image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1
- [38] Suhas Srinath, Shankhanil Mitra, Shika Rao, and Rajiv Soundararajan. Learning generalizable perceptual representations for data-efficient no-reference image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 22–31, 2024. 1
- [39] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 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*, pages 3667–3676, 2020. 5
- [40] Simeng Sun, Tao Yu, Jiahua Xu, Wei Zhou, and Zhibo Chen. Graphiqa: Learning distortion graph representations for blind image quality assessment. *IEEE Transactions on Multimedia*, 25:2912–2925, 2022. 2
- [41] Dacheng Tao, Xuelong Li, Wen Lu, and Xinbo Gao. Reduced-reference iqa in contourlet domain. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(6):1623–1627, 2009. 1
- [42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 8
- [43] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 6
- [44] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019. 5
- [45] Zhou Wang and Eero Simoncelli. Local phase coherence and the perception of blur. *Advances in neural information processing systems*, 16, 2003. 2
- [46] Zhou Wang, Hamid R Sheikh, and Alan C Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *Proceedings. International conference on image processing*, pages I–I. IEEE, 2002. 2
- [47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1
- [48] Zhihua Wang, Zhi-Ri Tang, Jianguo Zhang, and Yuming Fang. Toward a blind image quality evaluator in the wild by learning beyond human opinion scores. *Pattern Recognition*, 137:109296, 2023. 1
- [49] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 1
- [50] Kangmin Xu, Liang Liao, Jing Xiao, Chaofeng Chen, Haoning Wu, Qiong Yan, and Weisi Lin. Local distortion aware efficient transformer adaptation for image quality assessment. *arXiv preprint arXiv:2308.12001*, 2023. 2, 3, 5, 6
- [51] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa:

- Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 1
- [52] Guanghao Yin, Wei Wang, Zehuan Yuan, Chuchu Han, Wei Ji, Shouqian Sun, and Changhu Wang. Content-variant reference image quality assessment via knowledge distillation. *arXiv preprint arXiv:2202.13123*, 2022. 3
- [53] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 5, 6
- [54] Junyong You and Jari Korhonen. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1389–1393. IEEE, 2021. 3, 5
- [55] Guanghui Yue, Di Cheng, Leida Li, Tianwei Zhou, Hantao Liu, and Tianfu Wang. Semi-supervised authentically distorted image quality assessment with consistency-preserving dual-branch convolutional neural network. *IEEE Transactions on Multimedia*, 25:6499–6511, 2022. 3
- [56] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 2, 5
- [57] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 2, 5
- [58] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22302–22313, 2023. 2, 4, 5, 6
- [59] Heliang Zheng, Huan Yang, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning conditional knowledge distillation for degraded-reference image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10242–10251, 2021. 3
- [60] Yu Zhou, Leida Li, Shiqi Wang, Jinjian Wu, Yuming Fang, and Xinbo Gao. No-reference quality assessment for view synthesis using dog-based edge statistics and texture naturalness. *IEEE Transactions on Image Processing*, 28(9):4566–4579, 2019. 1
- [61] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Metaiq: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14143–14152, 2020. 2, 5
- [62] Mengmeng Zhu, Guanqun Hou, Xinjia Chen, Jiaying Xie, Haixian Lu, and Jun Che. Saliency-guided transformer network combined with local embedding for no-reference image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1953–1962, 2021. 8
- [63] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 3