Feature Denoising Diffusion Model for Blind Image Quality Assessment

Xudong Li[1], Yan Zhang[1*], Yunhang Shen[2], Ke Li[2], Runze Hu[3], Xiawu Zheng[1], Sicheng Zhao[4]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China
[2]Tencent Youtu Lab, Shanghai, China
[3]School of Information and Electronics, Beijing Institute of Technology, Beijing, China
[4]BNRist, Tsinghua University, Beijing, China
{lxd761050753, bzhy986, shenyunhang01, tristanli.sh, hrzlpk2015}@gmail.com,
zhengxiawu@xmu.edu.cn, schzhao@tsinghua.edu.cn

Abstract

Blind Image Quality Assessment (BIQA) aims to evaluate image quality in line with human perception, without reference benchmarks. Currently, deep learning BIQA methods typically depend on using features from high-level tasks for transfer learning. However, the inherent differences between BIQA and these high-level tasks inevitably introduce noise into the quality-aware features. In this paper, we take an initial step toward exploring the diffusion model for feature denoising in BIQA, namely Perceptual Feature Diffusion for IQA (PFD-IQA), which aims to remove noise from quality-aware features. Specifically, 1) we propose a Perceptual Prior Discovery and Aggregation module to establish two auxiliary tasks to discover potential low-level features in images that are used to aggregate perceptual textual prompt conditions for the diffusion model. 2) we propose a Perceptual Conditional Feature Refinement strategy, which matches noisy features to predefined denoising trajectories and then performs exact feature denoising based on textual prompt conditions. By incorporating a lightweight denoiser and requiring only a few feature denoising steps (e.g., just five iterations), our PFD-IQA framework achieves superior performance across eight standard BIQA datasets, validating its effectiveness.

1 Introduction

Image Quality Assessment (IQA) methods aim to match the human perception of image distortions (Wang et al. 2004). As a crucial low-level visual task, reliable IQA models are important for image-driven applications. Objective IQA includes Full-Reference IQA (Shi and Lin 2020), Reduced-Reference IQA (Tao et al. 2009), and Blind IQA (BIQA). As reference images are often unavailable, BIQA gains attention for tasks like image restoration (Banham and Katsaggelos 1997) and super-resolution (Dong et al. 2015).

Deep neural network-based BIQA models have made notable advancements (Bosse et al. 2017; Li et al. 2024b). These methods often utilize a straightforward training strategy, which includes pre-training on the large-scale ImageNet domain (Deng et al. 2009) and then fine-tuning on the IQA task to extract quality-aware features (Qin et al. 2023). Ideally, these features should capture both low-level distortions and high-level content (Zhao et al. 2023a). However, incorporating synthetic distortions as data augmentation (Hendrycks and Dietterich 2019) during pre-training for classification tasks may reduce the model's sensitivity to image degradation (Zhang et al. 2023; Hendrycks and Dietterich 2019), causing the model to overemphasize high-level semantics while neglecting critical low-level details (Li et al. 2024a). For example, as shown in Fig. 1, the baseline model emphasizes high-level information, such as the "human" in the red box, but overlooks low-level details like blur and geometric distortion in the yellow box, resulting in inaccurate quality predictions. Thus, not all pre-trained semantic features are beneficial; some may introduce quality-irrelevant noise, requiring careful filtering to improve BIQA.

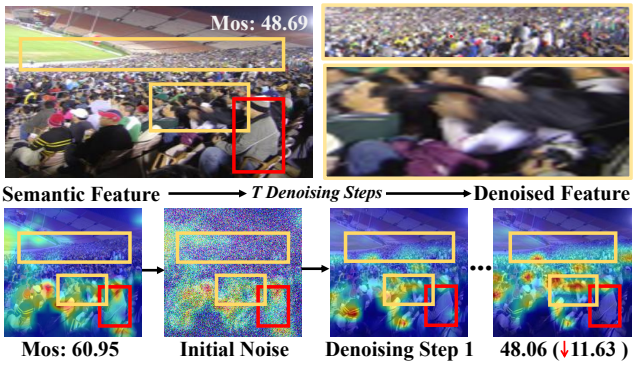

Figure 1: **Top Image:** the sample image. **Bottom Images:** Before and after diffusion denoising, the feature map significantly refines, effectively pinpointing areas with visible image quality degradation. The initial semantic focus is on "human", but after denoising, attention notably shifts to the fuzzy region (the yellow region with the crowd and arms), aligning more closely with Mean Opinion Scores (MOS).

Our motivation is illustrated in Fig. 1. We propose treating pre-trained semantic features as noise-corrupted quality-aware features and removing this noise with the help of diffusion models (DMs) (Rombach et al. 2022; Clark and Jaini 2024) which are particularly effective in recovering target data distribution from noisy input (Ye and Xu 2024). This

*Corresponding author.

effectiveness is attributed to the bidirectional nature of DMs: they add mixed Gaussian noise to simulate target noise distribution (Bishop 2006) in the diffusion process and learn to remove this noise via the denosier in the denoising process. While previous studies (Zhao et al. 2023b; Ye and Xu 2024) have demonstrated the efficacy of DMs in feature denoising for high-level tasks, their application to low-level BIQA remains largely unexplored. This gap can be attributed to two key challenges: **1)** The lack of ground truth distributions for quality-aware features complicates the definition of effective denoising trajectories. **2)** The inherent randomness of the diffusion process hinders detail features preservation (Liu et al. 2024), making precise feature denoising challenging.

In this work, we propose the Perceptual Feature Diffusion IQA (PFD-IQA) framework, which leverages a diffusion model to jointly perform the diffusion and denoising processes to discover potential noisy distributions of the quality-aware maps and learns to rectify them iteratively. By further incorporating low-level priors, the denoising process enhances sensitivity to both quality and distortion cues, enabling the effective refinement of quality-aware features.

Specifically, to address Challenge **1)**, we first define denoising trajectories using pre-trained teacher quality-aware features as target distributions and establish a latent DDIM-based (Rombach et al. 2022) denoiser to learn denoise noisy quality-aware features. We then introduce a Perceptual Conditional Feature Refinement Strategy, where student features are treated as noise-corrupted versions of the teacher's quality-aware features. The denoiser iteratively estimates and removes noise from these student representations, gradually restoring clean quality-aware features. To ensure accurate alignment with the initial reverse denoising steps, we incorporate a Noise-Level Alignment mechanism. This mechanism dynamically evaluates the noise in each student feature and injects appropriate Gaussian noise, ensuring that student features begin from the correct initial noise level during the denoising process. As a result, the reverse denoising process becomes more precise and efficient.

To address the Challenge **2)**, we draw inspiration from psychological insights (Hou et al. 2014), which suggest that humans perceive image quality in multiple dimensions (e.g., "fair quality" as "not bad" + "imperfect," or "mixed noise" as "impulse noise" + "white noise"). Building on this, we propose a Perceptual Prior Discovery and Aggregation mechanism, which identifies latent perceptual priors in images and aggregates them into distortion-aware and quality-aware textual prompts. During the reverse denoising process, the denoiser relies on these low-level prompts to reconstruct quality-aware features, thereby enhancing its ability to perceive and preserve distortion and quality information.

In summary, the contribution of this paper is threefold:

- This paper is the first to explore transforming BIQA challenges into a diffusion problem for feature denoising. We introduce a novel PFD-IQA approach that effectively filters out quality-irrelevant information from features.
- We propose a Perceptual Prior Discovery and Aggregation method that identifies and combines latent perceptual priors into perceptual textual prompts, guiding the denoising process to better refine quality-aware features.
- We propose a novel Perceptual Conditional Feature Refinement Strategy for BIQA. Specifically, we pre-define denoising trajectories by teacher pseudo-features. Using an adaptive Noise-Level Alignment module, we align student noise features with these trajectories and execute precise feature denoising based on prompt conditions.

2 Related Work

2.1 BIQA with Deep Learning

Early Blind Image Quality Assessment (BIQA) (Liu, van de Weijer, and Bagdanov 2017; Li et al. 2009) methods often relied on convolutional neural network (CNN) architectures. These approaches typically treated IQA as the downstream task of object recognition, following a pre-training and fine-tuning pipeline (Zhu et al. 2022). Such a strategy was effective because the pre-trained features partially overlapped with quality-aware features (Su et al. 2020). In recent years, Vision Transformer (ViT) (Dosovitskiy et al. 2021)-based BIQA methods have gained traction due to their ability to model non-local features. These methods generally adopt either hybrid (Xu et al. 2024; Golestaneh, Dadsetan, and Kitani 2022) or pure (Qin et al. 2023) transformer architectures, often relying on the CLS token for quality assessment. However, as the CLS token was originally designed for high-level content recognition rather than low-level IQA tasks, features inherited from classification pre-training may be noisy and suboptimal for IQA. Recently, some IQA methods (Fu et al. 2024; De, Mitra, and Soundararajan 2024) have leveraged the prior from the pre-trained Stable Diffusion model to improve BIQA performance and generalization. In contrast, our approach redefines the optimization of pre-trained semantic features as the feature denoising problem. Using a lightweight denoiser and a few sampling steps (e.g., 5 iterations), it achieves exceptional performance.

2.2 Probabilistic Diffusion Models

Diffusion models (Rombach et al. 2022) are probabilistic generative models that add noise to data and learn to reverse this process by predicting and removing the noise. Given an image z_0, Gaussian noise is added in the forward process as:

$$q(z_t|z_0) := \mathcal{N}(z_t|\sqrt{\bar{\alpha}_t}z_0, (1-\sqrt{\bar{\alpha}_t})I),$$

where z_t is the noisy data at time step $t \in \{0, 1, ..., T\}$, and $\bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_s)$ determines the noise variance schedule β (Ho, Jain, and Abbeel 2020; Shao et al. 2024). This process allows z_t to be expressed as:

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t,$$

where $\epsilon_t \sim \mathcal{N}(0, I)$. A neural network $\epsilon_\theta(z_t, t)$ is trained to predict ϵ_t by minimizing the L2 loss:

$$\mathcal{L}_{dif} := ||\epsilon_t - \epsilon_\theta(z_t, t)||_2^2.$$

During inference, the diffusion model reconstructs original sample z_0 from noisy data z_t iteratively as:

$$p_\theta(z_{t-1}|z_t) := \mathcal{N}(z_{t-1}; \epsilon_\theta(z_t, t), \sigma_t^2 I),$$

where σ_t^2 is the transition variance. Using DDIM (Song, Meng, and Ermon 2020), the denoising process accelerates, sampling $z_T \to z_{T-\Delta} \to ... \to z_0$, where Δ is the step size.

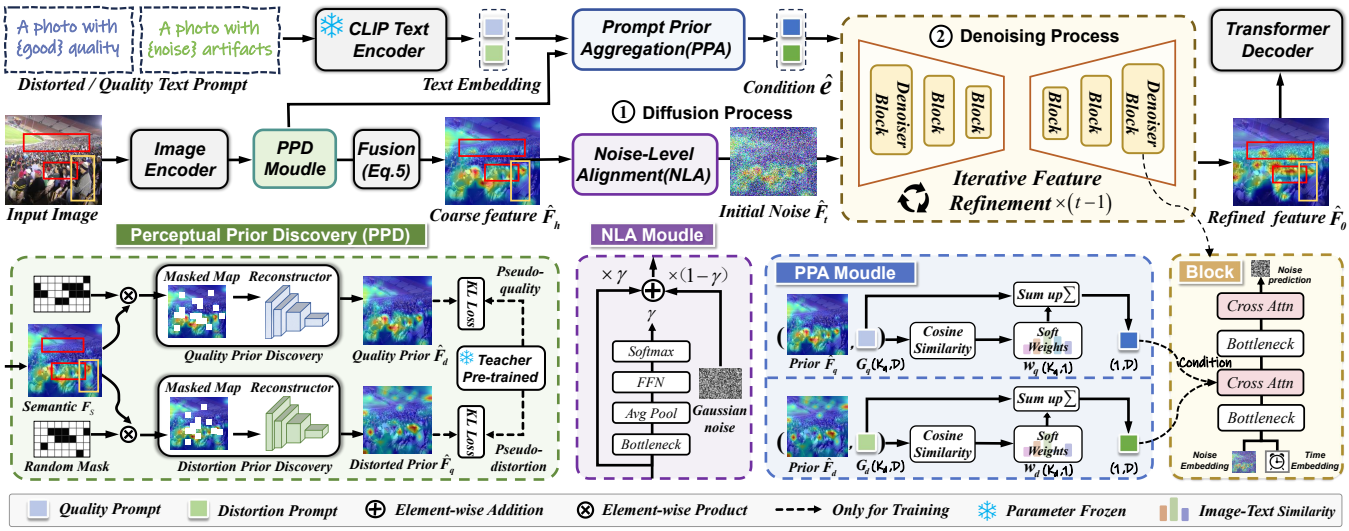

Figure 2: The image encoder initially extracts semantic features $\boldsymbol{F}_s$ and enhances them through feature fusion with distortion and quality priors from the PPD module (Sec. 3.2). These enhanced features $\hat{\boldsymbol{F}}_h$ are then aligned with specified initial noise $\hat{\boldsymbol{F}}_t$ to match the predefined denoising trajectory (Sec. 3.3 and Fig.3) using the NLA module (Sec.3.4). Concurrently, the PPA module (Sec.3.2) aggregates priors from the PPD module into prompt conditions $\hat{\boldsymbol{e}}$, guiding the lightweight denoiser to refine noisy features (Sec. 3.4). Finally, the refined features $\hat{\boldsymbol{F}}_0$ are fed into a transformer decoder to predict the quality score.

3 Methodology

3.1 Overview

Symbol Notations. Training data is $\mathcal{D} = \{\boldsymbol{x}, \boldsymbol{y}_g, \boldsymbol{p}_d, \boldsymbol{p}_q\}$, where $\boldsymbol{x}$ is the labeled image, $\boldsymbol{y}_g$ is the ground truth, and $\boldsymbol{p}_d$ and $\boldsymbol{p}_q$ are the soft labels for distortion type and quality level generated by teacher. Image and text features are $\boldsymbol{F}$ and $\boldsymbol{G}$.

Denoising Trajectories. As shown in Fig. 3, we use the teacher's quality-aware features $\boldsymbol{F}^{tea}$ to define denoising trajectories by performing joint diffusion and denoising processes, where diffusion captures latent noise distribution of the quality-aware features, and the denoising process trains the denoiser to predict these noises, enabling it to remove quality-irrelevant noise from the student features effectively.

Overview of Student. As shown in Fig 2, our PFD-IQA consists of two steps: (i) First, an image encoder extracts semantic features $\boldsymbol{F}_s$, which are then combined with the discovered quality and distortion priors by the proposed PPD module to generate a coarse quality-aware feature map $\hat{\boldsymbol{F}}_h$. (ii) Second, the NLA module diffuses the coarse features to the appropriate initial step $\hat{\boldsymbol{F}}_t$ in the diffusion process, ensuring more precise denoising. In the denoising process, the Perceptual Conditional Feature Refinement strategy refines these noisy features by using the denoiser to predict and remove noise, conditioned on text prompts from the PPA module. Finally, the refined features $\hat{\boldsymbol{F}}_0$ are fed into a transformer decoder (Qin et al. 2023) to predict the quality score.

Overview of Teacher. We use LIQE (Zhang et al. 2023) as the teacher for its ability to provide pseudo-labels for distortion types and quality levels, supervising the PPD module. It also supplies quality-aware pseudo-features to predefine the denoising trajectory (Sec. 3.3). LIQE's strong performance (Tab. 1) ensures the reliability of these pseudo-labels.

3.2 Perceptual Prior Discovery and Aggregation

Perceptual Prior Discovery (PPD). We introduce auxiliary tasks for classifying distortion types and quality levels to help the model learn different low-level priors, which contributes to robust IQA (Zhang et al. 2023). Inspired by the ability of masking mechanisms (Yang et al. 2022; He et al. 2022) to enhance local image perception, we further adopt a masking-and-reconstruction strategy. Specifically, a channel-wise random mask M_c is applied to the semantic features $\boldsymbol{F}_s$ from the ViT encoder, generating masked local features $\boldsymbol{F}_m$. Then, two lightweight reconstructors, $\mathcal{R}(\cdot)$, are trained to recover distorted priors $\hat{\boldsymbol{F}}_d$ and quality priors $\hat{\boldsymbol{F}}_q$ conditioned on local cues $\boldsymbol{F}_m$, enhancing the model's fine-grained perception. This process is formally defined as:

$$M_c = \begin{cases} 0, & \text{if } R_c < \beta, \\ 1, & \text{otherwise}, \end{cases} \quad \boldsymbol{F}_m = f_{\text{align}}(\boldsymbol{F}_s) \cdot M_c, \tag{1}$$
$$\hat{\boldsymbol{F}}_j = \mathcal{R}(\boldsymbol{F}_m) = W_{l2,j} \cdot \text{ReLU}\left(W_{l1,j}(\boldsymbol{F}_m)\right).$$

where R_c is a random number, c is the channel number, and β is the mask ratio (set to 0.45). The adaptation layer f_{align} is a 1×1 convolution. $j \in \{d, q\}$ denotes distorted priors $\hat{\boldsymbol{F}}_d$ or quality priors $\hat{\boldsymbol{F}}_q$. Each $\mathcal{R}(\cdot)$ comprises two 1×1 convolution layers (W_{l1}, W_{l2}) and a Batch Normalization (BN) layer.

Logits-Based Auxiliary Supervision. We use the pseudo-labels provided by the LIQE teacher (Zhang et al. 2023) to supervise the auxiliary tasks of quality and distortion classification. Following LIQE, images are categorized into $K_q = 5$ quality levels and $K_d = 11$ distortion types. Text prompts are defined as $\mathcal{T}_d = \{$"a photo with $\{d_i\}$ artifact."$\}$ and $\mathcal{T}_q = \{$"a photo with $\{q_i\}$ quality."$\}$, where d_i and q_i represent the i-th distortion type and quality level. For a

given prior $\hat{\boldsymbol{F}}_j$, we calculate its cosine similarity with each text embedding $\boldsymbol{G}_j = \mathcal{E}(\mathcal{T}_j)$, resulting in the logits $\hat{\boldsymbol{p}}_j$. The reconstructor $\mathcal{R}(\cdot)$ is then supervised using logits-based pseudo-labels $\boldsymbol{p}_d$ and $\boldsymbol{p}_q$ from the teacher, with Kullback-Leibler (KL) divergence loss, which is expressed as:

$$\hat{\boldsymbol{p}}_j = \frac{\hat{\boldsymbol{F}}_j \cdot \boldsymbol{G}_j^{\mathrm{T}}}{\|\hat{\boldsymbol{F}}_j\|_2 \|\boldsymbol{G}_j\|_2}, \quad j \in \{d, q\},$$
$$\mathcal{L}_{kl} = \sum_{j \in \{d,q\}} \mathcal{L}_{kl}^{j}\left(\boldsymbol{p}_j, \hat{\boldsymbol{p}}_j\right) = \sum_{j \in \{d,q\}} \boldsymbol{p}_j \log \frac{\boldsymbol{p}_j}{\hat{\boldsymbol{p}}_j}. \tag{2}$$

Prompt Prior Aggregation (PPA). Psychological research suggests that humans prefer using multidimensional natural language for qualitative assessments over quantitative ones (Hou et al. 2014). In practice, images are often described with terms like "fair" or "imperfect," conveying similar quality judgments. Furthermore, authentic images typically contain mixed distortions that predefined hard text descriptions cannot accurately capture. Inspired by this, we developed a soft-weighted aggregation method that uses natural language prompts to adaptively represent perceived image quality across multiple perceptual dimensions. Specifically, we calculate the logits for distortion and quality levels and then derive the corresponding weights w_j^i, where $j \in \{d, q\}$:

$$\boldsymbol{w}_{\boldsymbol{j}}^{\boldsymbol{i}} = \frac{\exp(\hat{\boldsymbol{p}}_j^i)}{\sum_{i=1}^{K_j} \exp(\hat{\boldsymbol{p}}_j^i)} \in (0, 1), \tag{3}$$

where $\hat{\boldsymbol{p}}_j^i$ is the i-th element of cosine similarity $\hat{\boldsymbol{p}}_j$. Next, for i-th element of prompt G_j^i, we obtain the adaptive perceptual textual prompt $\hat{\boldsymbol{e}}$ via the following weighted aggregation:

$$\hat{\boldsymbol{e}} = \frac{1}{2}\Big(\sum_{i=1}^{K_d} \boldsymbol{w}_d^i \boldsymbol{G}_d^i + \sum_{i=1}^{K_q} \boldsymbol{w}_q^i \boldsymbol{G}_q^i\Big). \tag{4}$$

Notably, $\hat{\boldsymbol{e}}$ effectively represents the complex mixture of distortions in authentic images as soft weights, offering finer-grained guidance for the subsequent feature refinement.

Perceptual Prior Fusion (PPF). To fuse the discovered priors $\hat{\boldsymbol{F}}_d$ and $\hat{\boldsymbol{F}}_q$ into a coarse quality-aware map, we apply a two-dimensional scaling modulation to the semantic feature $\boldsymbol{F}_s$, followed by two convolutional transformations. This process adjusts $\boldsymbol{F}_s$ using scaling and shifting parameters derived from additive features $\hat{\boldsymbol{F}}_{dq} = \hat{\boldsymbol{F}}_d + \hat{\boldsymbol{F}}_q$, producing the coarse quality map $\hat{\boldsymbol{F}}_h$ for better granularity fusion:

$$\hat{\boldsymbol{F}}_h = (\mathrm{conv}(\hat{\boldsymbol{F}}_{dq}) \times \mathrm{norm}(\boldsymbol{F}_s) + \mathrm{conv}(\hat{\boldsymbol{F}}_{dq})) + \boldsymbol{F}_s. \tag{5}$$

3.3 Predefined Denoising Trajectories

Lightweight Denoiser Architecture. Given the high dimensionality of ViT, performing the denoising process on features during training requires numerous iterations, leading to a substantial computational load. To mitigate this issue, we adopt a lightweight diffusion model (DM) $\epsilon_\theta(\cdot)$ as an alternative to the U-Net architecture, as shown in Fig. 3. This model consists of two bottleneck blocks from ResNet (He et al. 2016) and a 1×1 convolution. Additionally, a cross-attention layer (Rombach et al. 2022) is added after each bottleneck to aggregate textual prompt and image features.

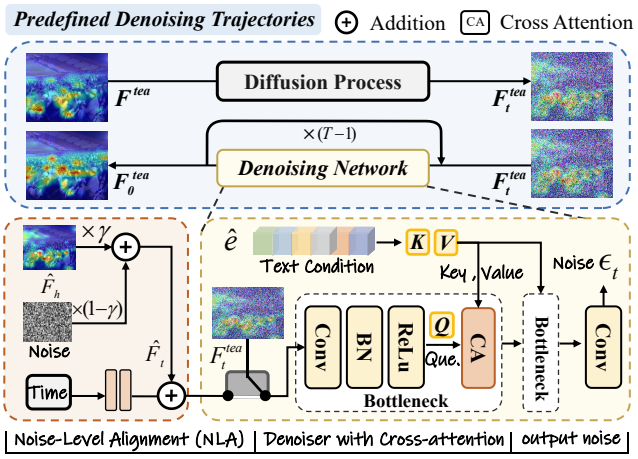

Figure 3: Overview of proposed predefined denoising trajectory, which begins with a teacher pseudo-feature for forward diffusion. In reverse denoising, image and text features are fused, and a denoiser is trained to predict the added noise accurately. For the student, the NLA module determines the initial time step for reverse denoising, which is then fed into the lightweight denoiser to eliminate noise iteratively.

Predefined Denoising Trajectories. We aim to iteratively optimize the noisy feature to attain quality-aware features. This process can be conceptualized as an approximation of the inverse feature denoising procedure. However, the features representing the ground truth are often unknown. Therefore, as depicted in Fig. 3, we introduce features $\boldsymbol{F}^{tea}$ extracted by a pre-trained teacher's encoder (Zhang et al. 2023) as pseudo-ground truth to pre-construct a denoising trajectory of quality-aware features. For the forward diffusion process, $\boldsymbol{F}_t^{tea}$ is a linear combination of the initial teacher feature $\boldsymbol{F}^{tea}$ and the noise variable $\epsilon_t \in \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$:

$$\boldsymbol{F}_t^{tea} = \sqrt{\bar{\alpha}_t}\boldsymbol{F}^{tea} + \sqrt{1 - \bar{\alpha}_t}\epsilon_t. \tag{6}$$

The parameter $\bar{\alpha}_t = \prod_{s=0}^{t}(1 - \beta_s)$ enables direct sampling of $\boldsymbol{F}_t^{tea}$ at any time step using a noise variance schedule β in DDIM. In the denoising process, a lightweight denoiser $\epsilon_\theta(\cdot)$, conditioned on textual prompt $\hat{\boldsymbol{e}}^{tea}$ (from Eq. 4), is trained to predict noise $\boldsymbol{\epsilon}_t$ by minimizing the ℓ_2 loss:

$$\mathcal{L}_{ldm} = \|\epsilon_t - \epsilon_\theta(\boldsymbol{F}_t^{tea}, \hat{\boldsymbol{e}}^{tea}, t)\|_2^2, \tag{7}$$

Once trained, the denoiser can iteratively estimate the noise present in the input features relative to the quality-aware features and remove it using the sampling formula (Eq. 9).

3.4 Perceptual Conditional Feature Refinement

Feature refinement is achieved through an NLA module and a feature denoiser. The NLA module aligns features with the teacher's predefined initial noise, while the denoiser iteratively removes noise to produce quality-aware features.

Noise-Level Alignment (NLA). We treat the feature representations extracted by students as noisy versions of the teacher's quality-aware features. However, the noise level, which represents the discrepancy between the teacher's

quality-aware features and student features, is unknown, making the precise alignment of the denoising trajectory complex and potentially leading to suboptimal feature denoising. As a result, identifying the optimal initial time step to initiate the diffusion process presents a challenging task. Inspired by previous work (Huang et al. 2023), we introduce an NLA module to match the noise level of student features with the initial noise level predefined by the teacher. As illustrated in Fig. 2, a simple convolutional layer learns a weight γ, which combines the coarse quality-aware feature $\hat{\boldsymbol{F}}_h$ with Gaussian noise to produce initial noise:

$$\hat{\boldsymbol{F}}_t = \gamma \odot \hat{\boldsymbol{F}}_h + (1-\gamma) \odot \mathcal{N}(0, I), \tag{8}$$

The noise $\hat{\boldsymbol{F}}_t$ matches the teacher's feature $\boldsymbol{F}_t^{tea}$ (Eq. 6) at the initial noise level for time step t, ensuring proper alignment of the denoising trajectory. This alignment helps the DM effectively predict and remove quality-irrelevant noise.

Iterative Feature Refinement. After obtaining the initial noise $\hat{\boldsymbol{F}}_t$, it is fed into the denoiser $\epsilon_\theta(\cdot)$. Conditioned on the textual prompts $\hat{\boldsymbol{e}}$, the denoiser iteratively predicts and removes quality-irrelevant noise, progressively refining the noisy features into the final quality-aware representation $\hat{\boldsymbol{F}}_0$:

$$p_\theta\left(\hat{\boldsymbol{F}}_{t-1} \mid \hat{\boldsymbol{F}}_t\right) := \mathcal{N}\left(\hat{\boldsymbol{F}}_{t-1}; \epsilon_\theta\left(\hat{\boldsymbol{F}}_t, \hat{\boldsymbol{e}}, t\right), \sigma_t^2 \boldsymbol{I}\right). \tag{9}$$

In practice, we find that the proposed lightweight denoiser effectively removes noise in fewer than 5 sampling iterations, achieving a speed over 200 times faster than DDPM.

Overall Loss. We use the features $\boldsymbol{F}^{tea}$ from the pre-trained teacher's pseudo-labels, to supervise the denoising process via ℓ_2 loss, ensuring stable feature refinement.

$$\mathcal{L}_{fea} = \|\hat{\boldsymbol{F}}_0 - \boldsymbol{F}^{tea}\|_2^2. \tag{10}$$

The overall loss during training is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{kl} + \lambda_2 \mathcal{L}_{ldm} + \lambda_3 \mathcal{L}_{fea} + \|\hat{\boldsymbol{y}} - \boldsymbol{y}_g\|_1. \tag{11}$$

Here, $\hat{y}$ denotes the predicted score of quality-aware feature $\hat{\boldsymbol{F}}_0$ by the transformer decoder, while $\boldsymbol{y}_g$ denotes the ground truth. The $\|\cdot\|_1$ refers to the ℓ_1 regression loss. In all experiments, we empirically set $\lambda_1 = 0.5$, $\lambda_2 = 1$, and $\lambda_3 = 0.01$. We detail the pipeline of our PFD-IQA in Algorithm 1.

Algorithm 1: Pseudocode for proposed PFD-IQA

Input: Image $\boldsymbol{x}$; label $\boldsymbol{y}_g$; Text $\mathcal{T}_q$ and $\mathcal{T}_d$; diffusion steps T; mode$\in$\{'train', 'infer'\}; student, teacher $\mathcal{N}_s, \mathcal{N}_t$; softmax σ
Output: Predicted quality score $\hat{y}$

$\left(\boldsymbol{p}_q, \boldsymbol{p}_d, \boldsymbol{F}^{tea}, \boldsymbol{G}_q, \boldsymbol{G}_d\right), \boldsymbol{F}_s = \mathcal{N}_t(\boldsymbol{x}, \mathcal{T}_q, \mathcal{T}_d), \mathcal{N}_s(\boldsymbol{x})$
Perceptual Prior Discovery (Sec. 3.2)
Get prior: $\hat{\boldsymbol{F}}_q, \hat{\boldsymbol{F}}_d = \text{Reconstructor}(\text{mask}(\boldsymbol{F}_s))$ Eq. 1
Prompt Prior Aggregation (Sec. 3.2)
Compute logits: $\hat{\boldsymbol{p}}_j = \cos(\hat{\boldsymbol{F}}_j, \mathcal{T}_j), j \in \{d, q\}$
Get prompt: $\hat{\boldsymbol{e}} = \sigma(\hat{\boldsymbol{p}}_d) \cdot \boldsymbol{G}_d + \sigma(\hat{\boldsymbol{p}}_q) \cdot \boldsymbol{G}_q$ Eq. 4
Perceptual Prior Fusion (Sec. 3.2)
Get coarse feature: $\hat{\boldsymbol{F}}_h = \text{Fusion}(\boldsymbol{F}_s, \hat{\boldsymbol{F}}_q + \hat{\boldsymbol{F}}_d)$ Eq. 5
Diffusion via Noise-Level Alignment (Sec. 3.4)
Get gaussian noise ratio: $\gamma = \text{NLA}(\hat{\boldsymbol{F}}_h)$, $\epsilon_t \sim \mathcal{N}(0, I)$
Get initial noise: $\hat{\boldsymbol{F}}_t = \gamma \hat{\boldsymbol{F}}_h + (1-\gamma)\epsilon_t$ Eq. 8
# Denoising via Lightweight Denoiser (Sec. 3.4)
for $t = T$ to 1 **do**
  Get refined feature: $\hat{\boldsymbol{F}}_{t-1} = \epsilon_\theta(\hat{\boldsymbol{F}}_t, t, \hat{\boldsymbol{e}}_{stu})$ Eq. 9
end for
Get quality score: $\hat{\boldsymbol{y}} = \text{Decoder}(\hat{\boldsymbol{F}}_0)$
if mode = 'train' **then**
  # Logits-Based Auxiliary Supervision
  Train $\mathcal{R}(\cdot)$: $\mathcal{L}_{kl} = \mathcal{L}_{kl}^q(\boldsymbol{p}_q, \hat{\boldsymbol{p}}_q) + \mathcal{L}_{kl}^d(\boldsymbol{p}_d, \hat{\boldsymbol{p}}_d)$ Eq. 2
  # Denoiser Training and Trajectory Definition
  Get prompt: $\hat{\boldsymbol{e}}_{tea} = \sigma(\boldsymbol{p}_d) \cdot G_d + \sigma(\boldsymbol{p}_q) \cdot G_q$ Eq. 4
  Add noise: $\boldsymbol{F}_t^{tea} = \sqrt{\bar{\alpha}_t}\boldsymbol{F}^{tea} + \sqrt{1-\bar{\alpha}_t}\epsilon_t$ Eq. 6
  for $t = T$ to 1 **do**
    Train $\epsilon_\theta(\cdot)$: $\mathcal{L}_{ldm} = \ell_2(\epsilon_t, \epsilon_\theta(\boldsymbol{F}_t^{tea}, t, \hat{\boldsymbol{e}}_{tea}))$ Eq. 7
  end for
  Improved stability: $\mathcal{L}_{fea} = \ell_2(\hat{\boldsymbol{F}}_0, \boldsymbol{F}^{tea})$ Eq. 10
  return $\lambda_1\mathcal{L}_{kl} + \lambda_2\mathcal{L}_{ldm} + \lambda_3\mathcal{L}_{fea} + \|\hat{\boldsymbol{y}} - \boldsymbol{y}_g\|_1$
end if

4 Experiments

4.1 Datasets and Evaluation Protocols

We evaluate the performance of the proposed PFD-IQA model on eight typical BIQA datasets, including four synthetic—LIVE (Sheikh, Sabir, and Bovik 2006), CSIQ (Larson and Chandler 2010), TID2013 (Ponomarenko et al. 2015), KADID (Lin, Hosu, and Saupe 2019)—and four authentic datasets—LIVEC (Ghadiyaram and Bovik 2015), KONIQ (Hosu et al. 2020), LIVEFB (Ying et al. 2020), SPAQ (Fang et al. 2020). LIVEC contains 1,162 mobile device images, SPAQ has 11,125 photos from 66 smartphones, KonIQ-10k includes 10,073 public images, and LIVEFB, the largest real-world dataset, has 39,810 images. The synthetic datasets involve original images with artificial distortions like noise and Gaussian blur. LIVE and CSIQ have 779 and 866 distorted images, respectively, with five and six distortion types. TID2013 and KADID contain 3,000 and 10,125 distorted images, covering 24 and 25 distortion types. We use Spearman's Rank Correlation Coefficient (SRCC) for prediction monotonicity and Pearson's Linear Correlation Coefficient (PLCC) for accuracy.

4.2 Implementation Details

For the student network, the image encoder is based on ViT-B from DeiT III (Touvron, Cord, and Jégou 2022) with a decoder depth of one, and the parameters of the text encoder are frozen. Our model is trained for 9 epochs with a learning rate of 8×10^{-5}, decaying by a factor of 10 every 3 epochs. The batch size is 16 for LIVEC and 64 for KonIQ. For each dataset, 80% of images are used for training and 20% for testing. This process is repeated 10 times to mitigate bias, and we report the average SRCC and PLCC. For synthetic distortion datasets, training and testing sets are divided by reference images to ensure content independence. For the teacher network, we retrained LIQE (Zhang et al. 2023) based on CLIP-B/16 (Radford et al. 2021) with default hyperparameters. As the performance is close to the original LIQE in Tab. 1, we did not list them separately.

Method	LIVE		CSIQ		TID2013		KADID		LIVEC		KonIQ		LIVEFB		SPAQ	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BRISQUE (Mittal, Moorthy, and Bovik 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
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 and 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
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
TReS (Golestaneh, Dadsetan, and Kitani 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	-	-
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
Re-IQA (Saha, Mishra, and Bovik 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	-	-	0.925	0.918
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
LIQE (Zhang et al. 2023)	0.951	0.970	0.939	0.936	-	-	0.931	0.930	0.910	0.904	0.908	0.919	-	-	-	-
LoDa (Xu et al. 2024)	0.979	0.975	-	-	0.901	0.869	0.936	0.931	0.899	0.876	0.944	**0.932**	**0.679**	**0.578**	0.928	0.925
PFD-IQA (ours)	**0.985**	**0.985**	**0.972**	**0.962**	**0.937**	**0.924**	**0.937**	**0.934**	**0.922**	**0.906**	**0.945**	0.930	0.667	0.572	**0.931**	**0.926**

Table 1: Performance comparison of average SRCC and PLCC, with bold indicating the best and underlines for the second-best.

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
HyperIQA	0.758	0.735	0.772	0.785	0.744	0.926
TReS	0.713	0.740	0.733	0.786	0.761	-
DEIQT	0.733	0.781	0.744	0.794	0.781	0.932
LoDa	0.763	**0.805**	0.745	0.811	-	-
PFD-IQA	**0.775**	0.783	**0.796**	**0.818**	**0.817**	**0.942**

Table 2: SRCC on the cross datasets validation. The best results are highlighted in bold, second-best is underlined.

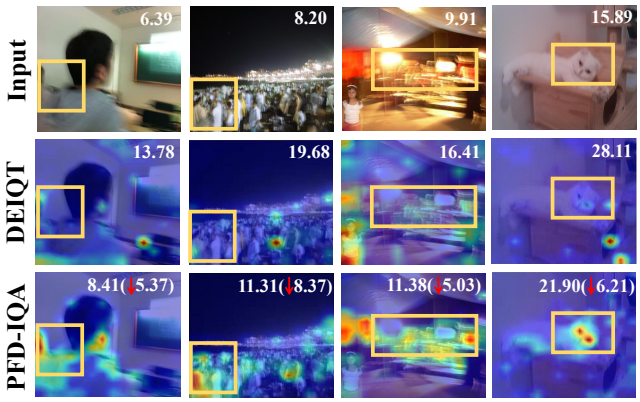

Figure 4: Activation maps using Grad-CAM. Scores in the image represent ground truths. Our PFD-IQA well-performs in distortion areas, resulting in accurate quality prediction.

4.3 Overall Prediction Performance Comparison

For the competing models, we either directly use the publicly available implementations or re-train them on our datasets using the training codes provided by the respective authors. Tab. 1 presents a comparative analysis between the proposed PFD-IQA and 14 state-of-the-art BIQA methods. These include methods that are not pre-trained on high-level tasks, such as Re-IQA (Saha, Mishra, and Bovik 2023) and MUSIQ (Ke et al. 2021), as well as those may incorporate pre-trained semantic noise, such as DEIQT (Qin et al. 2023) and LoDa (Xu et al. 2024). It is observed from these eight datasets that PFD-IQA achieves superior performance over other methods across most of the datasets. Since the images on these eight datasets cover various image content and distortion types, it is very challenging to consistently achieve leading performance on all these datasets. Accordingly, these observations confirm the effectiveness and superiority of PFD-IQA in characterizing image quality.

4.4 Generalization Capability Validation

We further evaluate the generalization ability of PFD-IQA by a cross-dataset validation approach, where the BIQA model is trained on one dataset and then tested on the others without any fine-tuning or parameter adaptation. Tab. 2 reports the experimental results of SRCC averages on the five datasets. As observed, PFD-IQA achieves the best performance on five out of six cross-datasets, achieving clear performance gains on the authentic dataset. This suggests that models like LoDa and DEIQT, which rely on pre-trained ViT, may introduce noise from high-level semantic features, limiting their cross-dataset performance.

4.5 Qualitative Analysis

We use Grad-CAM (Selvaraju et al. 2017) to visualize the feature attention maps, as illustrated in Fig. 4. Our results show that PFD-IQA effectively focuses on areas with significant quality degradation, while DEIQT tends to overemphasize semantic features, sometimes prioritizing irrelevant regions for quality assessment. For example, in the third column, DEIQT erroneously concentrates on the little girl in the bottom-left corner, neglecting the overexposure and blurriness in the center of the image. Moreover, Fig. 4 compares the quality predictions of PFD-IQA and DEIQT, demonstrating that PFD-IQA consistently outperforms DEIQT, particularly for images with moderate distortions.

Index	PDA	PFR	LIVE		CSIQ		TID2013		KADID		LIVEC	
			PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
a)			0.966	0.964	0.952	0.935	0.899	0.888	0.891	0.887	0.881	0.863
b)	✔		0.984	0.981	0.963	0.954	0.927	0.915	0.925	0.925	0.911	0.895
c)		✔	0.983	0.982	0.968	0.959	0.910	0.890	0.918	0.919	0.916	0.897
d)	✔	✔	**0.985**$_{+1.9\%}$	**0.985**$_{+2.1\%}$	**0.972**$_{+2.0\%}$	**0.962**$_{+2.7\%}$	**0.937**$_{+3.8\%}$	**0.924**$_{+3.6\%}$	**0.937**$_{+4.6\%}$	**0.934**$_{+4.7\%}$	**0.922**$_{+4.1\%}$	**0.906**$_{+4.3\%}$

Table 3: Ablation experiments about Perceptual Prior Discovery and Aggregation (PDA) and Perceptual Conditional Feature Refinement (PFR) module on different IQA datasets. Bold entries indicate the best performance.

Index	NLA	PPA	LIVEC		KonIQ		
			PLCC	SRCC	PLCC	SRCC	Std.
a)			0.881	0.863	0.929	0.914	±0.011
b)	✔		0.914	0.896	0.938	0.928	±0.005
c)		✔	0.918	0.900	0.940	0.928	±0.008
d)	✔	✔	**0.922**	**0.906**	**0.945**	**0.930**	**±0.004**

Table 4: Ablation study of the NLA and PPA module on authentic datasets. Bold entries indicate the best result.

4.6 Ablation Study

Perceptual Prior Discovery and Aggregation (PDA). As shown in Tab. 3, integrating the PDA module alone (considering only the PPD and PPF modules in this setup) to capture latent quality information and distortion priors (scenario ***b)***) leads to significant improvements in PLCC (1.1% to 3.4%) and SRCC (1.7% to 3.8%) across different IQA datasets. These results demonstrate the effectiveness of the proposed PDA module in enhancing the model's quality awareness.

Perceptual Conditional Feature Refinement (PFR). As shown in Table 3, integrating the PFR module alone, without the perceptual textual prompts from the PPA module (scenario ***c)***), improves PLCC by 1.1% to 3.5% and SRCC by 0.2% to 3.4% across datasets, demonstrating its effectiveness in feature optimization. Combining the PFR and PDA modules (scenario ***d)***) results in substantial gains in PLCC (1.9% to 4.6%) and SRCC (2.1% to 4.7%), highlighting their synergistic effect in enhancing robustness and accuracy.

NLA and PPA module. As shown in Tab. 4, the PPA module, which leverages perceptual text embedding cues, achieves significant improvements compared to the baseline. However, it exhibits certain stability issues. In contrast, the NLA module markedly reduces the model's standard deviation but limits performance gains without prompt conditions. When integrated, NLA aligns features with predefined denoising modules to mitigate randomness, while PPA refines perceptual features through precise image-text interaction. This joint collaboration between the two modules ultimately contributes to achieving optimal performance.

Number of Sampling Iterations. We employ DDIM (Song, Meng, and Ermon 2020) for sampling. The experiments emphasize how different sampling numbers impact performance. Tab. 5 demonstrates that single-step denoising significantly outperforms the baseline. We find that an iteration of 5 is adequate for effective performance in our approach.

Sampling Number T	LIVE		LIVEC		Avg.
	PLCC	SRCC	PLCC	SRCC	
1	0.982	0.980	0.913	0.890	0.941
3	0.984	0.982	0.918	0.899	0.945
5	**0.985**	**0.985**	**0.922**	**0.906**	**0.950**
10	0.983	0.983	0.921	0.901	0.947

Table 5: Ablation experiments about the number of sampling iterations. Bold entries indicate the best performance.

Condition	PDA Module				PFR Module			
Quality Prior	✗	✗	✔	✔	✗	✗	✔	✔
Distorted Prior	✗	✔	✗	✔	✗	✔	✗	✔
Prior-to-Prompt	✗	✗	✗	✗	✗	✔	✔	✔
SRCC Metric	0.863	0.887	0.888	**0.895**	0.897	0.898	0.900	**0.906**

Table 6: Ablation study on priors for PDA module and denoising conditions for PFR module on the LIVEC dataset.

Selection of Priors and Conditions. As shown in Tab. 6, we assess the impact of different discovered priors and denoising conditions. The results indicate that using the PDA module to fuse either distortion or quality priors improves performance (Columns 2–5). Additionally, incorporating perceptual prompts from the PPA module, which provides a more detailed description of image quality, further boosts performance (Columns 6–9). The most significant improvement occurs when all conditions are combined. These findings suggest that while each condition contributes individually, their combination is key to achieving optimal performance.

5 Conclusion

In conclusion, we propose a novel PFD-IQA framework using diffusion models for noise removal. It tackles key BIQA challenges with two modules: the Perceptual Prior Discovery and Aggregation module for enhanced feature representation and denoising conditions, and the Perceptual Conditional Feature Refinement Strategy for defining denoising trajectories without explicit benchmarks. By combining text prompts with perceptual priors and adaptive noise-level alignment, PFD-IQA refines quality-aware features with precision. Experimental results show that PFD-IQA achieves superior performance with a few sampling steps and a lightweight denoiser, outpacing previous methods.

Acknowledgments

This work was supported by National Science and Technology Major Project (No. 2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. U23A20383, No. U21A20472, 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

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

Bishop, C. M. 2006. Pattern recognition and machine learning. *Springer google schola*, 2: 1122–1128.

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

Clark, K.; and Jaini, P. 2024. Text-to-Image Diffusion Models are Zero Shot Classifiers. *Advances in Neural Information Processing Systems*, 36.

De, D.; Mitra, S.; and Soundararajan, R. 2024. GenzIQA: Generalized Image Quality Assessment using Prompt-Guided Latent Diffusion Models. *arXiv preprint arXiv:2406.04654*.

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

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

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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

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

Fu, H.; Wang, Y.; Yang, W.; and Wen, B. 2024. DP-IQA: Utilizing Diffusion Prior for Blind Image Quality Assessment in the Wild. *arXiv preprint arXiv:2405.19996*.

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

Golestaneh, S. A.; Dadsetan, S.; and Kitani, K. M. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1220–1230.

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

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

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

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

Hou, W.; Gao, X.; Tao, D.; and Li, X. 2014. Blind image quality assessment via deep learning. *IEEE transactions on neural networks and learning systems*, 26(6): 1275–1286.

Huang, T.; Zhang, Y.; Zheng, M.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2023. Knowledge Diffusion for Distillation. *arXiv preprint arXiv:2305.15712*.

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

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

Li, X.; Gao, T.; Hu, R.; Zhang, Y.; Zhang, S.; Zheng, X.; Zheng, J.; Shen, Y.; Li, K.; Liu, Y.; et al. 2024a. Adaptive Feature Selection for No-Reference Image Quality Assessment by Mitigating Semantic Noise Sensitivity. In *Forty-first International Conference on Machine Learning*, 27920–27941. PMLR.

Li, X.; Hu, R.; Zheng, J.; Zhang, Y.; Zhang, S.; Zheng, X.; Li, K.; Shen, Y.; Liu, Y.; Dai, P.; et al. 2024b. Integrating Global Context Contrast and Local Sensitivity for Blind Image Quality Assessment. In *Forty-first International Conference on Machine Learning*, 27920–27941. PMLR.

Li, X.; Tao, D.; Gao, X.; and Lu, W. 2009. A natural image quality evaluation metric. *Signal Processing*, 89(4): 548–555.

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

Liu, X.; van de Weijer, J.; and Bagdanov, A. D. 2017. RankIQA: Learning From Rankings for No-Reference Image Quality Assessment. In *The IEEE International Conference on Computer Vision (ICCV)*.

Liu, Y.; Ke, Z.; Liu, F.; Zhao, N.; and Lau, R. W. 2024. Diff-Plugin: Revitalizing Details for Diffusion-based Low-level Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4197–4208.

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

Pan, Z.; Zhang, H.; Lei, J.; Fang, Y.; Shao, X.; Ling, N.; and Kwong, S. 2022. 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.

Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30: 57–77.

Qin, G.; Hu, R.; Liu, Y.; Zheng, X.; Liu, H.; Li, X.; and Zhang, Y. 2023. Data-Efficient Image Quality Assessment with Attention-Panel Decoder. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saha, A.; Mishra, S.; and Bovik, A. C. 2023. Re-IQA: Unsupervised Learning for Image Quality Assessment in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5846–5855.

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

Shao, K.; Shao, J.; He, C.; and Hu, R. 2024. Class function-based adaptive disturbance observer for uncertain nonlinear systems. *International Journal of Systems Science*, 1–9.

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

Shi, C.; and Lin, Y. 2020. Full reference image quality assessment based on visual salience with color appearance and gradient similarity. *IEEE Access*, 8: 97310–97320.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

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

Tao, D.; Li, X.; Lu, W.; and Gao, X. 2009. Reduced-reference IQA in contourlet domain. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(6): 1623–1627.

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

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xu, K.; Liao, L.; Xiao, J.; Chen, C.; Wu, H.; Yan, Q.; and Lin, W. 2024. Boosting Image Quality Assessment through Efficient Transformer Adaptation with Local Feature Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2662–2672.

Yang, Z.; Li, Z.; Shao, M.; Shi, D.; Yuan, Z.; and Yuan, C. 2022. Masked generative distillation. In *European Conference on Computer Vision*, 53–69. Springer.

Ye, H.; and Xu, D. 2024. DiffusionMTL: Learning Multi-Task Denoising Diffusion Model from Partially Annotated Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27960–27969.

Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; and Bovik, A. 2020. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3575–3585.

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

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

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

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

Zhao, Y.; Ye, Q.; Wu, W.; Shen, C.; and Wan, F. 2023b. Generative prompt model for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6351–6361.

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

Zhu, Y.; Li, Y.; Sun, W.; Min, X.; Zhai, G.; and Yang, X. 2022. Blind Image Quality Assessment Via Cross-View Consistency. *IEEE Transactions on Multimedia*.