

DP-IQA: Utilizing Diffusion Prior for Blind Image Quality Assessment in the Wild

Honghao Fu^{*}, Yufei Wang^{*}, Wenhan Yang, *Member, IEEE*, Alex C. Kot, *Life Fellow, IEEE*, Bihan Wen[†], *Senior Member, IEEE*

Abstract—Blind image quality assessment (IQA) in the wild, which assesses the quality of images with complex authentic distortions and no reference images, presents significant challenges. Given the difficulty in collecting large-scale training data, leveraging limited data to develop a model with strong generalization remains an open problem. Motivated by the robust image perception capabilities of pre-trained text-to-image (T2I) diffusion models, we propose a novel IQA method, diffusion priors-based IQA (DP-IQA), to utilize the T2I model’s prior for improved performance and generalization ability. Specifically, we utilize pre-trained Stable Diffusion as the backbone, extracting multi-level features from the denoising U-Net guided by prompt embeddings through a tunable text adapter. Simultaneously, an image adapter compensates for information loss introduced by the lossy pre-trained encoder. Unlike T2I models that require full image distribution modeling, our approach targets image quality assessment, which inherently requires fewer parameters. To improve applicability, we distill the knowledge into a lightweight CNN-based student model, significantly reducing parameters while maintaining or even enhancing generalization performance. Experimental results demonstrate that DP-IQA achieves state-of-the-art performance on various in-the-wild datasets, highlighting the superior generalization capability of T2I priors in blind IQA tasks. To our knowledge, DP-IQA is the first method to apply pre-trained diffusion priors in blind IQA. The codes and checkpoints are available at <https://github.com/RomGai/DP-IQA>.

Index Terms—Blind IQA, diffusion prior, text-to-image model, knowledge distillation.

I. INTRODUCTION

ILLIONS of images are uploaded and spread across the internet daily [1]. Inevitably, some of these images are of poor quality, causing negative impressions due to their visual defects [2]. Image Quality Assessment (IQA) evaluates the visual quality of images from a human perspective, to ensure high-quality content for applications such as social media sharing and streaming [3]. Therefore, the robustness and generalization of IQA methods against various real-world distortions significantly impact the presentation of billions of images to the public. Blind IQA (BIQA) methods, also known as no-reference IQA, are crucial for evaluating image quality without reference images. In diverse and uncontrolled real-world environments (“in-the-wild”), BIQA is particularly necessary due to the unpredictable distortions present. Unlike

Honghao Fu, Yufei Wang, Alex C. Kot and Bihan Wen are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. (e-mail: {hfu006, yufei001, eackot, bihan.wen}@ntu.edu.sg)

Wenhan Yang are with the PengCheng Laboratory, China. (e-mail: yangwh@pcl.ac.cn)

^{*}These authors contributed equally.

[†]The corresponding author.

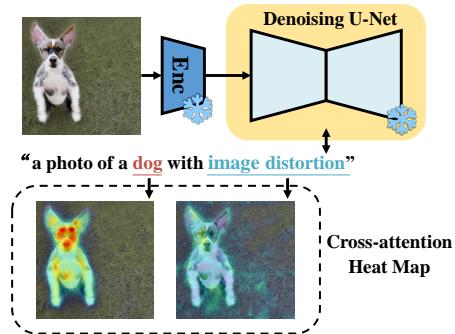


Fig. 1. The motivation of our work. Unlike commonly used classification priors for IQA, which only rely on image and category labels and focus on high-level, instance-level features, T2I models benefit from extensive, diverse training data that includes both high- and low-quality images with corresponding text prompts. As shown in the figure above, this enables T2I models to capture both high-level semantic features and low-level distortions simultaneously, making them a more effective prior for blind IQA.

methods that require reference images, BIQA directly predicts image quality, which is essential for handling authentic distortions. However, labeling the dataset to train BIQA models is laborious because it requires multiple volunteers to provide subjective scores for each image to avoid bias, resulting in a smaller scale of the dataset compared to other tasks like image classification [4], [5].

To increase the generalization ability of BIQA models under limited data, the majority of recent BIQA methods [3], [6]–[10] leverage priors from pre-trained image classification models. These priors emphasize high-level vision and consequently lack adequate low-level information, which creates potential barriers and increases the difficulty for the model in learning low-level features. This issue arises because, during classification training, images with similar high-level content but differing low-level quality are assigned the same label [11]. Furthermore, using networks pre-trained for classification does not align well with human visual perception of image quality [12]. Humans can recognize and classify objects in an image even if it is distorted, as long as the distortion is not too severe. Therefore, recent research [12]–[14] leverages the prior knowledge of visual-language multimodal models for BIQA tasks, reducing reliance on classification priors. An advanced approach involves constructing a set of text templates that describe both the high-level content and low-level quality of the input images, and utilizing the visual-language model CLIP [15] to obtain feature embeddings for both the image and corresponding text. The similarities among them are used as

metrics to further measure the image quality. However, recent research reveals that CLIP image encoder is largely insensitive to various distortion types [16], demonstrating effective performance only with a limited set of distortions (blurry, hazy, and rainy). In contrast, its text encoder proficiently manages related textual descriptions. This discrepancy results in a mismatch between the CLIP embeddings of distorted images and the clean descriptions of distortion types [16]. Furthermore, the CLIP image encoder compresses complex images into vectors, potentially leading to the loss of low-level information. Therefore, the current methods utilizing CLIP priors for BIQA still have limitations. This prompts us to explore whether BIQA could benefit from more ideal priors offered by other tasks and models.

As shown in Figure 1, inspired by the robust image perception capabilities of text-to-image (T2I) diffusion models, we propose leveraging diffusion priors for blind IQA (BIQA). While a few recent studies have explored using diffusion models [17], [18] for BIQA, they still rely on pre-trained classification models and do not fully utilize the large-scale pre-trained T2I priors. Priors from pre-trained T2I diffusion models have been effectively applied to high-level tasks such as image classification [19] and semantic segmentation [20], [21], as well as low-level tasks like super-resolution [22] and image restoration [23], [24]. This further confirms that diffusion priors encompass a rich blend of high-level and low-level information. Furthermore, employing a T2I model like Stable Diffusion (SD) [25] avoids processing distorted images through the CLIP image encoder, which is insensitive to various distortions. Instead, it only utilizes the CLIP text encoder to condition the T2I model, which can accurately embed text descriptions of image distortions. Additionally, incorporating negative prompts during inference with T2I models to prevent the generation of undesirable image content is a widely adopted engineering practice. These negative prompts typically include descriptions related to image quality, such as “blurry”, “low resolution”, “worst/low/normal quality”, and “JPEG artifacts.” This suggests that T2I models are capable of recognizing the components associated with these quality-related and distortion-related prompts in the images. However, despite these advantages, unlike IQA methods based on pre-trained classification models or CLIP, which can directly obtain feature vectors from the models’ output layer, how to effectively extract features for IQA tasks from T2I diffusion models remains an open problem.

In this paper, we explore the potential of T2I diffusion models and adapt them to better address in-the-wild BIQA with various unpredictable authentic distortions. We propose a novel BIQA method called diffusion prior-based IQA (DP-IQA). DP-IQA leverages a pre-trained SD model as the backbone, extracting multi-level features from the denoising U-Net at a specific timestep and decoding them to estimate image quality, without requiring a whole diffusion process. From a practical perspective, a text adapter is used to address the potential domain gap caused by our constant conditional embedding strategy, while an image adapter [26] supplements features from the original image to bypass the distortion information bottleneck of the variational autoencoder (VAE).

To more effectively utilize the T2I model’s image understanding and global modeling capabilities, DP-IQA processes the entire image without patch splitting, allowing for better extraction of semantic features. Unlike T2I models that require full image distribution modeling, our approach focuses on image quality assessment, which inherently requires fewer parameters. Consequently, we distill the knowledge from this model into a CNN-based student model, significantly reducing parameters to enhance its practicality in real-world applications. Experiments demonstrate that DP-IQA achieves state-of-the-art (SOTA) performance and superior generalization ability across various in-the-wild datasets. To the best of our knowledge, DP-IQA is the first method to apply T2I diffusion priors in BIQA. Our contributions are summarized as follows:

- We are the first to leverage the pretrained T2I diffusion model’s prior for blind IQA, specifically its strong ability to model semantic and low-level features simultaneously.
- We propose a framework that can better extract aesthetics-related features from activation values during the diffusion denoising step, resulting in a more compact and effective representation for subsequent prediction. Besides, the enhanced T2I diffusion priors are distilled into a lightweight model for enhanced applicability, achieving $\sim 3\times$ speed up and $\sim 14\times$ reduction in parameters under similar performance.
- The extensive experiments demonstrate the effectiveness and generalization ability of the proposed method on several in-the-wild benchmarks with authentic distortions.

II. RELATED WORKS

A. *Blind image quality assessment*

Traditional BIQA primarily leverages statistical features from the spatial and transform domains of images using natural scene statistics [27]–[29] and employs machine learning models for the regression of image quality score [30]–[33]. However, these methods often fail to capture high-level image information due to their reliance on specific feature computations. Recently, deep learning has advanced BIQA significantly [11], [34]–[39]. Initial methods used Convolutional Neural Networks (CNNs) to learn image quality features [40], [41], while recent works [6], [42] propose to leverage powerful Vision Transformer (ViT) [43] for better performance.

To address the challenge posed by the limited scale of IQA datasets hindering the models’ representational capabilities, utilizing priors from classification models pre-trained on larger-scale image datasets like ImageNet is a common practice [7]–[11], [44]–[52]. However, as discussed in the previous section, it exhibits significant differences from human visual perception habits. There are also some works avoid using pre-trained classification models. For example, early generative models such as Generative Adversarial Networks (GANs) have been applied to IQA tasks [53]–[55]. GAN-based methods typically reconstruct an undistorted image from a distorted one, then extract features from this process, or use the reconstructed image as a reference for IQA. Consequently, they require undistorted reference images during training, which limits their applicability to in-the-wild images without references.

More recent works, such as CLIP-IQA [13], LIQE [12] and IPCE [14], adopt the priors of vision-language model CLIP for BIQA [56]. They perform IQA by minimizing the cosine similarity between the CLIP embedding of the image and the CLIP embedding of text describing its content and quality. However, as stated in the previous section, the CLIP image encoder is not sensitive to a large number of distortion types, while its text encoder can accurately embed text describing these distortions, leading to a mismatch between image and text embeddings [16]. Therefore, applying CLIP priors to in-the-wild BIQA may still have limitations.

Recently, a few studies have applied diffusion models to BIQA. PFD-IQA [17] trains a diffusion model to denoise prior features of images obtained through pre-trained ViT and performed regression on the denoised features to predict quality scores. GenzIQA [57] utilizes cross-attention maps from a pretrained diffusion model as quality representations and employs prompt tuning to adapt the conditional embeddings for the IQA task. DiffV²IQA [18] trains a diffusion model on 2 small-scale synthetic distortion datasets to restore distorted images to high-quality images, and uses ViT and ResNet [58] to obtain the features of intermediate denoised images from the denoising process to predict quality scores. However, due to the poor performance of its self-trained diffusion model, the restored images significantly deviate from the original images, introducing new distortions not accounted for in the datasets' scoring system. Additionally, since the synthetic distortion datasets contains only a limited number of distortion types, the self-trained diffusion model lacks robustness to complex real-world distortions.

B. Diffusion model priors

Diffusion-based generative models excel in generating high-quality images with intricate scenes and semantics from textual descriptions, demonstrating a profound understanding of text and vision. The prior knowledge embedded in large-scale pre-trained diffusion models like SD has proven effective for high-level visual tasks such as image classification [19], semantic segmentation [20], [21], and depth estimation [21], [59]. Additionally, it has also been utilized in low-level tasks like super-resolution [22] and image restoration [23], [24], [60], showing impressive results. This indicates that the diffusion priors contain sufficient high-level and low-level information with no significant barriers between them. However, the challenge of fully harnessing its ability to represent visual features for image quality assessment and effectively utilizing its prior knowledge remains an open problem. Recent studies have provided compelling evidence [61] suggesting that the representational capability of diffusion models is predominantly derived from the denoising-driven process rather than the diffusion-driven process. Consequently, we focus on exploring strategies to effectively leverage prior knowledge embedded in the denoiser within the diffusion pipeline.

III. METHODOLOGY

A. Preliminary

Diffusion. As the backbone of our proposed DP-IQA, we first provide a brief introduction to the principles of diffusion

models. Let z_t be the random noise at the t -th timestep. Diffusion models transform z_t to the denoised sample z_0 by gradually denoising z_t to a less noisy z_{t-1} . The forward diffusion process is modeled as:

$$q(z_t | z_{t-1}) = \mathcal{N}(z_t; \sqrt{\alpha_t} z_{t-1}, (1 - \alpha_t)\mathbf{I}), \quad (1)$$

where $\{\alpha_t\}$ are fixed coefficients that determine the noise schedule. By defining $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, z_t can be obtained directly from z_0 [62]:

$$q(z_t | z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t} z_{t-1}, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (2)$$

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (3)$$

It makes sampling for any z_t more efficient. With proper re-parameterization, the training objective of diffusion models can be derived as [21], [63]:

$$\mathcal{L}_{DM} = \mathbb{E}_{z_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t(z_0, \epsilon), t; \mathcal{C})\|^2], \quad (4)$$

where ϵ_θ is a denoising autoencoder that is learned to predict ϵ given the conditional embedding \mathcal{C} . In our task, the denoising autoencoder ϵ_θ is a U-Net, z_t is a latent representation of a distorted image, which can also be regarded as a latent variable that has not been fully denoised from random noise. By controlling the conditional embedding \mathcal{C} , we enable the denoising U-Net to effectively extract different features from z_t , and thereby extract the prior knowledge required for the IQA task from a single timestep in the diffusion process.

B. Overview

We adapt the representation capabilities and priors of T2I diffusion models to BIQA **in the wild**, as illustrated in Figure 2. Specifically, the input image is first encoded with a pre-trained VAE encoder, then fed into the denoising U-Net of the pre-trained SD [25]. Concurrently, a CLIP encoder [15] converts text describing the image quality into conditional embeddings for the denoising U-Net. The input text is templated and consistent across all images. Meanwhile, text and image adapters are adopted to mitigate the domain gap caused by the constant conditional embedding strategy and correct the information loss caused by the VAE bottleneck. Subsequently, we extract feature maps from each stage of the U-Net's upsampling process, which are then fused and decoded by a well-designed Quality Feature Decoder (QFD). Finally, a Multi-Layer Perceptron (MLP) is employed to regress the image quality scores. Figure 3 provides details on the adapters and QFD. After obtaining the above teacher model, we distill the knowledge in the trained DP-IQA into an EfficientNet-based [64] student model, which is initialized with the official pre-trained weights, and its output structure is modified to align with the teacher model. The distillation process leverages two sources of supervision: (1) the output feature map from the QFD, and (2) the GT image quality scores.

Extracting diffusion priors from a single timestep. A pre-trained T2I diffusion model contains sufficient information to sample from the data distribution, including its low-level features and structures, as the model can be viewed as the learned gradient of data density [21]. With limited natural language supervision during pre-training, the T2I model also

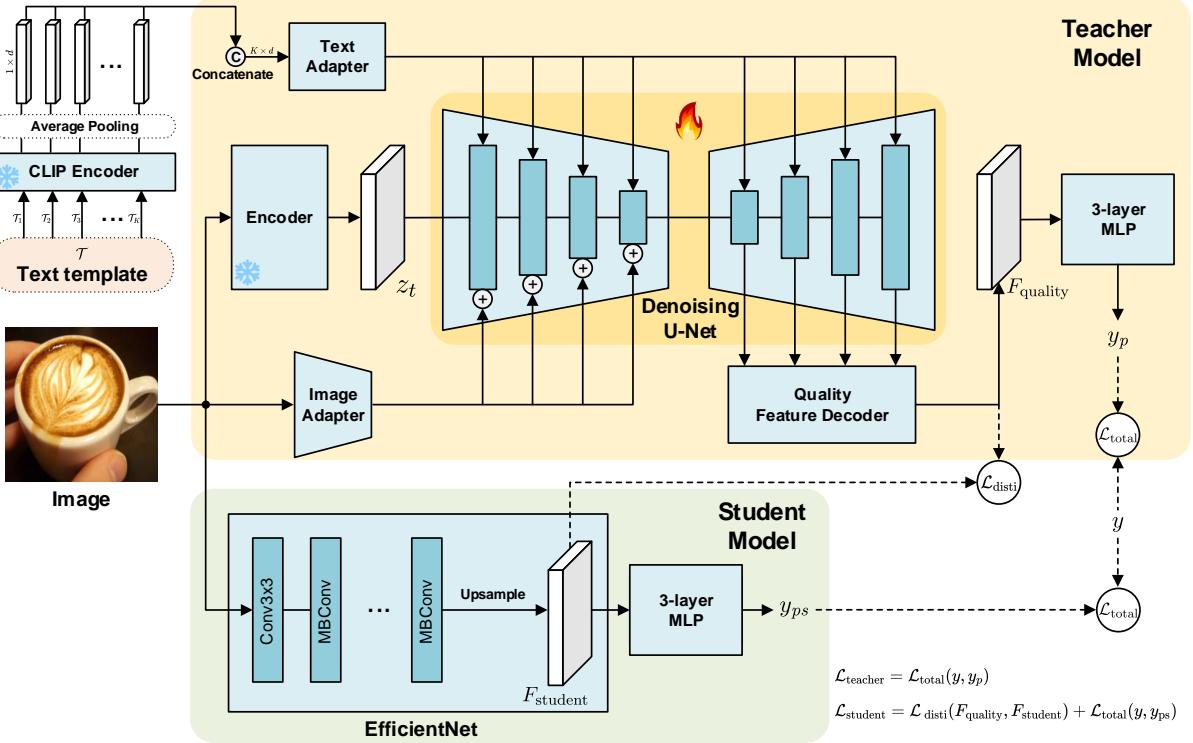


Fig. 2. Framework of DP-IQA and its corresponding student model by knowledge distillation. The pretrained CLIP encoder is first used to convert quality assessment-related text templates into embeddings, serving as the condition for the denoising U-Net. A text adapter is employed to bridge the gap introduced by fixed text templates, while an image adapter is incorporated to preserve low-level features essential for IQA. Finally, a Quality Feature Decoder (QFD) is designed to fuse features across multiple levels, producing the final feature maps, which are then passed through an MLP to obtain the final result. Unlike diffusion models, we generate the feature maps in a single feedforward pass by using a fixed timestep. To further improve the efficiency, we distill the knowledge from DP-IQA into a student model with EfficientNet as the backbone to further reduce parameters and increase inference speed. The loss functions are detailed in equation (11) and (12).

incorporates significant high-level knowledge. Recent work l-DAE [61] shows that the representational power of denoising diffusion models primarily stems from the denoising process rather than the diffusion itself. This suggests that a single-step denoising is sufficient to leverage the representation capability of the pretrained denoiser, without requiring a diffusion process. Thus, we adopt a single-timestep setting for our task. We utilize the pre-trained SD as our backbone. Assume we wish to utilize the diffusion priors expressed by the denoising U-Net ϵ_θ at timestep t . For an input image $x \in \mathbb{R}^{H \times W \times 3}$, it is encoded into latent representation z_t by a pretrained VAE. Then, from $\epsilon_\theta(z_t, t)$, we obtain the feature maps f_{up}^i at each upsampling stage, where $i = 1, 2, 3, 4$. The resulting set of feature maps $F_{\text{up}}^t = \{f_{\text{up}}^{t,1}, f_{\text{up}}^{t,2}, f_{\text{up}}^{t,3}, f_{\text{up}}^{t,4}\}$ is the prior features at t . Such a multi-level feature extraction strategy facilitates a more comprehensive representation of the image, as the upsampling process of the denoising U-Net transitions features from low-resolution, high-level semantics to high-resolution, fine-grained details. Moreover, since multi-level features from the downsampling stages are fused into the upsampling path via skip connections, this strategy also effectively captures information propagated during the downsampling process.

Text template. In a T2I diffusion model, text is converted into conditional embeddings by a text encoder to guide the denoising process. SD uses a CLIP encoder for embedding text. An appropriate text prompt is crucial for the denoiser to focus on the target features. We use a general text template

summarized by previous MLLM-based arts [12], [65], [66] to describe the image's content and quality as the text conditional input. The template is “a photo of a {scenes} with {distortion type} distortion, which is of {quality level} quality.” {Scenes} includes typical image subjects, {distortion type} covers common distortions, and {quality level} provides a general quality description, like “bad” or “good”. Both scenes and distortion types also offer general descriptions. For example, “a photo of an animal with realistic blur distortion, which is of bad quality.” However, real-world distortions are often complex and cannot be fully captured by a limited set of labels. To ensure robustness, we use the category “other” to represent distortions not explicitly defined in our templates. Assuming there are l_s scenes, l_d distortion types, and l_q quality levels, there are a total of $K = l_s \cdot l_d \cdot l_q$ combinations. We define \mathcal{T} as the set of all combinations, where \mathcal{T}_k is the k -th sentence in \mathcal{T} .

Constant conditional embedding. In T2I models, the text is typically split into a sequence of tokens, each of which is encoded into a conditional embedding. These embeddings are then used in the U-Net’s cross-attention mechanism to guide the model’s focus toward the content described in the text and its corresponding image features. Benefiting from this, our method does not require setting specific text template content for each input image. Instead, it inputs all the template combinations simultaneously. In the CLIP encoder E_C of the T2I diffusion model, an input prompt is split into multiple

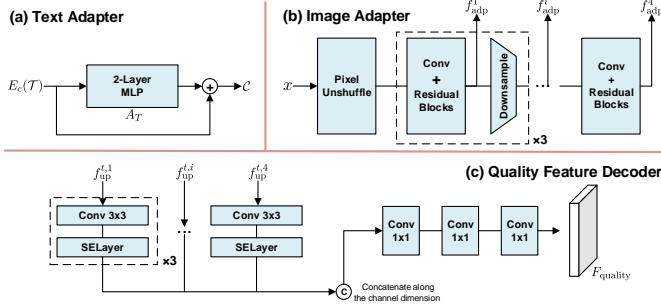


Fig. 3. Details of the (a) text adapter, (b) image adapter [67] and (c) quality feature decoder in DP-IQA. The image adapter is a lightweight module designed to complement the potential loss of degradation-specific information in the pretrained encoder of Stable Diffusion.

tokens (77 in SD by default, which can be modified). Define the output dimension of E_C as d , then each token is converted into an $1 \times d$ embedding. The embeddings of all tokens are concatenated ($77 \times d$) as the condition, influencing the attention mechanism. This allows us to treat each sentence in our template as a separate token, combining them into a universal constant condition embedding in our task, prompting the U-Net to be able to focus on all the distortion scenarios it needs to pay attention to. In practice, each sentence is first split into tokens, and the embeddings of all its tokens are average pooled to produce a vector with the same shape as the embedding of a single token ($1 \times d$), which represents the global embedding of a sentence. The pooled results of K sentences from the template \mathcal{T} are then concatenated to form an overall embedding with shape $K \times d$, which is simplified to $E_C(\mathcal{T}) \in \mathbb{R}^{K \times d}$. $E_C(\mathcal{T})$ will be used as a constant in our task to provide a universal conditional embedding. Therefore, we can apply all combinations of text template to an image at once, which helps the model better understand an image with multiple scenes and distortions

C. Diffusion prior-based IQA (DP-IQA)

Text adapter. However, our text template and conditional embeddings slightly differs from the standard strategy of pretrained SD, which may lead to the potential domain gap. Previous work [68] has shown that compared to learning task-specific prompts, introducing adapters can improve CLIP’s performance on new tasks beyond the pre-training setting while also reducing computational costs. Therefore, we use a text adapter [21], [69], [70] to mitigate this gap. CLIP-Adapter [68], which systematically demonstrates that compared to learning task-specific prompts, introducing adapters can improve CLIP’s performance on new tasks beyond the pre-training setting while also reducing computational costs. The text adapter consists of a two-layer MLP A_T , and takes $E_C(\mathcal{T})$ as input. The output of A_T is then added to $E_C(\mathcal{T})$ to obtain the adjusted conditional embedding $C \in \mathbb{R}^{K \times d}$. This process is:

$$C = E_C(\mathcal{T}) + A_T(E_C(\mathcal{T})). \quad (5)$$

Image adapter. Previous work has shown that VAEs act as a lossy compression method [71], typically encoding only

information that cannot be locally reconstructed into the latent space, such as long-range dependencies. Specifically, VAEs preserve only global structures in the latent variables, while local details are modeled directly by the decoder. This results in the loss of low-level details during the encoding process. In addition, some diffusion-based works focused on low-level vision have also found that VAEs are not robust to distortions under standard training conditions [72], [73]. However, finetuning a VAE adapted for a pretrained diffusion pipeline on low-level tasks is prohibitively expensive. To address this, we introduce an image adapter A_I that bypasses the VAE’s compression process and directly extracts features from the original image x as a supplement. These features are fed into the denoising U-Net’s downsampling process. Define the feature map at each downsampling stage as f_{down}^i , where $i = 1, 2, 3, 4$. The set of the feature maps at timestep t is $F_{\text{down}}^t = \{f_{\text{down}}^{t,1}, f_{\text{down}}^{t,2}, f_{\text{down}}^{t,3}, f_{\text{down}}^{t,4}\}$. Define the output of the image adapter as $A_I(x) = F_{\text{adp}}^i = \{f_{\text{adp}}^1, f_{\text{adp}}^2, f_{\text{adp}}^3, f_{\text{adp}}^4\}$, which is independent of the timestep t , and the size of f_{adp}^i is consistent with $f_{\text{down}}^{t,i}$. The process of feature supplementation by the image adapter is:

$$F_{\text{down}}^{t,i} = F_{\text{down}}^{t,i} + F_{\text{adp}}^i, \quad i = 1, 2, 3, 4. \quad (6)$$

Quality feature decoder (QFD). We design a CNN-based QFD D to decode the feature maps from the upsampling stages, and then regress the output of the decoder through an MLP to obtain the image quality score. QFD first accepts $f_{\text{up}}^{t,1}, f_{\text{up}}^{t,2}, f_{\text{up}}^{t,3}, f_{\text{up}}^{t,4}$ in F_{up}^t as input, and upsamples all of them to a size of 64×64 . Next, a convolution layer and a squeeze-and-excite (SE) layer are used to unify the channel number to 512 for each feature map, and the four feature maps are concatenated into a single feature map with 2048 channels. This concatenated feature map is then processed through four convolution layers to gradually reduce the number of channels to 512, 128, 32, and 8. The QFD finally outputs an image quality feature map of size $64 \times 64 \times 8$ as $F_{\text{quality}} = D(F_{\text{up}}^t)$. The F_{quality} is flattened into a one-dimensional vector and passed through a regression network R , which consists of a three-layer MLP, to perform score regression and obtain the predicted value y_p . The process is as follows:

$$F_{\text{quality}} = D(F_{\text{up}}^t) = D(f_{\text{up}}^{t,1}, f_{\text{up}}^{t,2}, f_{\text{up}}^{t,3}, f_{\text{up}}^{t,4}), \quad (7)$$

$$y_p = R(\text{Flatten}(F_{\text{quality}})). \quad (8)$$

Model optimization. Our model is trained in an end-to-end manner. The loss function consists of Mean Squared Error (MSE) loss \mathcal{L}_{mse} and Margin loss \mathcal{L}_{mgn} , which are commonly used for learning image quality score regression and ranking (i.e., distinguishing the quality relationship within a batch) in IQA. Assuming the batch size is n , the GT image quality score is y , the predicted value is y_p , and the standard deviation of y is σ_y , the loss functions are as follows:

$$\mathcal{L}_{\text{mse}} = \frac{1}{n} \|y - y_p\|_2^2, \quad (9)$$

$$\mathcal{L}_{\text{mgn}} = \frac{2 \sum_{i < j} \max(0, -\text{sign}(y_i - y_j) \cdot (y_{p_i} - y_{p_j}) + m)}{n(n-1)}, \quad (10)$$

TABLE I
LEARNING RATE DECAY AT WHICH EPOCH.

| Model | CLIVE | KonIQ | LIVEFB | SPAQ |
|---------|--------|-------|--------|------|
| Teacher | - | 5 | 2 | - |
| Student | 10, 25 | 5 | 4 | 6 |

TABLE II
THE VALUES OF THE NUMERIC VARIABLES DEFINED IN SEC. III

| Variable | Value | Explanation |
|-----------|-------|--|
| H | 512 | Height of the input image |
| W | 512 | Width of the input image |
| l_s | 11 | The number of elements in {scenes} |
| l_d | 35 | The number of elements in {distortion type} |
| l_q | 5 | The number of elements in {quality level} |
| K | 1925 | The total number of combinations of text templates |
| d | 768 | Output dimension of the CLIP encoder |
| λ | 0.25 | Coefficient used to control the margin |

where $m = \lambda\sigma_y$, $\lambda \in [0, 1]$. Therefore, the overall loss function $\mathcal{L}_{\text{total}}$ can be defined as:

$$\mathcal{L}_{\text{total}}(y, y_p) = L_{\text{mse}}(y, y_p) + L_{\text{margin}}(y, y_p). \quad (11)$$

This model is referred to as the “teacher model”, and its loss function can also be written as $\mathcal{L}_{\text{teacher}} = \mathcal{L}_{\text{total}}(y, y_p)$.

D. Knowledge distillation

Student model. Unlike T2I models that require full image distribution modeling which requires a large network capacity, our approach focuses on IQA, which inherently requires fewer parameters. To reduce the model’s parameters and increase inference speed, we propose distilling the feature distribution of DP-IQA into a student model. We use a lightweight EfficientNet [64] as the student model and adjust its output structure to align with that of the teacher model. By distillation, the student network only needs to learn the prior that corresponds to the image quality assessment.

Model optimization. The student model takes the image as input and uses the output feature map F_{quality} from the QFD as supervision to distill the image quality knowledge learned by the teacher model, we use the MSE shown in Equation (9) as the distillation loss $\mathcal{L}_{\text{disti}}$. Additionally, the student model is supervised by the GT image quality score y . Assuming the last feature map before the output layer of the student model is F_{student} , the predicted value of student model is p_{ps} , the loss function $\mathcal{L}_{\text{student}}$ for the student model can be defined as:

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{disti}}(F_{\text{quality}}, F_{\text{student}}) + \mathcal{L}_{\text{total}}(y, y_{\text{ps}}) \quad (12)$$

IV. EXPERIMENT

A. Datasets and evaluation metrics

Datasets. IQA datasets primarily consist of distorted images paired with quality scores. We assess our DP-IQA using four in-the-wild IQA datasets: CLIVE [74], KonIQ [4], LIVEFB (FLIVE) [36] and SPAQ [75], containing 1162, 10073, 11125, and 39810 authentically distorted (in-the-wild) images, respectively. Our research focuses on authentic distortion types,

TABLE III
DETAILS OF THE TEXT TEMPLATE WE USE IN THE EXPERIMENT.

| Word types | Details |
|-----------------|--|
| Scenes | animal, cityscape, human, indoor, landscape, night, plant, still_life, other |
| Distortion type | jpeg2000 compression, jpeg compression, motion, white noise, gaussian blur, fastfading, fnoise, lens, diffusion, shifting, color quantization, desaturation, oversaturation, underexposure, overexposure, contrast, white noise with color, impulse, multiplicative, jitter, white noise with denoise, brighten, darken, pixelate, shifting the mean, noneccentricity patch, quantization, color blocking, sharpness, realistic blur, realistic noise, realistic contrast change, other realistic, other |
| Quality level | bad, poor, fair, good, perfect |

therefore artificially distorted datasets [76] such as LIVE [77], CSIQ [78] and KADID [79], as well as artistic or stylized datasets, were not included in our study.

Evaluation metrics. Consistent with other works, we use Pearson’s linear correlation coefficient (PLCC) and Spearman’s rank-order correlation coefficient (SRCC) as performance evaluation metrics. PLCC measures the strength of a linear relationship between predicted and true values. Meanwhile, SRCC assesses the consistency of rank ordering between predicted and true values, focusing on monotonic relationships. Together, they provide a comprehensive evaluation, where higher values for both signify better performance.

B. Implementation

We implement our model using PyTorch and conduct training and testing on an A100 GPU. The version of stable diffusion is v1.5, while EfficientNet-B7 served as the backbone for the student model. We use Adam as the optimizer. The teacher model is trained with a batch size of at least 12, an initial learning rate of 10^{-5} , for up to 15 epochs, while the student model is trained with a batch size of at least 24, an initial learning rate of 10^{-4} and for up to 30. Besides, the validation step for CLIVE is 50 while for other datasets is 250. Due to varying dataset scales, learning rate decay differ slightly across datasets, as detailed in Table I, and the scheduler is MultiStepLR, decay factor is 0.2. We also provide detailed values of the numeric variables defined in Sec.3 in Table II. For data preprocessing, we resize in-the-wild images to 512×512 pixels without patch splitting. We randomly split datasets into training and testing sets in 8:2, and repeat the splitting process five times for all datasets and report the median results. Besides, We present the specific settings of the template in Table III.

C. Comparison against other methods

Overall comparison. We compare our method with 18 SOTA baselines¹ Table IV compares the performance of

¹Preprints and works w/o released code are not included in the comparison. Besides, we do not include comparisons with works that use customized experimental settings, such as joint training on multiple datasets or other conditions.

TABLE IV
COMPARISON OF OUR PROPOSED DP-IQA WITH SOTA BIQA ALGORITHMS ON AUTHENTICALLY DISTORTED (IN-THE-WILD) DATASETS. BOLD ENTRIES INDICATE THE TOP TWO RESULTS. '-' ARE NOT AVAILABLE PUBLICLY.

| Dataset | CLIVE | | KonIQ | | LIVEFB (FLIVE) | | SPAQ | |
|----------------|--------------|--------------|--------------|--------------|----------------|--------------|--------------|--------------|
| | Metrics | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC |
| DIIVINE [80] | 0.591 | 0.588 | 0.558 | 0.546 | 0.187 | 0.092 | 0.660 | 0.599 |
| BRISQUE [81] | 0.629 | 0.629 | 0.685 | 0.681 | 0.341 | 0.303 | 0.817 | 0.809 |
| ILNIQE [82] | 0.508 | 0.508 | 0.537 | 0.523 | 0.332 | 0.294 | 0.712 | 0.713 |
| BIECON [83] | 0.613 | 0.613 | 0.654 | 0.651 | 0.428 | 0.407 | - | - |
| MEON [40] | 0.710 | 0.697 | 0.628 | 0.611 | 0.394 | 0.365 | - | - |
| WaDIQaM [84] | 0.671 | 0.682 | 0.807 | 0.804 | 0.467 | 0.455 | - | - |
| DBCNN [37] | 0.869 | 0.851 | 0.884 | 0.875 | 0.551 | 0.545 | 0.915 | 0.911 |
| MetaIQA [39] | 0.802 | 0.835 | 0.856 | 0.887 | 0.507 | 0.540 | - | - |
| P2P-BM [36] | 0.842 | 0.844 | 0.885 | 0.872 | 0.598 | 0.526 | - | - |
| HyperIQA [49] | 0.882 | 0.859 | 0.917 | 0.906 | 0.602 | 0.544 | 0.915 | 0.911 |
| TIQA [42] | 0.861 | 0.845 | 0.903 | 0.892 | 0.581 | 0.541 | - | - |
| MUSIQ [6] | 0.746 | 0.702 | 0.928 | 0.916 | 0.661 | 0.566 | 0.921 | 0.918 |
| TReS [7] | 0.877 | 0.846 | 0.928 | 0.915 | 0.625 | 0.554 | - | - |
| DEIQT [8] | 0.886 | 0.861 | 0.934 | 0.921 | 0.645 | 0.557 | 0.921 | 0.914 |
| CLIP-IQA [13] | 0.832 | 0.805 | 0.909 | 0.895 | - | - | 0.866 | 0.864 |
| ReIQA [3] | 0.854 | 0.840 | 0.923 | 0.914 | - | - | 0.925 | 0.918 |
| SaTQA [85] | 0.903 | 0.877 | 0.941 | 0.930 | 0.676 | 0.582 | - | - |
| LIQE [12] | - | - | 0.912 | 0.928 | - | - | 0.919 | 0.922 |
| Q-Align [86] | - | - | 0.941 | 0.940 | - | - | 0.933 | 0.930 |
| LoDa [9] | 0.899 | 0.876 | 0.944 | 0.932 | 0.679 | 0.578 | 0.928 | 0.925 |
| Ours (student) | 0.902 | 0.875 | 0.944 | 0.926 | 0.671 | 0.567 | 0.923 | 0.920 |
| Ours (teacher) | 0.913 | 0.893 | 0.951 | 0.942 | 0.683 | 0.579 | 0.926 | 0.923 |

TABLE V

COMPARISON OF SRCC ON CROSS DATASETS SETTING, *i.e.*, WE TEST AND REPORT THE PERFORMANCE OF MODELS ON UNSEEN DATASETS. BOLD ENTRIES INDICATE THE TOP TWO RESULTS. '-' ARE NOT AVAILABLE PUBLICLY.

| Training on | LIVEFB | | CLIVE | | KonIQ | |
|----------------|--------------|--------------|--------------|--------------|-------|-------|
| | KonIQ | CLIVE | KonIQ | CLIVE | KonIQ | CLIVE |
| DBCNN | 0.716 | 0.724 | 0.754 | 0.755 | | |
| P2P-BM | 0.755 | 0.738 | 0.740 | 0.770 | | |
| HyperIQA | 0.758 | 0.735 | 0.772 | 0.785 | | |
| TReS | 0.713 | 0.740 | 0.733 | 0.786 | | |
| SaTQA | - | - | 0.788 | 0.791 | | |
| Q-Align | - | - | - | 0.853 | | |
| LoDa | 0.763 | 0.805 | 0.745 | 0.811 | | |
| Ours (student) | 0.767 | 0.758 | 0.781 | 0.830 | | |
| Ours (teacher) | 0.771 | 0.770 | 0.766 | 0.833 | | |

TABLE VI

THE RESULTS OF REGRESSION ON THE OUTPUT FEATURES OF DIFFERENT PRE-TRAINED BACKBONES WITH FROZEN PARAMETERS TO EVALUATE IMAGE QUALITY.

| Backbone | Learning Paradigm | CLIVE | | KonIQ | |
|-----------------------|-------------------|--------------|--------------|--------------|--------------|
| | | PLCC | SRCC | PLCC | SRCC |
| CLIP (b/16) [15] | | 0.758 | 0.765 | 0.828 | 0.800 |
| MAE (b/16) [87] | Self-supervised | 0.640 | 0.618 | 0.720 | 0.689 |
| DINOv2 (b/14) [88] | | 0.626 | 0.600 | 0.701 | 0.665 |
| ViT (b/16) [43] | | 0.524 | 0.502 | 0.682 | 0.646 |
| ResNet-50 [58] | Supervised | 0.770 | 0.767 | 0.837 | 0.821 |
| Stable Diffusion [25] | | 0.869 | 0.817 | 0.929 | 0.908 |

our method with others across four widely recognized in-the-wild datasets. Results for DEIQT [8] are based on our reproduction, while those for other methods are taken from the original papers of LoDa [9], TReS [7], and SaTQA [85]. The experimental findings demonstrate that our proposed method

TABLE VII
ABLATION ANALYSIS OF TEXT PROMPT (TP), CONSTANT CONDITIONAL EMBEDDING (CCE), TEXT ADAPTER (TA) AND IMAGE ADAPTER (IA) IN TEACHER MODEL. BOLD ENTRIES INDICATE THE BEST RESULTS.

| Dataset | Full | w/o TP | w/o CCE | w/o TA | w/o IA |
|---------|---------------------------|-------------|-------------|-------------|-------------|
| | PLCC SRCC | PLCC SRCC | PLCC SRCC | PLCC SRCC | |
| CLIVE | 0.913 0.893 | 0.867 0.871 | 0.898 0.878 | 0.907 0.881 | 0.904 0.875 |
| KonIQ | 0.951 0.942 | 0.929 0.931 | 0.937 0.928 | 0.941 0.940 | 0.946 0.932 |

achieves SOTA performance on the CLIVE (LIVEC), KonIQ, and LIVEFB (FLIVE) datasets, underscoring its effectiveness across diverse real-world scenarios. Additionally, on the SPAQ dataset, our method delivers highly competitive results that are nearly on par with the best-performing approach. Furthermore, the student model exhibits only a marginal decline in performance across these datasets. This outcome aligns with expectations and highlights the success of the distillation process in effectively transferring knowledge.

Generalization ability. The practical value of a model is closely tied to its generalization capability. In Table V, we evaluate our model's generalization through cross-dataset zero-shot performance on three in-the-wild datasets. Training is done on one dataset, and testing on unseen datasets. We compare our method with SOTA baselines, and the results show superior generalization in most cases. Additionally, the student model performs similarly to the teacher model but with far fewer parameters, demonstrating its practical value. For larger datasets (KonIQ, LIVEFB, and SPAQ), the student model's performance closely matches the teacher model, indicating effective knowledge distillation. However, for smaller dataset (CLIVE), the student model outperforms the teacher model. This may be because the teacher model is too large and poses a higher risk of overfitting on very small datasets, whereas the

TABLE VIII

ABLATION ANALYSIS OF THE SETTINGS OF TIMESTEP FOR TEACHER MODEL. BOLD ENTRIES INDICATE THE BEST RESULTS.

| Dataset | Timestep | | | | |
|-----------|---------------------------|---------------------------|-------------|--------------------|-------------|
| | 1 | 5 | 10 | 20 | 50 |
| PLCC SRCC | PLCC SRCC | PLCC SRCC | PLCC SRCC | PLCC SRCC | PLCC SRCC |
| CLIVE | 0.913 0.893 | 0.913 0.893 | 0.912 0.879 | 0.913 0.879 | 0.907 0.871 |
| KonIQ | 0.951 0.942 | 0.947 0.939 | 0.945 0.936 | 0.946 0.936 | 0.942 0.931 |

TABLE IX

ABLATION ANALYSIS OF THE MULTI-LEVEL FEATURE EXTRACTION STRATEGY, WHERE THE TIMESTEP $t = 1$. BOLD ENTRIES INDICATE THE BEST RESULTS.

| Dataset | Full | only $f_{\text{up}}^{t,1}$ | only $f_{\text{up}}^{t,2}$ | only $f_{\text{up}}^{t,3}$ | only $f_{\text{up}}^{t,4}$ |
|---------|---------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| | PLCC SRCC | PLCC SRCC | PLCC SRCC | PLCC SRCC | PLCC SRCC |
| CLIVE | 0.913 0.893 | 0.867 0.812 | 0.874 0.845 | 0.879 0.841 | 0.869 0.817 |
| KonIQ | 0.951 0.942 | 0.923 0.903 | 0.937 0.921 | 0.941 0.927 | 0.929 0.908 |

use of the student model significantly ameliorated this issue, improving generalizability.

Effectiveness of different priors. In Table VI, we use linear probing to assess the prior effectiveness of various pretrained backbones for IQA. We freeze the backbone parameters and train only a linear regressor on features from supervised models like ResNet-50 [58] and ViT [43], and self-supervised ones such as CLIP [15], MAE [87], and DINOv2 [88]. Results show that diffusion-based representations are more effective for IQA. We hypothesize that this is because the learning strategy of the T2I diffusion model allows its prior to be viewed as the learned gradient of data density, which contains rich low-level information. Additionally, with natural language supervision, the model also incorporates significant high-level semantic knowledge. Such combined prior enables the T2I diffusion model to achieve better performance in IQA tasks.

D. Ablation

Text prompt and adapters. As shown in Table VII, we explore the impact of text prompt and constant conditional embedding strategy, and “w/o CCE” means using the description in the template that best matches the current image content as input, rather than using all templates. Additionally, we also conduct ablation studies on the text and image adapters. When there is no text prompt (w/o TP), the text adapter was not activated by default. The results indicate that the text prompt, constant conditional embedding strategy, image adapter, and text adapter play positive roles in overall performance.

Timesteps. We observe the impact of different timestep settings on model performance. As shown in Table VIII, using smaller timesteps is generally more advantageous. Therefore, we set $t = 1$ in our other experiments. However, considering that the model’s prior knowledge is embedded in its learned parameters, the choice of time steps does not have a decisive impact in the case of full parameter fine-tuning.

Multi-level features. As shown in Table IX, we conduct ablation analysis on the multi-level feature extraction strategy,

TABLE X

ABLATION ANALYSIS OF FEATURE EXTRACTED FROM EACH LAYER, WHERE THE TIMESTEP $t = 1$. BOLD ENTRIES INDICATE THE BEST RESULTS.

| Dataset | Full | w/o $f_{\text{up}}^{t,1}$ | w/o $f_{\text{up}}^{t,2}$ | w/o $f_{\text{up}}^{t,3}$ | w/o $f_{\text{up}}^{t,4}$ |
|---------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| | PLCC SRCC |
| CLIVE | 0.913 0.893 | 0.909 0.891 | 0.904 0.875 | 0.897 0.874 | 0.910 0.893 |
| KonIQ | 0.951 0.942 | 0.951 0.939 | 0.947 0.941 | 0.945 0.936 | 0.949 0.942 |

TABLE XI

ABLATION ANALYSIS OF THE DISTILLATION LOSS $\mathcal{L}_{\text{DISTI}}$. USING THE DISTILLATION CAN SIGNIFICANTLY IMPROVE THE PERFORMANCE THAN TRAINING USING L_{TEACHER} UNDER THE SAME LIGHTWEIGHT BACKBONE. BOLD ENTRIES INDICATE THE BEST RESULTS.

| Dataset | Distilled student | | w/o distillation loss | |
|---------|-------------------|--------------|-----------------------|-------|
| | PLCC | SRCC | PLCC | SRCC |
| CLIVE | 0.902 | 0.875 | 0.717 | 0.715 |
| KonIQ | 0.944 | 0.926 | 0.881 | 0.841 |

where we use only one layer of features for image quality assessment. The experimental results show that using features from a single layer leads to a performance drop, which highlights the importance of multi-scale features. Moreover, as shown in Table X, we find that each level of features positively impact the results, with $f_{\text{up}}^{t,2}$ and $f_{\text{up}}^{t,3}$ being potentially more important. However, although the impact of features from different layers varies, discarding any layer leads to performance degradation, indicating that the contributions of features from different layers are partially orthogonal and indispensable.

Distillation. As shown in Table XI, we conduct ablation analysis on the distillation. Experimental results indicate that distillation effectively enhances the performance of the student model. The results show the effectiveness of distilling enhanced priors from DP-IQA compared with training from scratch, which also demonstrates that the outstanding performance of the student model does not solely rely on its own architecture and pre-trained weights. A comparison of the number of parameters and inference speed is shown in Table XII, where student model achieves similar performance with $\sim 3\times$ speed up and $\sim 14\times$ size reduction.

Training strategies and pre-trained priors. As shown in Table XIII, we conducted an ablation study to assess the effects of training strategies and pretrained priors. For training strategies, we examined freezing the pre-trained U-Net and training QFD and adapters, as well as applying LoRA [89]. Although these approaches did not surpass SOTA methods, they achieved reasonable performance, indicating the value of the pre-trained SD prior for IQA. However, compared to full-parameter fine-tuning, training from scratch (only training QFD and adapters or using LoRA) required significantly more steps to converge, while offering limited computational savings and encountering notable performance bottlenecks. Moreover, the results highlight the importance of leveraging pretrained SD weights, as models without them struggled to extract meaningful features from limited IQA data, further demonstrating the effectiveness of pre-trained priors.



Fig. 4. Saliency maps generated by DP-IQA on in-the-wild images from the KonIQ-10k dataset. The model effectively prioritizes complex structures and semantically significant regions, aligning with human visual perception. Additionally, the model does not exhibit significant overfitting to noise, demonstrating strong noise resistance.

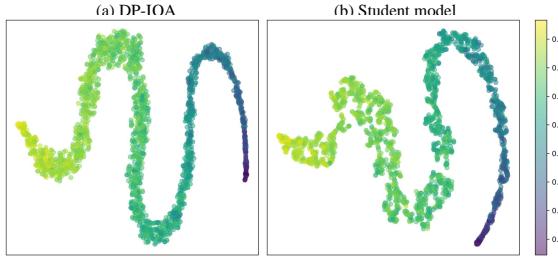


Fig. 5. t-SNE visualization of the feature embeddings from the last hidden layer of DP-IQA and its student models. The input images (points in the subplots) are from the test set of Koniq-10k, with the color gradient representing their ground truth scores (from 0 to 1, with higher values indicating better image quality).

V. DISCUSSION

A. Visual saliency

The cross-attention map in Figure 1 shows that the T2I diffusion model can attend to both high-level semantics and low-level distortions through the text prompt, indicating its reliable prior knowledge. However, it does not explain how the model evaluates perceptual quality or assigns scores. Understanding this process is crucial for assessing whether the fine-tuned model aligns with human visual perception and its sensitivity to noise in small-scale BIQA datasets, which is vital for generalization. To better illustrate the model's behavior, we visualize saliency maps to show which regions influence its decision-making. These saliency maps highlight pixel-level regions that affect the model's output, offering different insights than the cross-attention maps.

As shown in Figure 4, the model tends to focus on complex structures, semantically important objects, and areas with significant color or brightness variation. This aligns with human visual perception, where attention is drawn to high-contrast or semantically meaningful regions. The model's ability to focus on these key features suggests it has learned important cues for IQA tasks. Additionally, we observed no significant overfitting to noise, highlighting the model's strong noise resistance and

TABLE XII
THE AVERAGE TIME SPENT PER IMAGE ON OUR HARDWARE PLATFORM AND THE NUMBER OF PARAMETERS BETWEEN OUR TEACHER AND STUDENT MODEL. BOLD ENTRIES INDICATE THE BEST RESULTS.

| Model | Time (s/image) | Params |
|-------------------|----------------|---------------|
| DP-IQA (teacher) | 0.023 | 1.19B |
| Distilled student | 0.006 | 81.01M |

TABLE XIII
ABLATION STUDY ON THE IMPACT OF PRETRAINED PRIORS AND TRAINING STRATEGIES ON DP-IQA'S PERFORMANCE.

| Trainable Params | Pretrained U-Net | CLIVE | | KonIQ | |
|-------------------|------------------|--------------|--------------|--------------|--------------|
| | | PLCC | SRCC | PLCC | SRCC |
| QFD | ✓ | 0.884 | 0.849 | 0.929 | 0.908 |
| Adapters+QFD | ✓ | 0.890 | 0.829 | 0.935 | 0.917 |
| LoRA+QFD | ✓ | 0.903 | 0.858 | 0.943 | 0.928 |
| Adapters+LoRA+QFD | ✓ | 0.907 | 0.872 | 0.946 | 0.931 |
| Full | ✓ | 0.913 | 0.893 | 0.951 | 0.942 |
| Full | ✗ | 0.551 | 0.569 | 0.786 | 0.784 |

its excellent cross-dataset generalization, which enhances its adaptability and reliability in diverse scenarios.

B. t-SNE Visualization

As illustrated in Figure 5, t-Distributed Stochastic Neighbor Embedding (t-SNE) is employed to visualize the feature embeddings from the final hidden layers of DP-IQA and its corresponding student models, using the complete set of test images from the Koniq-10k dataset as input. A gradient of colors is used to represent the ground truth (GT) scores, facilitating the exploration of the relationship between the learned features and image quality. As a non-linear dimensionality reduction technique, t-SNE projects high-dimensional features into a two-dimensional space, preserving the local structural relationships among data points. If images with similar GT scores are clustered in proximity, it suggests that the model has effectively learned features that are strongly correlated with image quality. The resulting scatter plot exhibits a well-defined, continuous color gradient, which signifies a strong correspondence between the GT scores and the spatial arrange-

ment of the feature embeddings, thereby demonstrating the model's capacity to discern subtle variations in image quality.

C. Limitations

As shown in Table XII, we have distilled the knowledge of DP-IQA into a relatively lightweight model that is both easy to deploy and fast. However, further reducing its size while maintaining prediction accuracy and generalization remains a significant challenge. Further investigation is needed to develop more effective approaches for distilling image-quality-related prior knowledge from teacher models into lightweight student models. Although DP-IQA is designed for in-the-wild scenarios, we also evaluate its performance in synthetic distortion scenarios. We observed that the DP-IQA teacher model may overfit on small-scale synthetic datasets. As synthetic distortion is not the focus of this paper, further details are provided in the supplementary materials.

VI. CONCLUSION

In this paper, we propose a novel BIQA method based on large-scale pre-trained diffusion priors for in-the-wild images, named DP-IQA. It leverages pre-trained SD as the backbone, extracting multi-level features from the denoising U-Net during the upsampling process at a specific timestep and decoding them to estimate image quality, without requiring a diffusion process. To alleviate the computational burden of diffusion models in practical applications, we distill the knowledge from DP-IQA into a smaller EfficientNet-based model. Experimental results show that DP-IQA achieves SOTA on various in-the-wild datasets and demonstrates the best generalization capabilities. We believe our exploration can provide a new technical direction for future works and inspire future efforts to more effectively leverage diffusion priors for better assessment of image perceptual quality.

REFERENCES

- [1] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *IEEE Trans. Image Process.*, vol. 31, pp. 4149–4161, 2022.
- [2] T.-Y. Chiu, Y. Zhao, and D. Gurari, "Assessing image quality issues for real-world problems," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3646–3656.
- [3] A. Saha, S. Mishra, and A. C. Bovik, "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5846–5855.
- [4] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.
- [5] T. Xiang, Y. Yang, and S. Guo, "Blind night-time image quality assessment: Subjective and objective approaches," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1259–1272, 2019.
- [6] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5148–5157.
- [7] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1220–1230.
- [8] G. Qin, R. Hu, Y. Liu, X. Zheng, H. Liu, X. Li, and Y. Zhang, "Data-efficient image quality assessment with attention-panel decoder," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 2091–2100.
- [9] K. Xu, L. Liao, J. Xiao, C. Chen, H. Wu, Q. Yan, and W. Lin, "Boosting image quality assessment through efficient transformer adaptation with local feature enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 2662–2672.
- [10] L. Agnolucci, L. Galteri, M. Bertini, and A. Del Bimbo, "Arniqa: Learning distortion manifold for image quality assessment," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 189–198.
- [11] K. Zhao, K. Yuan, M. Sun, M. Li, and X. Wen, "Quality-aware pre-trained models for blind image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22302–22313.
- [12] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14071–14081.
- [13] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 2555–2563.
- [14] F. Peng, H. Fu, A. Ming, C. Wang, H. Ma, S. He, Z. Dou, and S. Chen, "Aigc image quality assessment via image-prompt correspondence," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops*, vol. 6, 2024.
- [15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8748–8763.
- [16] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Controlling vision-language models for multi-task image restoration," in *12th Int. Conf. Learn. Represent.*, 2023.
- [17] X. Li, J. Zheng, R. Hu, Y. Zhang, K. Li, Y. Shen, X. Zheng, Y. Liu, S. Zhang, P. Dai *et al.*, "Feature denoising diffusion model for blind image quality assessment," *arXiv preprint arXiv:2401.11949*, 2024.
- [18] Z. Wang, B. Hu, M. Zhang, J. Li, L. Li, M. Gong, and X. Gao, "Diffusion model based visual compensation guidance and visual difference analysis for no-reference image quality assessment," *arXiv preprint arXiv:2402.14401*, 2024.
- [19] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, "Your diffusion model is secretly a zero-shot classifier," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2206–2217.
- [20] J. Tian, L. Aggarwal, A. Colaco, Z. Kira, and M. Gonzalez-Franco, "Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion," *arXiv preprint arXiv:2308.12469*, 2023.
- [21] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 5729–5739.
- [22] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *arXiv preprint arXiv:2305.07015*, 2023.
- [23] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai, "Generative diffusion prior for unified image restoration and enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9935–9946.
- [24] L. Guo, C. Wang, W. Yang, S. Huang, Y. Wang, H. Pfister, and B. Wen, "Shadowdiffusion: When degradation prior meets diffusion model for shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14049–14058.
- [25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
- [26] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4296–4304.
- [27] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, 2010.
- [28] ———, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, 2011.
- [29] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, pp. 32–32, 2017.
- [30] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, 2014.

- [31] M. A. Saad, A. C. Bovik, and C. Charrier, "A dct statistics-based blind image quality index," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 583–586, 2010.
- [32] A. Sadiq, I. F. Nizami, S. M. Anwar, and M. Majid, "Blind image quality assessment using natural scene statistics of stationary wavelet transform," *Optik*, vol. 205, p. 164189, 2020.
- [33] L. Xu, J. Li, W. Lin, Y. Zhang, L. Ma, Y. Fang, and Y. Yan, "Multi-task rank learning for image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 9, pp. 1833–1843, 2016.
- [34] D. Ghadiyaram and A. C. Bovik, "Blind image quality assessment on real distorted images using deep belief nets," in *2014 IEEE Global Conf. Signal Inf. Process.* IEEE, 2014, pp. 946–950.
- [35] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1733–1740.
- [36] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3575–3585.
- [37] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, 2018.
- [38] H. Chen, X. Chai, F. Shao, X. Wang, Q. Jiang, X. Meng, and Y.-S. Ho, "Perceptual quality assessment of cartoon images," *IEEE Trans. Multimedia*, vol. 25, pp. 140–153, 2021.
- [39] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiq: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14 143–14 152.
- [40] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, 2017.
- [41] D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu, and Y. Zhang, "Blind predicting similar quality map for image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6373–6382.
- [42] J. You and J. Korhonen, "Transformer for image quality assessment," in *2021 IEEE Int. Conf. Image Process.* IEEE, 2021, pp. 1389–1393.
- [43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [44] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, 2017.
- [45] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal Image Video Process.*, vol. 12, pp. 355–362, 2018.
- [46] F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognit.*, vol. 81, pp. 432–442, 2018.
- [47] D. Varga, D. Saupe, and T. Szirányi, "Deeprn: A content preserving deep architecture for blind image quality assessment," in *2018 IEEE Int. Conf. Multimedia Expo.* IEEE, 2018, pp. 1–6.
- [48] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "Graphiq: Learning distortion graph representations for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 25, pp. 2912–2925, 2022.
- [49] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3667–3676.
- [50] X. Yang, F. Li, and H. Liu, "Ttl-iqa: Transitive transfer learning based no-reference image quality assessment," *IEEE Trans. Multimedia*, vol. 23, pp. 4326–4340, 2020.
- [51] B. Hu, G. Zhu, L. Li, J. Gan, W. Li, and X. Gao, "Blind image quality index with cross-domain interaction and cross-scale integration," *IEEE Trans. Multimedia*, 2023.
- [52] N. C. Babu, V. Kannan, and R. Soundararajan, "No reference opinion unaware quality assessment of authentically distorted images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2459–2468.
- [53] K.-Y. Lin and G. Wang, "Hallucinated-iqa: No-reference image quality assessment via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 732–741.
- [54] Y. Zhu, H. Ma, J. Peng, D. Liu, and Z. Xiong, "Recycling discriminator: Towards opinion-unaware image quality assessment using wasserstein gan," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 116–125.
- [55] H. Ren, D. Chen, and Y. Wang, "Ran4iq: Restorative adversarial nets for no-reference image quality assessment," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
- [56] S. Srinath, S. Mitra, S. Rao, and R. 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*, 2024, pp. 22–31.
- [57] D. De, S. Mitra, and R. Soundararajan, "Genziqa: Generalized image quality assessment using prompt-guided latent diffusion models," *arXiv preprint arXiv:2406.04654*, 2024.
- [58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [59] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," *arXiv preprint arXiv:2312.02145*, 2023.
- [60] J. Xiao, R. Feng, H. Zhang, Z. Liu, Z. Yang, Y. Zhu, X. Fu, K. Zhu, Y. Liu, and Z.-J. Zha, "Dreamclean: Restoring clean image using deep diffusion prior," in *12th Int. Conf. Learn. Represent.*
- [61] X. Chen, Z. Liu, S. Xie, and K. He, "Deconstructing denoising diffusion models for self-supervised learning," *arXiv preprint arXiv:2401.14404*, 2024.
- [62] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," *arXiv preprint arXiv:2112.03126*, 2021.
- [63] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.
- [64] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. Mach. Learn.* PMLR, 2019, pp. 6105–6114.
- [65] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai *et al.*, "Q-bench: A benchmark for general-purpose foundation models on low-level vision," in *The Twelfth International Conference on Learning Representations*.
- [66] Z. Zhang, H. Wu, E. Zhang, G. Zhai, and W. Lin, "Q-bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [67] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4296–4304.
- [68] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.
- [69] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [70] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *Int. J. Comput. Vis.*, vol. 132, no. 2, pp. 581–595, 2024.
- [71] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," *arXiv preprint arXiv:1611.02731*, 2016.
- [72] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao, and C. Dong, "Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 669–25 680.
- [73] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao, and C. Dong, "Diffbir: Towards blind image restoration with generative diffusion prior," *arXiv preprint arXiv:2308.15070*, 2023.
- [74] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, 2015.
- [75] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3677–3686.
- [76] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," in *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI* 16. Springer, 2020, pp. 633–651.
- [77] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.

- [78] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, pp. 011 006–011 006, 2010.
- [79] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqo database," in *11th Int. Conf. Quality Multimedia Experience*. IEEE, 2019, pp. 1–3.
- [80] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, 2012.
- [81] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [82] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, 2015.
- [83] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 1, pp. 206–220, 2016.
- [84] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, 2017.
- [85] J. Shi, P. Gao, and J. Qin, "Transformer-based no-reference image quality assessment via supervised contrastive learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4829–4837.
- [86] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun *et al.*, "Q-align: teaching lmms for visual scoring via discrete text-defined levels," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 54 015–54 029.
- [87] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
- [88] M. Oquab, T. Darcret, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.
- [89] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.