Toward Generalized Image Quality Assessment: Relaxing the Perfect Reference Quality Assumption

Du Chen[1,3,*], Tianhe Wu[2,3,*], Kede Ma[2,†], and Lei Zhang[1,3,†]
[1]The Hong Kong Polytechnic University [2]City University of Hong Kong [3]OPPO Research Institute
csdud.chen@connet.polyu.hk, {tianhewu, kede.ma}@cityu.edu.hk, cslzhang@comp.polyu.edu.hk

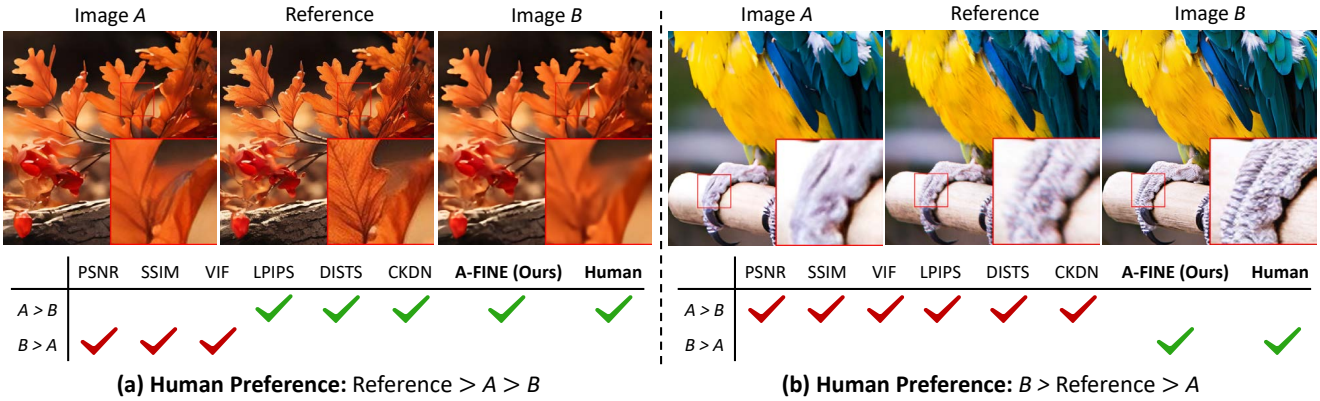

Fig. 1. *With the reference image in the middle, which image, A or B, has better perceived visual quality*? The proposed A-FINE generalizes and outperforms standard FR-IQA models under both perfect and imperfect reference conditions. Zoom in for better visibility.

Abstract

Full-reference image quality assessment (FR-IQA) generally assumes that reference images are of perfect quality. However, this assumption is flawed due to the sensor and optical limitations of modern imaging systems. Moreover, recent generative enhancement methods are capable of producing images of higher quality than their original. All of these challenge the effectiveness and applicability of current FR-IQA models. To relax the assumption of perfect reference image quality, we build a large-scale IQA database, namely DiffIQA, containing approximately $180,000$ *images generated by a diffusion-based image enhancer with adjustable hyper-parameters. Each image is annotated by human subjects as either worse, similar, or better quality compared to its reference. Building on this, we present a generalized FR-IQA model, namely* ***A****daptive* ***FI****delity-****N****aturalness* ***E****valuator (A-FINE), to accurately assess and adaptively combine the fidelity and naturalness of a test image. A-FINE aligns well with standard FR-IQA when the reference image is much more natural than the test image. We demonstrate by extensive experiments that A-FINE surpasses standard FR-IQA models on well-established IQA datasets and our newly created DiffIQA. To further validate A-FINE, we additionally construct a super-resolution IQA benchmark (SRIQA-Bench), encompassing test images derived from ten state-of-the-art SR methods with reliable human quality annotations. Tests on SRIQA-Bench re-affirm the advantages of A-FINE. The code and dataset are available at* https://tianhewu.github.io/A-FINE-page.github.io/.

1. Introduction

Image Quality Assessment (IQA) plays an indispensable role in the digital image lifecycle, from acquisition, transmission, and reproduction, to storage [42]. The objective of IQA is to develop computational models that mimic the Human Visual System (HVS) in perceiving image quality [45], which can be broadly classified into two categories based on the availability of reference images: Full-Reference IQA (FR-IQA) [5, 6, 44, 45, 55, 58] and No-Reference IQA (NR-IQA) [25, 43, 48, 51, 60, 61]. FR-IQA evaluates a test image by comparing it to a reference image, which is assumed to be of perfect quality, while NR-IQA assesses the test image quality without needing the reference.

Over the past two decades, FR-IQA has experienced significant progress, with the paradigm shifted from measuring

*Equal contribution.
†Corresponding authors.

(a) Reference image (b) Enhanced image

Fig. 2. **(a)** Reference image from CSIQ [15] and **(b)** its corresponding enhanced image by a recent generation-based image enhancer, SeeSR [47].

error visibility [4] to assessing structural similarity [45], and more recently, to unifying structural and textural similarity [5]. This evolution has led to the development of several representative methods, including SSIM [45], FSIM [55], LPIPS [58], and DISTS [5]. These FR-IQA models have been rapidly adopted as standard evaluation criteria, alongside the traditional peak signal-to-noise ratio (PSNR) [42], for measuring the progress in various image processing tasks. Moreover, there is a growing trend of employing these models as loss functions for perceptual optimization of image restoration algorithms based on deep neural networks (DNNs) [3, 7, 16, 19, 33, 59].

Most FR-IQA models operate under the assumption that the reference image is of perfect quality. However, this assumption is problematic as digital imaging systems face practical hardware and software limitations, making it extremely difficult (if not impossible) to capture perfect-quality images. This is particularly true for natural scenes that exhibit great spatiotemporal complexity, high dynamic range, and wide color spectrum. As a result, many reference images in existing IQA datasets are of subpar quality. Moreover, the image quality could be rescued and even improved using modern generative image enhancement techniques [32, 36, 46, 49, 52] (see Fig. 2 for a visual example).

The violation of the perfect reference quality assumption undermines the reliability and applicability of standard FR-IQA models in providing useful quality estimates. Fig. 3 shows a motivating example. We embed images in a perceptually uniform space, where the perceived quality of a test image is computed by its Euclidean distance to the perfect-quality image (*i.e.*, *Image A*). When the reference *Image D* of non-perfect quality is used, it is inherently incapable of assessing *Images B* and *C* with higher quality. Additionally, it may also struggle to accurately evaluate *Images E* and *F* of the same worse quality (*i.e.*, lying on the same level set). In this case, standard FR-IQA models may be biased toward *Image E*, which is closer to the imperfect reference.

Very limited subjective testing [29] and objective modeling [62] studies have been reported on the quality assessment of enhanced images under imperfect reference conditions. These studies are becoming increasingly obsolete as they primarily focused on simple and synthetic scenarios (*e.g.*, image interpolation [50] and Gaussian image denoising [53]), and the employed enhancers frequently fail to yield images with improved perceptual quality.

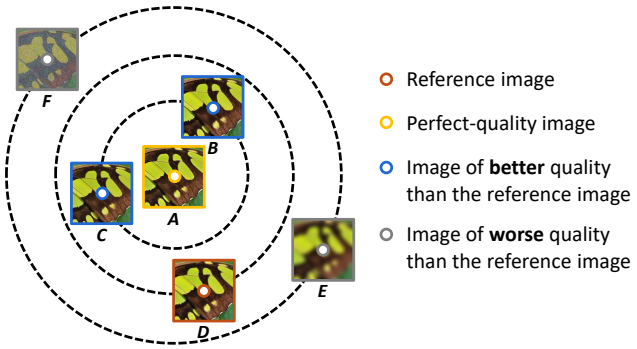

Fig. 3. Standard FR-IQA models tend to fail when the reference image is of non-perfect quality. In this visualization, images are embedded in a perceptually uniform space, where the perceived quality of a test image is described by its Euclidean distance to the perfect-quality image. Images located on the same dashed circles are perceived to have identical visual quality.

To relax the perfect reference quality assumption, we first establish a large-scale IQA database, named DiffIQA, which comprises approximately $180,000$ images with *worse*, *similar*, and *better* quality relative to their corresponding references. DiffIQA is generated by adapting the recent pixel-aware stable diffusion (PASD) method [49] into a powerful image enhancer, while also adjusting its hyper-parameters and perturbing the input reference images. We then invite human subjects to categorize each image as having worse, similar, or better perceptual quality compared to its reference using incomplete paired comparison. Moreover, we present a generalized FR-IQA model by adaptively weighting a DISTS-like image fidelity term and an image naturalness term, both of which share the same feature extraction backbone. The resulting **A**daptive **FI**delity-**N**aturalness **E**valuator (A-FINE) can be end-to-end optimized, and gracefully reverted to standard FR-IQA models when the reference image is much more natural than the test image. To further evaluate A-FINE, we construct an SR-based IQA benchmark, named SRIQA-Bench, comprising $1,000$ images generated by ten SR methods and annotated using complete paired comparison.

In summary, the contributions of this paper include

- A large-scale IQA database, DiffIQA, breaking the perfect reference quality assumption;
- A generalized FR-IQA model, A-FINE, outperforming existing methods under both perfect and imperfect reference conditions;
- An extensive experimental demonstration on the effectiveness of A-FINE on standard IQA datasets [10, 15, 20, 26], and our newly created DiffIQA and SRIQA-Bench.

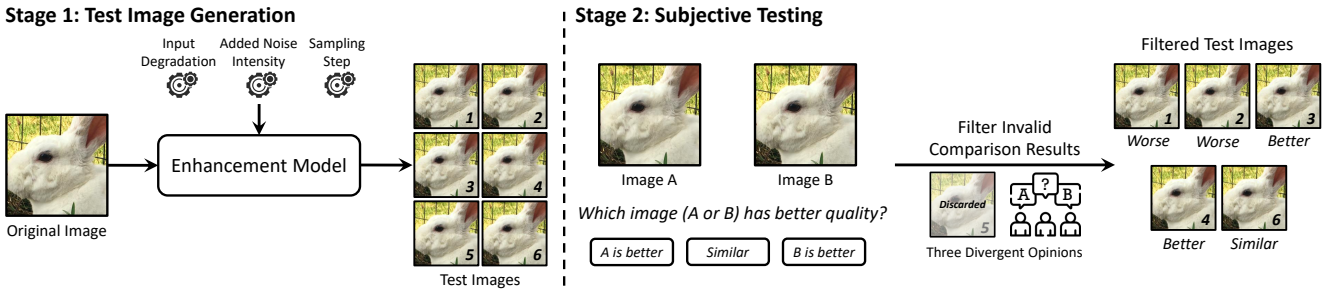

Fig. 4. DiffIQA is constructed in two stages. In Stage 1, we adapt PASD [49] to a generative image enhancer (see the Appendix for more details) to produce images of varying perceptual quality, some of which are perceived better than the original. In Stage 2, we conduct subjective experiments using incomplete paired comparison, followed by raw subjective data filtering.

2. Related Work

FR-IQA Datasets. The creation of FR-IQA datasets generally starts by selecting a set of reference images of "perfect" quality, to which multiple synthetic distortions at various intensity levels are applied. Subjective testing is then conducted to gather mean opinion scores (MOSs) as the ground-truth quality annotations for the distorted images.

The LIVE dataset [30] is the first successful public-domain IQA dataset, containing 29 reference images and five types of distortions, annotated using a single-stimulus continuous quality rating method. CSIQ [15] maintains a similar dataset size but enhances annotation efficiency using multi-stimulus continuous quality rating. TID2013 [26] extends the distortion scope to 25 types, and utilizes an incomplete pairwise comparison method based on the Swiss tournament system, constrained to pairs of the same underlying visual content. The KADID-10K [20] dataset has 81 reference images, yielding $10,125$ distorted images rated using double-stimulus absolute category rating on a crowdsourcing platform. The Waterloo Exploration Database [22] expands the number of reference images to $4,744$, and introduces three computational tests for evaluating IQA models without reliance on subjective experimentation. BAPPS [58] and PIPAL [10] broaden the scope of synthetic distortion scenarios by incorporating algorithm-dependent distortions from DNN-based image restoration and enhancement algorithms (see Table 1).

Standard FR-IQA for Distorted Images. Mean squared error (MSE, along with its derivative PSNR) and mean absolute error (MAE) have been dominantly used, yet they fail to match human perception of visual quality. Within this error visibility paradigm, various remedies have been proposed, including VDP [4] and its HDR extensions [23, 24], MAD [15], and LPIPS [58]. A major paradigm shift occurred in 2004 with the introduction of SSIM [45], which prioritizes structural similarity over error visibility. SSIM has been extended for multiscale processing [44] and transformed/feature domain analysis [55]. Leveraging pretrained DNN-based features, DISTS [5] and its locally adaptive version [6] unify the structural and textural similarity. Standard FR-IQA models rely on comparing the test image against a perfect-quality reference image, and thus they fall short in quality assessment of enhanced images.

Generalized FR-IQA for Enhanced Images. Although not initially intended for this purpose, the information-theoretic VIF [29] is one of the first FR-IQA methods to handle cases when the test image visually outperforms the reference. PCQI [37] takes a structural similarity approach, and gives credit to image patches with improved local contrast. Yeganeh and Wang [50] leveraged the low-resolution image to evaluate interpolated image quality based on a natural scene statistical model, while Zhang *et al*. [53] used the noisy image to predict denoised image quality through empirical Bayes estimation. CKDN [62] aligns the degraded image with the reference in feature space, allowing its features to act as a reference proxy for assessing restored images. These models, developed and tested under simplistic, constrained scenarios, tend to struggle on the proposed DiffIQA and SRIQA-bench, which feature test images with improved quality compared to their corresponding references. The proposed A-FINE is designed as a generalized FR-IQA model, which can be end-to-end optimized to perform well under both perfect and imperfect reference conditions.

3. Proposed Dataset: DiffIQA

This section describes the construction of DiffIQA, including test image generation by our diffusion-based image enhancer and subjective testing for collecting quality annotations, as illustrated in Fig. 4.

3.1. Generative Image Enhancer

To generate test images with diverse quality, we adapt the PASD method [49], which is initially designed for realistic single-image SR and personalized stylization, into a generative image enhancer. Specifically, we feed the input image to a lightweight convolutional network to generate the control signal for the ControlNet [57], and employ the pretrained Stable-Diffusion [28] as the backbone to enhance

Table 1. Comparison of DiffIQA and SRIQA-Bench against existing representative FR-IQA datasets.

Dataset	# of Ref. Images	# of Test Images	Distortion / Enhancement Type	Image Resolution	# of Human Annotations	Perfect Reference Quality Assumption
LIVE [30]	29	779	Simulated	480×720 to 768×512	25k	Necessary
CSIQ [15]	30	866	Simulated	512×512	25k	Necessary
TID2013 [26]	25	3k	Simulated	512×384	500k	Necessary
KADID-10K [20]	81	10.1k	Simulated	512×384	303.8k	Necessary
PIPAL [10]	250	29k	Simulated / GAN-based	288×288	1.1m	Necessary
BAPPS [58]	187.7k	375.4k	Simulated / DNN-based	64×64	484.3k	Necessary
DiffIQA (Ours)	29.9k	177.3k	Diffusion-based	512×512	537.6k	Not Necessary
SRIQA-Bench (Ours)	100	1.1k	DNN- / GAN- / Diffusion-based	512×512	55k	Not Necessary

the image. In the backward diffusion process, we incorporate pixel-aware cross-attention [49] to facilitate interactions between generative features in diffusion UNet and the control features from ControlNet. The proposed enhancer is trained on the widely-adopted DF2K_OST [38, 40], DIV8K [11], FFHQ [13], and LSDIR [17] datasets. To diversify output image quality, half of the input images are subject to slight blind degradations [40], allowing the enhancer to produce outputs with *worse*, *similar*, and *better* visual quality compared to the original. The trainable components of our enhancer—the lightweight convolutional network and the pixel-aware cross-attention modules—are optimized to predict the noise added to the input latent. More details regarding the network architecture and training procedure are presented in the Appendix.

3.2. Construction of DiffIQA

Test Image Generation. We gathered original input images from three sources: 1) $1,200$ from the DF2K dataset [40]; 2) $1,000$ from the Internet under the license of Creative Commons; and 3) 640 captured using mobile phones or digital cameras. These were cropped to 512×512 with an overlap of less than 128 pixels, leading to a total of $29,868$ images as inputs to our trained generative enhancer. During inference, we randomly 1) applied the same degradations as used during training, 2) augmented the initial image latent with additive Gaussian noise of varying intensities, and 3) adjusted the sampling steps within range $[20, 1000]$ to generate images with diverse quality levels. We produced six test images for each input, totaling $179,208$ images. Additional details are provided in the Appendix.

Subjective Testing. We employed an incomplete paired comparison method, where subjects were shown a reference image alongside a test image of the same visual content in random spatial order. They were asked to infer the relative quality of the two images by choosing one from three options: the left image is of *worse*, *similar*, *better* perceived quality compared to the right one. A diverse group of 240 subjects, including 132 males and 108 females aged between 18 and 42, contributed to this study. A total of $179,208$ image pairs were evaluated, with each subject assigned $2,240$ pairs, organized into multiple 30-minute sessions to mitigate visual fatigue. All subjects completed the assigned sessions within two weeks, and the entire subjective testing spanned four months. Each image pair was rated by a minimum of three annotators, and the average time taken for each comparison was about 3.40 seconds. More details regarding the subjective experimental setups can be found in the Appendix.

3.3. DiffIQA Statistics

We gathered a total of $537,624$ quality annotations, with $232,285$ (43.20%) labeled as *worse*, $85,671$ (15.94%) as *similar*, and $219,668$ (40.86%) as *better* compared to the reference. This verifies the capability of our generative enhancer to improve the perceived quality of original input images. The potential labeling discrepancies among the three subjects were resolved by majority voting. When there is a tie (*i.e.*, one *worse*, one *similar* and one *better* vote), the image pair is marked as outlier and removed. As a result, we discard $1,889$ (1.05%) invalid annotations, leading to $76,515$ (42.70%) *worse*, $24,654$ (13.76%) *similar*, and $76,150$ (42.49%) *better* quality labels.

As summarized in Table 1, the proposed DiffIQA dataset possesses three unique features that distinguish it from existing FR-IQA datasets. First, it is large-scale in terms of the number of test images (at the standard resolution of 512×512) and the number of quality annotations. Second, it features diffusion-based distortions, which exhibit visual characteristics distinct from those produced by regression-based or GAN-based image restoration and enhancement methods. Third, it is generalized to include test images with better quality than their references, relaxing the perfect reference quality assumption.

4. Proposed FR-IQA Model: A-FINE

In this section, we introduce the design of A-FINE as a generalized FR-IQA model, followed by a detailed description of its training procedure. The overall system diagram of A-FINE is illustrated in Fig. 5.

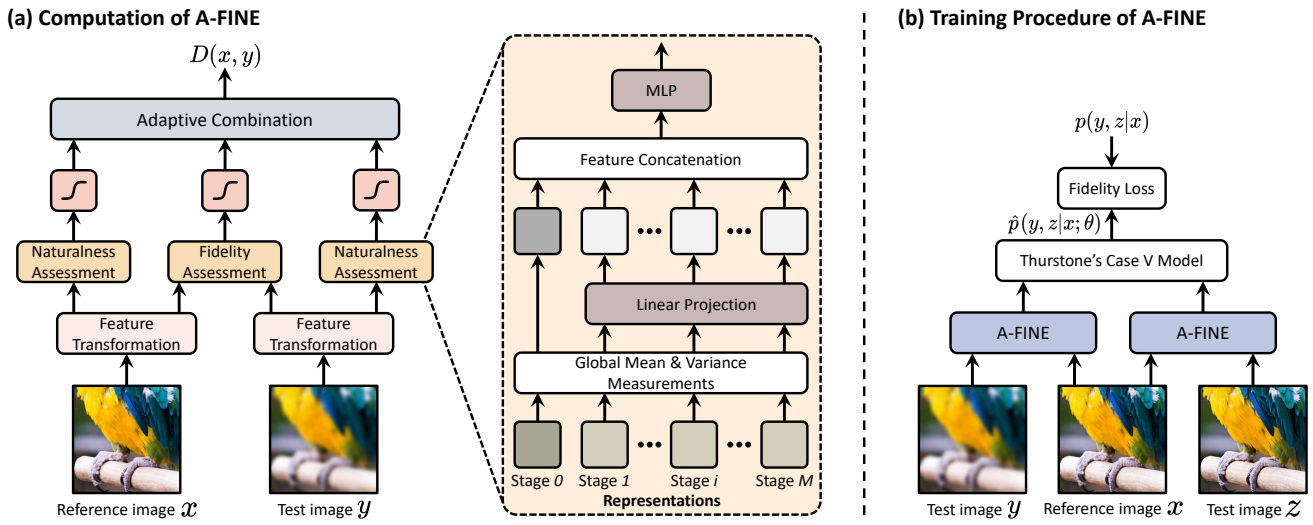

Fig. 5. System diagram of the proposed A-FINE and its pairwise learning-to-rank training procedure. A-FINE leverages a shared feature transformation to make image fidelity and naturalness measurements, which are adaptively combined to produce the final quality score.

4.1. Computation of A-FINE

Given a reference image $x \in \mathbb{R}^{H\times W\times 3}$, which may be of lower quality than the test image $y \in \mathbb{R}^{H\times W\times 3}$, we aim to learn a generalized FR-IQA model, $D(x,y)$, to evaluate the perceptual quality of y relative to x without the perfect reference quality assumption. Drawing inspiration from the maximum a posterior (MAP) estimation, A-FINE is designed as an adaptive linear combination of an image fidelity term, $F(x,y)$, and an image prior term, $N(y)$:

$$D(x,y) = F(x,y) + \lambda(x,y)N(y), \tag{1}$$

where $\lambda(x,y) \geq 0$ serves as the adaptive weighting function. The value of $D(x,y)$ can either be positive or negative, with a smaller value indicating better-predicted quality of y. Correspondingly, smaller values of $F(x,y)$ and $N(y)$ denote better-predicted fidelity and naturalness.

Intuitively, $\lambda(x,y)$ can be designed as a function reflecting the relative naturalness of the two images. If the naturalness score of the reference image, $N(x)$, is significantly lower than that of the test image, $N(y)$, the term $F(x,y)$ should largely drive the quality prediction, thereby reducing $D(x,y)$ to be a standard FR-IQA model. Conversely, when the test image y appears more natural than x, as in the case when x is degraded, $D(x,y)$ should depend more on the naturalness assessment of the test image itself, *i.e.*, $N(y)$. Thus, we define $\lambda(x,y)$ as

$$\lambda(x,y) = \exp\left(k(N(x) - N(y)\right), \tag{2}$$

where $k \geq 0$ is a learnable scale parameter.

To instantiate the image fidelity term $F(x,y)$, we employ a DISTS-like [5, 6] approach:

$$F(x,y) = 1 - \sum_{i=0}^{M}\sum_{j=1}^{N_i} F\left(x_j^{(i)}, y_j^{(i)}\right), \tag{3}$$

where

$$F\left(x_j^{(i)}, y_j^{(i)}\right) = \alpha_{ij} L\left(x_j^{(i)}, y_j^{(i)}\right) + \beta_{ij} S\left(x_j^{(i)}, y_j^{(i)}\right). \tag{4}$$

Here, $x_j^{(i)}$ and $y_j^{(i)}$ represent the feature maps extracted from the j-th channel of the i-th stage of the backbone network, corresponding to x and y, respectively. M and N_i denote the total number of stages and the number of channels in the i-th stage, respectively. $L\left(x_j^{(i)}, y_j^{(i)}\right)$ and $S\left(x_j^{(i)}, y_j^{(i)}\right)$ measure the global texture and structure similarity [45], respectively:

$$L(x_j^{(i)}, y_j^{(i)}) = \frac{2\mu_{x_j}^{(i)}\mu_{y_j}^{(i)} + c_1}{\left(\mu_{x_j}^{(i)}\right)^2 + \left(\mu_{y_j}^{(i)}\right)^2 + c_1}, \tag{5}$$

$$S(x_j^{(i)}, y_j^{(i)}) = \frac{2\sigma_{x_j y_j}^{(i)} + c_2}{\left(\sigma_{x_j}^{(i)}\right)^2 + \left(\sigma_{y_j}^{(i)}\right)^2 + c_2}, \tag{6}$$

where $\mu_{x_j}^{(i)}$, $\mu_{y_j}^{(i)}$, $(\sigma_{x_j}^{(i)})^2$, $(\sigma_{y_j}^{(i)})^2$, and $\sigma_{x_j y_j}^{(i)}$ denote the global means and variances of $x_j^{(i)}$ and $y_j^{(i)}$, as well as the global covariance between $x_j^{(i)}$ and $y_j^{(i)}$, respectively. c_1 and c_2 are two small positive constants to prevent numerical instability when the denominators approach zero. The

weights $\{\alpha_{ij}, \beta_{ij}\}$ are positive and learnable, satisfying the constraint $\sum_{i=0}^{M} \sum_{j=1}^{N_i} (\alpha_{ij} + \beta_{ij}) = 1$.

In contrast to DISTS [5], which uses a fixed, pre-trained VGG network [31] for feature representation, A-FINE adopts a more advanced Vision Transformer (ViT) as the backbone, specifically the CLIP ViT-B/32@224 [27]. We interpolate position embeddings to accommodate images of arbitrary resolutions. Additionally, we fine-tune all backbone parameters, denoted as ϕ, along with the linear weights in Eq. (4), exploiting the transferability of ViT features to improve quality prediction.

To instantiate the image naturalness term $N(\cdot)$, parameterized by φ, we reuse the CLIP ViT backbone for computing $F(x, y)$. Global average and variance pooling are applied to the feature maps at each stage, resulting in a stage-wise feature vector of size 768×2. This vector is then linearly projected into a 128-dimensional space using a shared projection matrix across all stages. The projected multi-stage features are concatenated and fed to a multilayer perceptron (MLP), composed of two fully connected layers and a Gaussian error linear unit (GELU) activation in between, with dimensions $3 \times 2 + 128 \times 12 \to 128 \times 6 \to 1$, to compute the naturalness score[1].

To stabilize training, we follow [30] and incorporate a separate four-parameter monotonic logistic function for both $F(x, y)$ and $N(y)$ as part of our model computation:

$$F_\eta(x, y) = \frac{\eta_1 - \eta_2}{1 + \exp\left(-\frac{F(x,y)-\eta_3}{|\eta_4|}\right)} + \eta_2 \tag{7}$$

and

$$N_\gamma(y) = \frac{\gamma_1 - \gamma_2}{1 + \exp\left(-\frac{N(y)-\gamma_3}{|\gamma_4|}\right)} + \gamma_2, \tag{8}$$

where η_1 and γ_1 are set to 2, and η_2 and γ_2 are set to -2, respectively, defining the upper and lower bounds of the non-linear mappings. The learnable parameters constitute $\{\eta_3, \eta_4, \gamma_3, \gamma_4\}$.

4.2. Training Procedure of A-FINE

Inspired by UNIQUE [60], we adopt a similar pairwise learning-to-rank approach to optimize the parameters in A-FINE, collectively denoted as $\theta = \{\phi, \varphi, \alpha, \beta, k, \eta, \gamma\}$. In particular, given a triplet (x, y, z), where x is the reference image, and y and z are two test images with the same underlying visual content as x, we derive the ground-truth ranking label based on the relative quality of y and z:

$$p(y, z|x) = \begin{cases} 1 & \text{if } Q(y|x) > Q(z|x) \\ 0.5 & \text{if } Q(y|x) = Q(z|x) \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

[1]The term "3×2" corresponds to the global means and variances computed from the three color channels of the input test image.

where $Q(y|x)$ and $Q(z|x)$ represent the MOS of y and z relative to x, respectively. Under Thurstone's Case V model [34], we assume that the perceptual quality of a test image follows a Gaussian distribution, where the mean is estimated by the proposed A-FINE, and the variance is fixed to one. This enables us to compute the probability that y is perceived better than z given x by:

$$\hat{p}(y, z|x; \theta) = \Phi\left(\frac{D(x, y; \theta) - D(x, z; \theta)}{\sqrt{2}}\right), \tag{10}$$

where $\Phi(\cdot)$ represents the standard Gaussian cumulative distribution function. Following [60], we adopt the fidelity loss [35] for end-to-end optimization:

$$\begin{aligned} \ell(y, z|x; \theta) = 1 &- \sqrt{p(y, z|x)\hat{p}(y, z|x; \theta)} \\ &- \sqrt{(1 - p(y, z|x))(1 - \hat{p}(y, z|x; \theta))}. \end{aligned} \tag{11}$$

5. Experiments

In this section, we first describe the experimental setups, followed by the construction of SRIQA-Bench. Then, we compare A-FINE against several existing FR-IQA models across various IQA datasets, including the proposed DiffIQA and SRIQA-Bench.

5.1. Experimental Setups

Training Details of A-FINE. Since our DiffIQA dataset is designed for scenarios where test images can have equal or higher quality than the reference, we combine it with TID2013 [26], KADID-10K [20], and PIPAL [10] to enhance the training of A-FINE. Following standard practice, we partition each dataset into training, validation, and test sets in a 7:1:2 ratio, ensuring content independence. The training of A-FINE proceeds in three phases. In Phase 1, we perform a warm-up training for the image naturalness term $N(\cdot)$, in which we fine-tune the ViT backbone parameters ϕ, and train the linear projection matrix and the MLP prediction head, with parameters φ. In Phase 2, the fine-tuned ViT backbone is frozen, and the training focuses on optimizing the linear weights $\{\alpha_{ij}, \beta_{ij}\}$ associated with the fidelity term $F(\cdot, \cdot)$. In Phase 3, the complete A-FINE model is refined by optimizing the scale parameter k in the adaptive weighting function $\lambda(\cdot, \cdot)$ and the parameters $\{\eta_i, \gamma_i\}$ in the two nonlinear mappings, while keeping all other fixed.

Training is carried out by employing AdamW [21] as the optimizer, with a weight decay factor of 10^{-3} and initial learning rates of 5×10^{-6} for Phase 1, 5×10^{-4} for Phase 2 and 1×10^{-3} for Phase 3, subject to cosine annealing scheduling with a period of $10,000$ iterations. The training minibatch size on a single GPU is set to 128, and we train A-FINE on four NVIDIA V100 GPUs. The maximum number of training iterations are set to $40,000$, $40,000$, and $10,000$ for Phases 1, 2, and 3, respectively.

Table 2. Accuracy (%) results of FR-IQA models on the test sets of TID2013 [26], KADID-10K [20], PIPAL [10], and the proposed DiffIQA. The term "Combined" indicates the combination of TID2013, KADID, PIPAL, and DiffIQA. The suffix "-FT" means that the model is fine-tuned on this combined dataset. The top two results are highlighted in **bold** and underlined, respectively.

Scenario	Method	Training Dataset	TID2013	KADID	PIPAL	Average	DiffIQA Ref < Test	DiffIQA Ref > Test	DiffIQA Average	All Average
Standard	PSNR	N.A.	75.8	74.8	70.7	72.2	18.2	92.1	45.6	58.9
	SSIM [45]	N.A.	68.9	74.0	72.1	72.4	20.1	93.0	47.1	60.0
	MS-SSIM [44]	N.A.	83.4	81.8	72.5	75.9	20.1	93.0	47.1	61.5
	FSIM [55]	N.A.	86.0	83.4	76.2	79.0	20.2	93.1	47.2	63.1
	VSI [56]	N.A.	87.3	84.8	76.2	79.5	19.7	93.1	46.9	63.2
	LPIPS [58]	BAPPS	78.7	77.0	74.3	75.4	23.7	94.7	50.0	62.7
	LPIPS-FT	Combined	72.5	78.2	71.7	73.6	35.4	91.6	55.6	64.6
	DISTS [5]	KADID	78.4	81.4	75.3	77.2	21.4	94.8	48.6	62.9
	DISTS-FT	Combined	78.4	81.9	72.1	75.3	38.2	89.5	56.7	66.0
	AHIQ [14]	PIPAL	74.6	76.4	79.3	78.1	34.1	88.1	54.1	66.1
	AHIQ-FT	Combined	81.0	79.7	74.9	76.7	78.4	73.8	76.7	76.7
	TOPIQ [1]	KADID	**90.4**	**94.3**	80.5	**85.1**	22.1	**95.1**	49.1	67.1
	TOPIQ-FT	Combined	78.9	85.0	79.0	80.6	78.6	74.2	77.0	78.8
Generalized	VIF [29]	N.A.	78.5	75.2	72.4	73.7	20.0	92.8	46.9	60.3
	PCQI [37]	N.A.	66.6	65.4	56.7	59.9	17.3	90.3	44.3	52.1
	SFSN [63]	N.A.	75.6	70.5	69.8	70.5	19.5	89.6	45.4	58.0
	CKDN [62]	PIPAL	76.9	70.9	79.8	77.1	33.3	82.4	51.4	64.3
	CKDN-FT	Combined	75.0	80.1	68.1	72.0	**79.4**	71.0	76.4	74.2
	A-FINE (Ours)	Combined	88.1	88.3	**81.0**	83.6	78.5	82.3	**79.9**	**81.8**

Competing FR-IQA Models. We compare A-FINE against nine standard FR-IQA models: 1) PSNR, 2) SSIM [45], 3) MS-SSIM [44], 4) FSIM [55], 5) VSI [56], 6) LPIPS [58], 7) DISTS [5], 8) AHIQ [14] and 9) TOPIQ [1], and four generalized FR-IQA models for enhanced images: 10) VIF [29], 11) PCQI [37], 12) SFSN [63] and 13) CKDN [62]. For a more fair comparison, we present the performance of the original and, when applicable, the fine-tuned versions (indicated by the suffix "-FT") of these models on the same combined dataset used to train A-FINE.

5.2. SRIQA-Bench

To further verify the effectiveness of A-FINE, we constructed an SR-based IQA benchmark, named **SRIQA-Bench**. We first compiled 100 original images covering a wide range of natural scenes and subjected them to common degradations [40, 54] to generate input low-resolution images. We then adopted two regression-based SR methods: 1) SwinIR [18] and 2) RRDB [39], and eight generation-based SR methods: 3) Real-ESRGAN [40], 4) BSRGAN [54], 5) HGGT [2], 6) SUPIR [52], 7) SeeSR [47], 8) StableSR [36], 9) SinSR [41] and 10) OSEDiff [46] to produce ten SR images for each input. Generally speaking, diffusion-based SR methods outperform GAN-based methods with more plausible textures, which in turn are more effective than regression-based SR methods with more realistic and sharper structures.

The subjective testing protocol is identical to the one described in Sec. 3, except that we performed a complete paired comparison experiment involving $\binom{11}{2} = 55$ pairs per input. To ensure rating reliability, each pair was assessed by at least ten subjects. A total of 40 subjects—comprising 25 males and 15 females aged between 21 and 39—took part in this study. More details about dataset construction can be found in the Appendix.

5.3. Main Results

Within-Dataset Results. Table 2 shows the accuracy results on the test sets of TID2013 [26], KADID-10K [20], PIPAL [10], and the proposed DiffIQA, from which we have several key observations. First, existing FR-IQA models, standard or generalized, exhibit noticeable performance declines on DiffIQA when the assumption of perfect reference quality is not met. Second, models based on *surjective* feature transformations (*e.g.*, AHIQ [14], TOPIQ [1], and CKDN [62]) demonstrate more pronounced improvements after fine-tuning on the combined dataset compared to models based on *injective* transformations (*e.g.*, LPIPS [58] and DISTS [5]). This suggests that DiffIQA is beneficial for enhancing generalized FR-IQA. Nonetheless, fine-tuned models typically show a performance drop on standard IQA datasets. Last, A-FINE achieves the highest average results, attributed to its adaptive weighting of image fidelity and naturalness terms.

Results on SRIQA-Bench. Table 3 presents the accuracy results on SRIQA-Bench. Most models perform adequately for regression-based SR methods. This is mainly because they have limited capabilities in creating plausible structures and textures. Consequently, the resulting SR images exhibit clearly inferior quality compared to their references.

Table 3. Accuracy (%) results on SRIQA-Bench. Pairs are formed within each subcategory and across the entire dataset.

Method	Regression-based	Generation-based	All
PSNR	80.7	41.7	34.7
SSIM [45]	83.0	45.3	37.4
MS-SSIM [44]	83.0	45.6	37.6
FSIM [55]	85.3	49.5	41.0
VSI [56]	81.3	50.1	41.2
LPIPS [58]	82.0	63.9	65.8
LPIPS-FT	84.7	63.8	72.2
DISTS [5]	83.3	66.6	72.4
DISTS-FT	**86.0**	63.9	71.4
AHIQ [14]	83.7	70.0	68.4
AHIQ-FT	71.0	71.5	69.6
TOPIQ [1]	83.7	63.9	67.0
TOPIQ-FT	78.3	73.0	77.7
VIF [29]	85.3	47.1	39.0
PCQI [37]	79.0	39.8	32.2
SFSN [63]	80.3	48.4	39.9
CKDN [62]	45.0	60.1	47.4
CKDN-FT	76.7	64.3	59.1
A-FINE (Ours)	83.3	**78.9**	**82.4**

In contrast, generation-based SR methods can produce output images of much higher quality, challenging the perfect reference quality assumption. Fine-tuned models on DiffIQA show clearly improved performance, and the proposed A-FINE demonstrates the strongest generalization to SRIQA-Bench, confirming its effectiveness in evaluating enhanced image quality.

5.4. Ablation Studies

Backbone. We try different backbone networks to implement A-FINE, including VGG16 [31], ImageNet-trained ResNet50 [12], CLIP-trained ResNet50 [27], ImageNet-trained ViT-B/32 [8], and CLIP-trained ViT-B/32 [27]. From the results in Table 4, it is evident that A-FINE benefits from more sophisticated computation (global attention over local convolution) and stronger backbone pretrained on more data (0.4B image-text pairs over 1M images).

Training Dataset. We train A-FINE only on DiffIQA or only on the combined standard IQA datasets TID2013 [26], PIPAL [10] and KADID-10k [20]. Table 5 lists the results, from which we find that training on the combined DiffIQA and standard IQA datasets using the pairwise learning-to-rank approach yields the best average performance with the strongest cross-dataset generalization.

Training Strategy. We last compare our three-phase training strategy with a baseline that trains all parameters of A-FINE simultaneously in a single phase. As shown in Table 6, single-phase training proves less effective in optimizing individual image fidelity and naturalness terms and adaptively balancing these two. In contrast, our proposed three-phase training strategy stabilizes the training dynamics, resulting in improved generalization.

Table 4. Ablation study on backbone networks. The accuracy values in the "Standard" column are averaged across the test sets of TID2013 [26], KADID-10K [20], and PIPAL [10]. The results of TOPIQ-FT are included for reference.

Backbone	Standard	DiffIQA	SRIQA-Bench		
			Reg.	Gen.	All
TOPIQ-FT	80.6	77.0	78.3	73.0	77.7
VGG16	77.6	77.0	79.0	75.0	79.8
ResNet50 (ImageNet)	74.8	69.6	84.7	70.7	77.2
ResNet50 (CLIP)	76.1	71.1	**85.2**	70.3	75.6
ViT-B/32 (ImageNet)	81.0	77.7	81.3	75.5	80.4
ViT-B/32 (CLIP)	**83.6**	**79.9**	83.3	**78.9**	**82.4**

Table 5. Ablation study on training datasets.

Training Dataset	Standard	DiffIQA	SRIQA-Bench		
			Reg.	Gen.	All
Standard	**84.1**	65.6	**86.7**	71.8	78.7
DiffIQA	70.6	79.6	78.3	72.9	76.0
Combined	83.6	**79.9**	83.3	**78.9**	**82.4**

Table 6. Ablation study on training strategies.

Training Strategy	Standard	DiffIQA	SRIQA-Bench		
			Reg.	Gen.	All
Single-Phase	79.7	79.6	79.3	75.1	77.9
Three-Phase	**83.6**	**79.9**	**83.3**	**78.9**	**82.4**

6. Conclusion and Limitations

We explored the problem of generalized FR-IQA, which relaxes the assumption of perfect reference image quality. From a data perspective, we introduced DiffIQA and SRIQA-bench to train and test generalized FR-IQA models, respectively. From a model perspective, we developed A-FINE, which adaptively combines an image fidelity term and an image naturalness term. We hope our data and model will inspire researchers in related fields to engage with the important research topic of generalized FR-IQA in the era of deep generative models.

Limitations. A-FINE represents one of the early efforts in generalized FR-IQA, paving the way for several promising research avenues. First, it can perform effectively when the perceptual quality of the reference image is reasonably high, even if not perfect. However, if the reference image quality is poor, the image fidelity term in A-FINE could introduce significant bias in quality prediction, which is worth deep investigation. Second, A-FINE can be viewed as a specific instance within the broader family of *asymmetric* distance measures. Exploring the optimal functional form for such measures in the context of generalized FR-IQA remains an open question. Third, we empirically observed a performance trade-off when evaluating the quality of distorted versus enhanced images. Identifying the best trade-off from both data and model perspectives presents an intriguing direction for future exploration.

Acknowledgments

This work was partly supported by the Hong Kong ITC Innovation and Technology Fund (9440379 and 9440390), and fully supported by the PolyU-OPPO Joint Innovative Research Center. We sincerely thank the volunteers who participated in our subjective study. A Human Subjects Ethics Committee approved the study and all participants signed consent forms beforehand.

References

[1] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. 7, 8

[2] Du Chen, Jie Liang, Xindong Zhang, Ming Liu, Hui Zeng, and Lei Zhang. Human guided ground-truth generation for realistic image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14082–14091, 2023. 7

[3] Du Chen, Zhengqiang Zhang, Jie Liang, and Lei Zhang. SSL: A self-similarity loss for improving generative image super-resolution. In *ACM International Conference on Multimedia*, pages 3189–3198, 2024. 2

[4] Scott J Daly. Visible differences predictor: An algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, pages 2–15, 1992. 2, 3

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020. 1, 2, 3, 5, 6, 7, 8

[6] Keyan Ding, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma. Locally adaptive structure and texture similarity for image quality assessment. In *ACM International Conference on Multimedia*, pages 2483–2491, 2021. 1, 3, 5

[7] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4):1258–1281, 2021. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[10] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S Ren, and Chao Dong. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651, 2020. 2, 3, 4, 6, 7, 8

[11] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. DIV8K: Diverse 8K resolution image dataset. In *IEEE/CVF International Conference on Computer Vision Workshop*, pages 3512–3516, 2019. 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 4

[14] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. Attentions help CNNs see better: Attention-based hybrid image quality assessment network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1140–1149, 2022. 7, 8

[15] Eric C Larson and Damon M Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):1–21, 2010. 2, 3, 4

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 2

[17] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc. Van Gool. LSDIR: A large scale dataset for image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023. 4

[18] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using Swin Transformer. In *IEEE/CVF International Conference on Computer Vision Workshop*, pages 1833–1844, 2021. 7

[19] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 2

[20] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10K: A large-scale artificially distorted IQA database. In *IEEE International Conference on Quality of Multimedia Experience*, pages 1–3, 2019. 2, 3, 4, 6, 7, 8

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 6

[22] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26 (2):1004–1016, 2016. 3

[23] Rafal K Mantiuk, Scott J Daly, Karol Myszkowski, and Hans-Peter Seidel. Predicting visible differences in high dynamic range images: Model and its calibration. In *Human Vision and Electronic Imaging X*, pages 204–214, 2005. 3

[24] Rafal K Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics*, 30(4):1–14, 2011. 3

[25] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 1

[26] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and Kuo J. C.-C. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30: 57–77, 2015. 2, 3, 4, 6, 7, 8

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 6, 8

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[29] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15 (2):430–444, 2006. 2, 3, 7, 8

[30] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. 3, 4, 6

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 6, 8

[32] Lingchen Sun, Rongyuan Wu, Jie Liang, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability and efficiency of diffusion models for content consistent super-resolution. *arXiv preprint arXiv:2401.00877*, 2023. 2

[33] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-LoRA approach. *arXiv preprint arXiv:2412.03017*, 2024. 2

[34] Louis L Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, 1927. 6

[35] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: A ranking method with fidelity loss. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 383–390, 2007. 6

[36] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 2, 7

[37] Shiqi Wang, Kede Ma, Hojatollah Yeganeh, Zhou Wang, and Weisi Lin. A patch-structure representation method for quality assessment of contrast changed images. *IEEE Signal Processing Letters*, 22(12):2387–2390, 2015. 3, 7, 8

[38] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 4

[39] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision Workshops*, pages 1–16, 2018. 7

[40] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *IEEE/CVF International Conference on Computer Vision Workshops*, pages 1905–1914, 2021. 4, 7

[41] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. SinSR: Diffusion-based image super-resolution in a single step. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024. 7

[42] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. 1, 2

[43] Zhou Wang and Alan C Bovik. Reduced-and no-reference image quality assessment. *IEEE Signal Processing Magazine*, 28(6):29–40, 2011. 1

[44] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems and Computers*, pages 1398–1402, 2003. 1, 3, 7, 8

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1, 2, 3, 5, 7, 8

[46] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024. 2, 7

[47] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 25456–25467, 2024. 2, 7

[48] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1191–1200, 2022. 1

[49] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image

super-resolution and personalized stylization. In *European Conference on Computer Vision*, pages 74–91, 2024. 2, 3, 4

[50] Hojatollah Yeganeh, Mohammad Rostami, and Zhou Wang. Objective quality assessment of interpolated natural images. *IEEE Transactions on Image Processing*, 24(11):4651–4663, 2015. 2, 3

[51] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. *arXiv preprint arXiv:2501.11561*, 2025. 1

[52] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024. 2, 7

[53] Chen Zhang, Wu Cheng, and Keigo Hirakawa. Corrupted reference image quality assessment of denoised images. *IEEE Transactions on Image Processing*, 28(4):1732–1747, 2018. 2, 3

[54] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 7

[55] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 1, 2, 3, 7, 8

[56] Lin Zhang, Ying Shen, and Hongyu Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10): 4270–4281, 2014. 7, 8

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 1, 2, 3, 4, 7, 8

[59] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. RankSRGAN: Generative adversarial networks with ranker for image super-resolution. In *IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 2

[60] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 1, 6

[61] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 1

[62] Heliang Zheng, Huan Yang, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning conditional knowledge distillation for degraded-reference image quality assessment. In *IEEE/CVF International Conference on Computer Vision*, pages 10242–10251, 2021. 2, 3, 7, 8

[63] Wei Zhou and Zhou Wang. Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity. In *ACM International Conference on Multimedia*, pages 934–942, 2022. 7, 8