#### **Rethinking Image Evaluation in Super-Resolution**

Shaolin Su<sup>1</sup> Josep M. Rocafort<sup>1,2</sup> Danna Xue<sup>1,2</sup>, David Serrano-Lozano<sup>1,2</sup> Lei Sun<sup>3</sup> Javier Vazquez-Corral<sup>1,2</sup> <sup>1</sup>Computer Vision Center <sup>2</sup>Universitat Autonoma de Barcelona <sup>3</sup>INSAIT, Sofia University St. Kliment Ohridski



Figure 1. We show that even Ground Truth (GT) images in existing SR datasets [6, 33] can show relatively poor quality. As a result, image metrics tend to favor outputs that more resemble the reference GTs (middle), even when they are perceptually poorer (left side), leading to contradictory evaluations with human preferences (right side). Please zoom in for better comparisons.

#### Abstract

While recent advancing image super-resolution (SR) techniques are continually improving the perceptual quality of their outputs, they can usually fail in quantitative evaluations. This inconsistency leads to a growing distrust in existing image metrics for SR evaluations. Though image evaluation depends on both the metric and the reference ground truth (GT), researchers typically do not inspect the role of GTs, as they are generally accepted as 'perfect' references. However, due to the data being collected in the early years and the ignorance of controlling other types of distortions, we point out that GTs in existing SR datasets can exhibit relatively poor quality, which leads to biased evaluations. Following this observation, in this paper, we are interested in the following questions: Are GT images in existing SR datasets 100% trustworthy for model evaluations? How does GT quality affect this evaluation? And how to make fair evaluations if there exist imperfect GTs? To answer these questions, this paper presents two main contributions. First, by systematically analyzing seven state-of-the-art SR models across three real-world SR datasets, we show that SR performances can be consistently affected across models by low-quality GTs, and models can perform quite differently when GT quality is controlled. Second, we propose a novel perceptual quality metric, Relative Quality Index (RQI), that measures the relative

quality discrepancy of image pairs, thus issuing the biased evaluations caused by unreliable GTs. Our proposed model achieves significantly better consistency with human opinions. We expect our work to provide insights for the SR community on how future datasets, models, and metrics should be developed.

#### 1. Motivation

Image super-resolution aims at reconstructing a highresolution (HR) image from a low-resolution (LR) observation. The advance of recent image processing techniques [19, 20, 31, 34, 40, 47] has enabled realistic SR models to not only super-resolve images but also restore other types of degradations in the real world, such as noise, blurriness, and compression [31, 40]. However, recent studies have still found an inconsistency between human perceptual and model quantitative evaluations, i.e. models that achieve better visual quality can easily fail under existing evaluation metrics such as PSNR, SSIM, and LPIPS [16, 30, 34]. As a result, a growing distrust in existing image metrics emerges, and researchers have to conduct self-organized user studies [30, 34] or adopt varying image metrics [35, 37] to demonstrate their performances. In this paper, aside from 'blaming' the metrics, we aim to investigate and rethink the issue. so that a fair understanding of the SR evaluations can be reached and further justifications can be explored.

Given an output image  $I_{HR}$  from a SR model, the evaluation of  $I_{HR}$  can be generally measured by:

$$Q = D(I_{HR}, I_{GT}), \tag{1}$$

where  $I_{GT}$  is the GT image served as reference, and  $D(\cdot)$  can be any similarity metric such as PSNR and LPIPS. Thus, the evaluation results depend both on the choice of D and the quality of  $I_{GT}$ . However, while a variety of image metrics are studied, few have investigated the roles of GTs during evaluation. Thus, it remains to ask: are GTs really 100% trustworthy? Or are image metrics, not the only one responsible for the growing inconsistency between perceptual and quantitative evaluations?

In Figure 1, we show GT images (center) from two SR datasets RealSR [6] and DRealSR [33]. These GTs show relatively poor quality, due to the device limitation —images were captured several years ago— or the lack of careful control over other types of distortions. Meanwhile, advancing state-of-the-art (SOTA) SR methods (*e.g.* diffusion-based models [34, 37]) can remove multiple degradations and produce finer details, often producing outputs (right) that surpass the GT images in perceptual quality. As a result, when applying an existing similarity metric  $D(\cdot)$ , the lower-quality outputs from models (left) will instead show higher similarity to the GT, leading to contradictory evaluations that images presenting superior quality are performing worse.

This observation motivates us to rethink the current de facto evaluations in SR. Thus, we investigate how the existence of unreliable GTs affects the evaluations of current SR models and how to make fair evaluations with imperfect GTs. To do so, we first test seven representative SOTA SR models on three datasets and evaluate how their performances are affected by the quality of GTs. We observe an inherent connection between model performances and GT quality —despite concrete models or evaluation metrics, which sheds light on how future models and datasets should be developed. Second, we propose a simple yet effective solution to correct the unfairness caused by imperfect GTs. Specifically, we propose a relative IQA scheme, Relative Quality Index (RQI), to measure two arbitrary images that may have varying quality, instead of treating GT as the perfect reference. Our newly defined metric outperforms existing IQA metrics on both user-based opinions and public benchmarks.

In summary, the main contributions of this paper include:

• We point out the existence of imperfect GTs in current widely used SR datasets and then systematically analyze how GT quality affects model evaluations, demonstrating the inherent connection between model performance and GT quality, and providing valuable insights into the future development of models and datasets. We propose a novel perceptual quality metric, RQI, to improve SR evaluations with unreliable GTs. RQI assesses relative quality discrepancies between image pairs, regardless of their quality levels, addressing the limitations of poor GTs. Experimental results on both user opinions and public benchmarks demonstrate its superiority.

#### 2. Related Work

#### 2.1. Image super-resolution

Image super-resolution reconstructs HR images from LR inputs, serving as a cornerstone for medical imaging, satellite analysis, video enhancement, etc. [8-11, 13, 14, 19, 20, 30, 31, 34, 37, 40, 45, 46, 48]. Early approaches [8, 11, 13, 14, 20, 47] operate under the assumption of a predefined degradation process, such as bicubic downsampling or blurring with a known prior. However, their effectiveness is often constrained in real-world scenarios where noise or compression artifacts may occur. To handle practical SR, BSR-GAN [40] and RealESRGAN [31] assume complex degradation in LR images and adversarially train models to remove multiple degradations. Transformer-based methods further improve HR quality by capturing long dependencies between pixels: SwinIR [19] utilizes window attention in Swin Transformer [22] for global dependencies, while HAT [9] optimizes hierarchical feature integration. With the emergence of diffusion techniques [15], diffusion-based SR models [30, 34, 37] are able to learn an even more powerful image representation and effectively handle various distortions in LR images.

However, while the perceptual quality of SR outputs keeps improving, recent research finds a lingering inconsistency between perceptual quality and quantitative evaluations of SR models [30, 34, 37]. As a result, either cumbersome user studies or varying image metrics are sought to justify SR model performances. In this paper, we investigate the issue and propose a solution that better correlates with human perception.

#### 2.2. Image quality metrics

Image quality metrics are widely used as evaluation criteria for image processing systems. In most cases, the evaluations assume GTs are accessible for varying tasks, and thus, full-reference IQA (FR-IQA) approaches are applied. These approaches measure the similarity between the target and GT image to evaluate the performance. Common metrics include distortion-based metrics such as PSNR, SSIM [32] and FSIM [41], and recent perceptual-based metrics such as LPIPS [44] and DISTS [12]. However, since FR-IQA assumes references are perfect, only absolute similarity/dissimilarity is measured, thus failing in handling cases where reference images are perceptually poor.

Noticing the inconsistency between human perception

and some FR-IQA metrics, recent SR studies also adopt noreference IQA (NR-IQA) metrics for evaluations. Widely applied metrics including Natural Scene Statistics (NSS) based metrics such as NIQE [24], IL-NIQE [42] and PI [5], and deep-learning based metrics such as MUSIQ [17], MANIQA [36] and Clip-IQA [29]. Since these metrics evaluate image quality without any reference, it is still doubtful if they make reliable predictions. As a result, current SR studies [30, 34, 37] widely conduct user studies to obtain convincing evaluations.

Note that some IQA datasets (*e.g.* BAPPS [43] and PieAPP [26]) employ 2AFC (Two Alternative Forced Choice) strategy to collect user preferences over a pair of images, which resembles the proposed relative quality scheme. However, the logic is still different. In 2AFC, given a pair of images, users are asked to compare which one is closer to the reference, in contrast, in RQI, we only compare which image is relatively better than the other one without any reference, since we assume references can also show poor quality. Based on this assumption, we develop a different training strategy from traditional FR- or NR- metrics, showing its superiority over current schemes.

#### 3. How GTs Affect SR Model Evaluations

We focus on the classic  $\times 4$  SR task and evaluate seven representative SOTA SR models, including two GANbased models RealESRGAN [31], BSRGAN [40], two transformer-based models SwinIR [19], HAT [9], and three difussion-based models StableSR [30], SeeSR [34] and PASD [37]. We evaluate all the models according to their official implementations on test sets from 3 real-world SR datasets, i.e. DIV2K-wild [2], DRealSR [33] and RealSR [6]. We first employ the NR-IQA metric KonIQ++ [27] to assess the quality and degradation of GTs from three datasets, and show their distributions in Figure 2. As can be seen from Figure 2. a), even for GT images, their quality can be limited as the highest quality score can reach 100. The issue is more apparent for DRealSR and RealSR, where most quality scores fall below 60. For degradation (Figure 2. b)-d)), while the contrast problem is small, blur and noise can exist in some images, potentially bringing biases for evaluations. In Figure 3, we intuitively show GT samples from three datasets, where they suffer from different degradation including blur, noise and vague details.

After assessing the quality of GTs in three datasets, we investigate how their quality affects the evaluations of SR models. In Figure 4, we gradually discard images with the lowest GT quality according to the KonIQ++ scores from all three datasets and evaluate the model performance on the remaining images. We discard from 0% to 80% images in total and show average PSNR, SSIM, LPIPS, and DISTS scores with the amount of discarded images. From Figure 4, we make several observations and discussions:



(c) Noise level distribution (d) Contrast problem distribution Figure 2. GT Quality and degradation distributions in three SR datasets. All scores range from 0 to 100, (a): a higher quality score indicates better quality, (b)-(d): higher degradation scores indicate larger distortions.



Figure 3. GT samples from RealSR [6], DRealSR [33] and DIV2K [2] dataset. The images suffer from blur, vague details, and noise problems respectively. Please zoom in for better view.

1. A challenging image will always be challenging. We observe that in all the models, similar performance fluctuations occur when the same image is discarded, indicating that challenging images are always challenging regardless of SR models. This observation also provides insight into how to improve future SR models by troubleshooting those images that are challenging across models.

2. High quality GTs are consistently challenging for SR models. As low-quality GTs are discarded from evaluations, we observe a consistent performance drop for all the models on all the metrics (decrease in PSNR, SSIM, and increase in LPIPS and DISTS). We also show in Supplementary Material that this phenomenon is not incidental, *i.e.*, the consistent drop would not happen if we randomly discard images. This observation indicates the evaluation of SR models can be inherently affected by the quality of GTs, where a high-quality GT can usually lead to lower performances. We attribute this to the loss of details and vague structures in low-quality GTs (see Figure 3), making it easy to achieve high quantitative evaluations on both distortion and perception metrics. The phenomenon further indicates



Figure 4. We show how evaluations of 7 SR models change when low-quality GT are gradually discarded from the testing datasets.

the limitations of current SR models in generating finer details and producing very high-quality outputs.

**3.** Different evaluation results can be reached when GT quality is controlled. We can observe this change by perceptual metrics LPIPS and especially DISTS, where model rankings can change dramatically according to the GT quality. A clear example is SeeSR [34], which moves from Rank #6 to Rank #2 by LPIPS and from Rank #6 to Rank #1 by DISTS when looking only at images with higher GT quality. This trend agrees with our assumption that model outputs can be better than GTs, resulting in those methods obtaining lower rankings when poorer GTs are included. This also indicates the existence of biased evaluations with unreliable GTs, as a solution, in Section 4, we propose RQI to relieve the impact of low-quality GTs.

4. The perception-distortion tradeoff [4] also exists. We can see that HAT [9] and PASD [37], the two topperforming models by distortion metrics PSNR and SSIM, perform relatively low on perception metrics LPIPS and DISTS. In contrast, well-performing models by perceptual metrics (SwinIR [19] and SeeSR [34]) achieve low performances on PSNR and SSIM. However, it is worth thinking, in cases where GT quality is low, is it still convincing to compute pixel-wise distortion metrics as fidelity evaluations? For GTs with distortions, the fidelity metrics will reward methods that fail to remove distortions. We believe controlling high-quality GT in future SR datasets and developing adaptive fidelity metrics can relieve the issue.

The analysis above demonstrates that low-quality GTs can introduce biased SR model evaluations. Therefore, with the current evaluation paradigm, it is crucial to guarantee GT quality during evaluations to ensure fairness. However, even if it is possible to evaluate models only on high-quality GTs in existing datasets by filtering out the 'unqualified' ones, the test samples will be reduced, thus affecting the reliability of evaluations. To alleviate these problems, in the following section, we further investigate how to make fair evaluations with 'imperfect' GTs.

#### 4. How to Fairly Evaluate with Imperfect GTs

Our solution is straightforward and simple: Since we do not recognize GTs as perfect references, we allow cases in which model outputs (*i.e.* target images) can achieve better quality than the GT. Therefore, instead of measuring absolute differences between target and reference (GT) images as adopted in previous FR metrics [12, 32, 44], we serve GTs as anchors and measure the relative quality from target images to GTs (*i.e.* either better or worse). We name the proposed scheme RQI (Relative Quality Index), and define it with the following property:

$$RQI(I_{HR}, I_{GT}) \begin{cases} > 0, \ I_{HR} \text{ has better quality.} \\ < 0, \ I_{GT} \text{ has better quality.} \end{cases}$$
(2)

This design is different from the current FR-IQA scheme in three aspects, shown in Figure 5:

- 1. RQI is an order-sensitive metric. Thus,  $RQI(I_A, I_B)$ and  $RQI(I_B, I_A)$  can lead to opposite results. In contrast, in existing FR metrics [12, 32, 44],  $D(I_A, I_B)$  and  $D(I_B, I_A)$  have the same results, as they suppose that one of the inputs always shows the best quality.
- 2. Any two images are considered as pairs during training. For current FR-IQA scheme, given one reference image  $I_0$  and a sequence of its distorted images  $\{I_1, I_2, ..., I_n\}$ , only image pairs  $\{I_0, I_i\}, i \in [1, n]$  are constructed to train the model. In RQI, as we consider that GTs also contain distortions, we select arbitrary two images from the sequence  $\{I_0, I_1, ..., I_n\}$  to construct training pairs  $\{I_i, I_j\}, i, j \in [0, n], i \neq j$ . This not only yields to a larger amount of training samples  $(n^2 \text{ vs. } n)$ , but also covers more complex cases in which both GTs and target images may contain varying degradations.
- 3. *RQI calculates discrepancies*. For training labels, we calculate quality discrepancy between  $I_i$  and  $I_j$ , *i.e.*  $q_i q_j$ , whereas FR-IQA metrics adopt  $q_i$  as quality label for image pair  $\{I_0, I_i\}$ .

Training an IQA model under the RQI scheme is simple: we select random image pairs (containing the same content) from any existing IQA dataset and calculate their quality differences to use them as discrepancy labels (either positive or negative). Then, the ordered image pairs are fed to the model to predict the discrepancy value. Thus, we can train an arbitrary FR-IQA model on any IQA dataset under the proposed scheme. In practice, we train three different IQA models AHIQ [18], MANIQA [36], and TOPIQ [7], on



(a) Traditional FR-IQA scheme (b) The proposed RQI scheme

Figure 5. The proposed RQI scheme differs from traditional FR-IQA scheme in three aspects: 1. RQI is order-sensitive. 2. We substitute reference image  $I_0$  to any image  $I_i$  in the distorted image sequence. 3. Relative quality discrepancy is used as label.

three different datasets Kadid-10K [21], PieAPP [26] and PIPAL [16]. Since MANIQA [36] is an NR-IQA model, we modify it to extract features both from the target image and the reference image, and then concatenate the features before the transposed attention block to fuse them.

For all datasets, we construct labels for arbitrary image pairs and normalize them to [-1, 1]. We also remove the activation functions in the last regression layer to ensure all of the models produce negative values. We didn't train the models on the perceptual IQA dataset BAPPS [43] since the  $64 \times 64$  image patch resolution is too small for the selected IQA models. We train all models following their official configurations, and we select the best-performing models by looking at their performances on the validation split. During testing, we fed the target image and GT to the model in order. We also downscale the input 2 times and select random crops from different scales to ensure multiscale features are captured from high-resolution inputs. The mean score of the crops is reported as the RQI value.

#### **5.** Evaluations

In this section, we conduct two main types of evaluations: a subjective user study and quantitative comparisons regarding the consistency of image metrics.

#### 5.1. User study and discussion

To analyze how existing image metrics correlate with human opinions, we first conduct a comprehensive user study to collect subjective scores on different SR models and their corresponding GTs. Specifically, we collect user opinions from seven selected SR models, tested on five SR test datasets DIV2K-wild [2], DRealSR [33], RealSR [6], Set5 [3] and Set14 [39].

Our experiment consists of a two-alternative forced choice (2AFC) paradigm. The observers were asked to select the HR image that showed better perceptual quality from a pair of images containing the same content. For each source scene, the users compared eight images (7 model outputs + GT). For each comparison, we collect opinions



Figure 6. User statistics of the best quality HR image in four SR datasets.

from at least 15 users, all of whom have passed the Ishihara test to avoid color blindness. Then, we use Thurstone's model [28] to reconstruct a ranking from the user's opinion.

We carefully control the experiment in a matte dark room where the monitor is the only light source. Images are shown in a 3K resolution monitor set to sRGB. We randomize the order and placement (left or right) of image pairs to reduce potential biases. For the DRealSR dataset, since the HR images are more than 4K resolutions, we only compare the center-cropped images to avoid scaling effects. In total, we collect user scores on  $8 \times 312 = 2,496$  images, far exceeding image opinions collected in previous SR studies [30, 34, 37].

In Figure 6, we show statistics of the best-ranked images by the users from all test sets. Since Set4 and Set15 are similar and only contain a few images, we combine these two datasets for evaluation in the rest of the paper. From the statistics, we can extract several findings.

**1.** There exist model outputs better than GTs. For all the datasets, there exist model outputs that are perceptually better than GTs, and the percentage increases when the dataset contains poorer GTs (a small percentage from DIV2K and more than a half from Set5&Set14).

**2. Diffusion-based models are preferred**. Most of the preferred model outputs are from diffusion-based models, indicating that the diffusion-based scheme has a better capability to recover distortions and reconstruct finer details.

**3.** SeeSR [34] is the preferred method by users. Among all the competing models SeeSR produces most of the preferred images, which is also consistent with our observation in Figure 3, where perceptual metrics tend to favor SeeSR when low-quality GTs are discarded for evaluations.

In Figure 7, we show the Thurstone scores for all the datasets. We can extract similar observations. More discussions about SR models' behavior can be found in the supplementary material.

#### 5.2. The effectiveness of the RQI scheme

We calculate the Spearman Rank order Correlation Coefficient (SRCC) between IQA model predictions and user opinions to show how model evaluations correlate with human perception. Specifically, for each source image, we obtain model predictions and Thurstone scores on the 7 SR



Figure 7. Average user scores on different SR models (including GT) in four SR testing datasets.

Table 1. The effectiveness of the proposed RQI scheme. We train varying IQA models on different datasets using traditional FR-IQA and the RQI scheme (with subscript 'R'). SRCC consistency with user opinions are reported.

Train Set	Kadid	Kadid <sub>R</sub>	PieAPP	PieAPP <sub>R</sub>	PIPAL	PIPALR				
Model	AHIQ									
DIV2K	0.365	0.506	0.459	0.515	0.431	0.626				
RealSR	0.196	0.181	0.095	0.284	0.452	0.474				
DRealSR	0.267	0.350	0.244	0.378	0.413	0.467				
Set5&14	0.292	0.426	0.203	0.378	0.280	0.472				
Model	MANIQA									
DIV2K	0.502	0.55	0.573	0.570	0.624	0.744				
RealSR	0.238	0.258	0.098	0.208	0.470	0.504				
DRealSR	0.343	0.343	0.285	0.417	0.372	0.529				
Set5&14	0.472	0.504	0.386	0.483	0.544	0.588				
Model	TOPIQ									
DIV2K	0.462	0.490	0.374	0.414	0.490	0.561				
RealSR	0.101	0.233	0.107	0.133	0.277	0.328				
DRealSR	0.164	0.282	0.003	0.322	0.042	0.357				
Set5&14	0.060	0.334	0.035	0.010	0.025	0.271				

model outputs and calculate their SRCC. For all images in a dataset, the averaged SRCC value is reported.

As stated in Section 4, we train three IQA models (AHIQ [18], MANIQA [36] and TOPIQ [7]) on three IQA datasets (Kadid-10K [21], PieAPP [26] and PIPAL [16]), both in traditional FR-IQA setting and in our proposed RQI setting (with subscript 'R'). In total 18 models are tested on user scores collected for four SR datasets.

Table 1 shows the results and we make several observations. First, we observe a consistent SRCC improvement on all four testing sets by training under the RQI scheme, regardless of which IQA model or training set is applied. This indicates the effectiveness of the scheme. Second, by calculating the average improvement for each testing dataset, the improvements on DIV2K, RealSR, DRealSR, and Set5&Set14 are 0.077, 0.085, 0.146, and 0.138, respectively. The improvements on DRealSR and Set5&Set14 are relatively larger, consistent with our observation in Figure 6, resulting that when more outputs are preferred than GTs, the RQI becomes more effective. Third, among three selected IQA models, we observe MANIQA [36] achieves better consistency, while among all training sets, models trained on PIPAL [16] perform better. Therefore, we select the best-performing model  $RQI_{MANIQA}$  trained on the PIPAL dataset for analysis in the rest of the paper.

#### 5.3. Analysis of existing metrics

We further analyze how current evaluation metrics correlate with human perception. The selected metrics include widely used distortion metrics SSIM [32] and PSNR, recent deep-learning based perception metrics DISTS [12] and LPIPS [44], traditional NR-IQA metrics NIQE [24] and PI [5], recent deep-learning based NR-IQA metrics Clip-IQA [29] and MANIQA [36], and the proposed RQI. We calculate, for each source image, SRCC and Pearson Linear Correlation Coefficient (PLCC) between metric evaluations and user opinions. We report the averaged values for each dataset. In addition, since in SR comparisons, the bestperforming model is more valued, we also compute the prediction accuracy for the best-performing model within each dataset (Winning Rate). The results are shown in Table 2. Based on this table, we can make a detailed analysis of current image evaluation metrics.

1. Traditional distortion image metrics do not correlate with human perception. SSIM [32] and PSNR, though surprising, are making opposite evaluations with human perception on all the datasets. As we collect user opinions purely on their perceptual impressions of the images, the results further demonstrate the perception-distortion trade-off [4]. However, we believe the inconsistency partly stems from the issue of unreliable GTs, therefore, we urge for high-quality GTs in future SR datasets.

2. Deep-learning-based perceptual metrics perform similarly. Both DISTS [12] and LPIPS [44] show some consistency on the DIV2K and Set5&Set14 datasets but fail on the RealSR and DRealSR datasets. This is probably due to the poor image quality in the RealSR dataset and the larger image resolution size in the DRealSR dataset.

**3. Traditional NR-IQA metrics do have good consistency.** Two traditional NR-IQA metrics, NIQE [24] and PI [5], show good consistency with human perception. This indicates the power of traditional NSS features for image quality measurement.

4. Deep-learning based NR-IQA also present some degree of consistency. The two deep-learning based NR-

Dataset	Criterion	SSIM	PSNR	DISTS	LPIPS	NIQE	PI	Clip-IQA	MANIQA	RQI
DIV2K [2]	SRCC	-0.348	-0.079	0.610	0.415	0.516	0.565	0.593	0.554	0.744
	PLCC	-0.360	-0.124	0.627	0.452	0.492	0.549	0.593	0.553	0.785
	Win Rate	0.05	0.19	0.44	0.41	0.44	0.57	0.50	0.47	0.65
RealSR [6]	SRCC	-0.220	-0.116	0.048	0.008	0.263	0.317	0.377	0.187	0.504
	PLCC	-0.289	-0.160	0.027	-0.031	0.282	0.325	0.437	0.212	0.484
	Win Rate	0.04	0.05	0.12	0.11	0.47	0.51	0.58	0.32	0.49
DRealSR [33]	SRCC	-0.354	-0.355	-0.102	-0.141	0.240	0.303	0.268	0.284	0.529
	PLCC	-0.409	-0.405	-0.129	-0.143	0.222	0.301	0.268	0.331	0.603
	Win Rate	0.02	0.01	0.10	0.04	0.44	0.46	0.38	0.35	0.53
Set5&Set14 [3, 39]	SRCC	-0.321	-0.204	0.403	0.282	0.578	0.466	0.642	0.437	0.664
	PLCC	-0.387	-0.239	0.414	0.293	0.527	0.506	0.683	0.443	0.673
	Win Rate	0.06	0.06	0.35	0.24	0.41	0.29	0.29	0.24	0.35

Table 2. Consistency evaluations of current metrics with human perception. SRCC, PLCC and winning rate are reported.

IQA models, Clip-IQA [29] and MANIQA [36], also achieve user consistency to some extend.

From 3. and 4. we should notice that all the four NR-IQA metrics (NIQE [24], PI [5], Clip-IQA [29] and MANIQA [36]) are out performing the FR-IQA metrics on RealSR, DRealSR and Set5&Set14 datasets. Since GT quality in these datasets can be poor, the result further demonstrates the inconsistency can be attributed to the poor quality of the reference GT images.

**5. RQI outperforms the other metrics.** RQI achieves the best consistency with human perception in most cases, and we attribute its superior performance to the training scheme that covers complicate cases where both target and reference image contain distortions. For more qualitative analysis of existing metrics, please refer to Section 5.5.

#### 5.4. Evaluation on other SR-IQA benchmarks

We further evaluate all the metrics on other two SR-IQA benchmarks BSD-SR [23], QADS [49], and one general IQA dataset Kadid-10K [21]. The BSD-SR dataset contains 30 source images and 1,620 HR images generated from 6 scales, the QADS dataset contains 20 source images and 980 HR images with interpolation factors 2, 4, and 8, and the Kadid-10K dataset contains 81 source images and 10,125 distorted images. Images in BSD-SR and QADS are produced by varying SR methods, while the Kadid-10K dataset contains images generated from different distortions. All datasets collect Mean Opinion Scores (MOS) from users as perception scores. We test the metrics on the whole data of three datasets and report SRCC and PLCC values. Both mean consistency over each source image and overall consistency are reported. Note that we do not evaluate DISTS [12] on the Kadid-10K dataset since the model is pre-trained on this same dataset. We show the results in Table 3 and make the following observations.

For the two SR-IQA datasets (BSD-SR and QADS), traditional metrics such as SSIM [32] achieve good consistency when evaluating outputs from the same source images. We attribute this to the simple SR models (interpo-

lation, dictionary-based, and early CNN models) collected in the two datasets. Therefore, it is relatively easier to distinguish image quality in early SR models. Yet their performances are limited in evaluating all images across contents. Second, the FR-IQA metrics generally perform better than NR-IQA metrics, which seems to be contradictory with the results in Table 2. However, HR images in these two datasets are produced by earlier SR models, mostly showing poorer quality than GTs. As GTs in these datasets are serving as 'better' references, FR-IQA metrics are able to make better predictions than metrics that do not consider reference images. Third, among all the metrics, RQI can perform stably across all datasets and criteria. Note that we train RQI only by relative quality discrepancy, but it still achieves good consistency across image contents, on different SR scales, and even on the general IQA task. All the results validate the generalized ability of the RQI metric.

#### 5.5. Qualitative analysis

Figure 8 shows how existing distortion FR-IQA metrics, perception FR-IQA metrics, and NR-IQA metrics can fail. Among them, the distortion FR-IQA metric SSIM [32] (the up-left case) tends to favor blur regions (RealESRGAN) more than textures (SeeSR), which also matches previous findings [44]. For NR-IQA metrics, though they perform well in evaluating overall image quality, they fail when image semantics change from GTs due to their lack of reference. As shown up-right, the multi-model based NR-IQA metric ClipIQA [29] fails to evaluate character consistency for BSRGAN and PASD, since it cannot refer to correct semantics. Finally, perceptual FR-IQA metrics can easily fail when GT quality is poor (two cases at the bottom). In these cases, we show that when GT contains either vague details or distortions, the two perceptual FR-IQA metrics LPIPS [44] and DISTS [12] favor blurred results (HAT) more than sharp ones (SeeSR) since they are perceptually closer to the reference GTs. In contrast, the proposed RQI is perceptually accurate, aware of semantic consistency, and avoids biased evaluations caused by unreliable GTs.

SSIM PSNR DISTS LPIPS NIOE PI Dataset Criterion Clip-IQA MANIQA ROI 0.949 0.945 0.947 0.901 0.875 0.793 SRCCmean 0.668 0.789 0.901 0.945 0.940 0.950 0.910 0.893 0.786 0.815 0.901 PLCCmean 0.664 BSD-SR [23] 0.703 0.759 SRCC<sub>all</sub> 0.617 0.438 0.827 0.624 0.639 0.849 0.842 **PLCC**<sub>all</sub> 0.625 0.454 0.708 0.760 0.840 0.828 0.611 0.643 0.868 0.704 SRCC<sub>mean</sub> 0.927 0.727 0.887 0.832 0.420 0.760 0.842 0.912 0.910 PLCCmean 0.923 0.562 0.869 0.823 0.398 0.704 0.685 0.826 QADS [49] SRCC<sub>all</sub> 0.552 0.193 0.703 0.619 0.394 0.708 0.489 0.759 0.828 **PLCC**<sub>all</sub> 0.547 0.213 0.706 0.618 0.327 0.651 0.488 0.733 0.822 SRCCmean 0.649 0.261 0.809 0.393 0.406 0.558 0.548 0.669 -**PLCC**<sub>mean</sub> 0.633 0.247 0.801 0.379 0.376 0.480 0.531 0.651 Kadid-10K [21] SRCC<sub>all</sub> 0.595 0.231 0.741 0.435 0.474 0.534 0.574 0.666 0.720 0.548 **PLCC**<sub>all</sub> 0.585 0.229 0.389 0.425 0.485 0.649 BSRGAN PASD GT RealESRGAN SeeSR GT

Table 3. Consistency evaluations of image metrics on three IQA benchmarks. BSD-SR [23] and QADS [49] are two SR-IQA datasets, and

Kadid-10K [21] is a general IQA dataset. The best and second best performances are in **bold** and underscore.



Figure 8. We show different cases where existing metrics fail. Up-left: failure case for the distortion FR-IQA metric SSIM [32]. Up-right: failure case for the NR-IQA metric ClipIQA [29]. Bottom: failure cases for perception FR-IQA metrics LPIPS [44] and DISTS [12]. As a comparison, RQI handles all the cases correctly. All scores are normalized to [0,1] for easier comparisons. Please zoom in for better view.

#### 6. Discussion

We point out the existence of poor-quality GTs in existing SR datasets and raise discussions about how future SR research can be developed. First, we urge careful consideration in building future SR datasets, where GT quality should be meticulously controlled. Second, we observe certain images can be consistently challenging for existing SR models, suggesting a further development of the models upon the challenging cases. Third, with imperfect GTs in current SR datasets, we propose RQI to fairly evaluate models, and we expect the same consideration can be taken when future metrics are developed. Last, it will also be interesting to explore if the same issue exists in other low-level image processing tasks such as deblurring [25], denoising [1], and

### 7. Conclusion

In this paper, we study the emerging inconsistency between perceptual and quantitative evaluations that bothers current SR models. We point out the poor quality of GTs is also responsible for this inconsistency. After demonstrating how GT quality affects SR model evaluations, we propose a simple yet effective scheme, RQI, to relieve the evaluation biases caused by imperfect GTs. We show that the proposed RQI can achieve better consistency with human perception data, both on a new user study and on existing benchmarks. We believe our research will shed light on both future SR research and evaluations.

deraining [38], *etc.* and we leave this as a future work.

#### Acknowledgment

This work was supported by Grant PID2021-128178OB-I00 funded by MCIN/AEI/10.13039/501100011033, ERDF "A way of making Europe", the Departament de Recerca i Universitats from Generalitat de Catalunya with ref. 2021SGR01499. Shaolin Su was supported by the HORI-ZON MSCA Postdoctoral Fellowships funded by the European Union (project number 101152858). Danna Xue was supported by the grant Càtedra ENIA UAB-Cruïlla (TSI-100929-2023-2) from the Ministry of Economic Affairs and Digital Transition of Spain. David Serrano-Lozano was supported by the FPI grant from Spanish Ministry of Science and Innovation (PRE2022-101525). Lei Sun was partially funded by the Ministry of Education and Science of Bulgaria's support for INSAIT as part of the Bulgarian National Roadmap for Research Infrastructure.

#### Appendix

## A. More analysis of how GT quality affects SR evaluations

In the main paper, we employed the KonIQ++ [27] model to assess GT quality and showed there exists an inherent connection between GT quality and model evaluations. In this part, we show that similar observation can also be made from other IQA model predictions, and that this connection is not occasional. We employ another NR-IQA model MANIQA [36] to assess GT quality, and discard images based on its prediction. The evaluation results are shown in Figure 9, where we can see a consistent drop in model performances. In Figure 10, we show that by randomly discarding images, this consistent performance drop will not occur. Figures are shown in unified scale.

#### **B.** Implementation details

The selected models are trained following their official implementations, *i.e.* for all the models, the learning rate is set to  $10^{-4}$  with weight decay  $10^{-5}$ . The batch size is set to 4 for AHIQ [18], and 8 for MANIQA [36] and TOPIQ [7]. AHIQ [18] and MANIQA [36] randomly crop image patches with size 224, while TOPIQ [7] randomly crop image patches with size 384. The crops are randomly flipped during training for augmentation. We split the training datasets with 20% images according to contents as the validation set, and select the best-performing models upon their performances on the validation set.

During testing, we randomly crop patches from the inputs. The patches are cropped from the same regions of the target image and GT. Since we crop patches from 3 different scales, in each scale, 20 patches are cropped and the final score is averaged across all patches. The downscaling and cropping will only operate on images that are larger than the required input size of the models (*i.e.* 224 for AHIQ [18] and MANIQA [36], and 384 for TOPIQ [7]).

#### C. More discussions about SR models

In this section, we discuss the behaviors of SR models (mainly diffusion-based models) in more detail. We show qualitative examples in which model outputs are perceptually better than GTs, and then briefly analyze the three diffusion-based models StableSR [30], SeeSR [34] and PASD [37].

#### C.1. More examples of perceptually poorer GTs and better model outputs

In Figures 11 - 14, we show cases in which model outputs are better than GTs from datasets DIV2K [2], RealSR [6], DRealSR [33], Set5 [3] and Set14 [39] respectively. As shown, GTs across datasets exhibit noise, blur, compression artifacts, and vague details. For DIV2K [2], the overall GT quality is better, thus less suffers from the problems. However, for Set5&Set14, where the images were collected in early ages, current models can easily produce better outputs than GT quality. This further indicates that, with the advancing of methods, it is also necessary to update corresponding datasets and evaluation metrics to match the renewing technology.

# C.2. More discussions about diffusion-based SR models and how they affect human perception

Though generally, diffusion-based models can produce perceptually better images, from user preferences, we also find that the three evaluated diffusion-based models can perform differently from each other. In this part, we briefly discuss how the models perform and how the results affect the perceptual evaluations.

- StableSR [30] leans to increase image contrast by sharpening the edges. In cases where edges are thin, the overall image contrast is increased, leading to better perceptual quality. However, when images contain strong edges, StableSR tends to over-process the edges which brings artifacts, leading to poorer perceptual evaluations (see Figure 15).
- SeeSR [34] produces perceptually better images in most cases, due to its strong generation power. However, we find it sometimes changes image semantics, affecting user choices during experiments (see Figure 16).
- PASD [37] is also capable of recovering high-quality details, however, the model leans to produce out-of-focus effects (see Figure 17). During experiments, we find this effect can happen on the wrong subjects in the image, therefore leading to poorer perceptual quality.

Note that even with different behaviors of the diffusion-



Figure 9. Model evaluation results when discard low quality GTs according to MANIQA [36] predictions. Similar observations can also be made to those of the main paper.



Figure 10. Model evaluation results when randomly discard images. The connection between GT quality and model evaluations does not exist in this scenario.

based models, the proposed RQI can still achieve good consistency with humans on varying cases.

#### **D.** Ablation study

In this section, we show more ablation results of the proposed RQI scheme. Since we propose training RQI with image pairs that contain arbitrary distortions, the images contain distortions across types and levels. Therefore, we compare training RQI with image pairs containing the same type of distortions, denoted as RQI<sub>single distortion</sub>. We also compare testing RQI on single-scale images, instead of cropping multi-scale patches, denoted as RQI<sub>single-scale</sub>. The models are compared with our full model on user opinions collected from four datasets DIV2K [2], RealSR [6], DRealSR [33], and Set5&Set14 [3, 39]. RQI<sub>single-scale</sub> is not tested on Set5&Set14, since the image resolutions are low and cannot be down-scaled. The results are shown in Table 4.

From Table 4, several observations can be made. First, there is a significant performance drop when training RQI merely on same-distortion image pairs. We attribute this to the different types of distortions that GT and model outputs may contain. GTs usually contain common distortions (noise, blur, compression artifacts *.etc*), while model outputs are introducing algorithms-caused distortions, featuring different representations from each other. However, by training RQI on arbitrary distortions, the model is capable of measuring quality that covers complex distortions, showing superior performance on all four datasets. Second,

we also observe a slight performance improvement on RealSR [6], but a relatively larger performance gain on DIV2K [2] and DRealSR [33] when testing on multi-scale patches. Since image resolutions in DIV2K [2] and DRealSR [33] are relatively larger, evaluating RQI in multi-scale not only captures detailed texture quality but also measures structure and semantic consistency, leading to better alignment with human perception.

#### E. More qualitative comparisons

In this section, we provide more qualitative comparisons to show how different types of metrics can fail and the RQI results. For easier comparison, all scores are normalized to [0, 1], and a higher score indicates better visual quality. Figure 18 shows two distortion-based FR-IQA metrics SSIM [32] and PSNR, where they favor blurry regions more than detailed textures, leading to contradictory predictions with human perception. Figure 19 shows four NR-IQA metrics PI [5], NIQE [24], Clip-IQA [29] and MANIQA [36], all of which can fail on subtle structure or semantic changes due to the lack of references. Figure 20 shows two perceptionbased FR-IQA metrics LPIPS [44] and DISTS [12]. As can be seen, the two metrics suffer from poor GT quality, thus can not make correct evaluations when model outputs show better quality. As a comparison, ROI makes correct evaluations on all the above cases, showing its superiority as a reliable image metric for SR evaluations.



Figure 11. GT samples (left side) with relatively poorer quality from DIV2K dataset [2], and model outputs (right side) that show better quality than GTs.



Figure 12. GT samples (left side) with relatively poorer quality from RealSR dataset [6], and model outputs (right side) that show better quality than GTs.



Figure 13. GT samples (left side) with relatively poorer quality from DRealSR dataset [33], and model outputs (right side) that show better quality than GTs.



Figure 14. GT samples (left side) with relatively poorer quality from Set5 [3] and Set14 [39], and model outputs (right side) that show better quality than GTs.



Figure 15. We show that StableSR [30] tends to increase image contrast by sharpening the edges. This may increase perceptual quality for thin edges (left case) but produce ringing artifacts for strong edges (right case), reducing image quality in contrast.



(a) GT(b) SeeSR(c) PASD(d) GT(e) SeeSR(f) HATFigure 16. We show that though SeeSR [34] improves image quality in most cases, it alters image semantics in some cases, affecting user<br/>perception.



Figure 17. We show that PASD [37] tends to produce out-of-focus effect (left side). When this effect is produced on wrong subjects (right side), the perceptual quality will be reduced.



Figure 18. Distortion-based FR-IQA metrics SSIM [32] and PSNR tend to favor blurry regions over textures, leading to contradictory predictions with human perception.

Dataset	DIV2K [2]		RealSR [6]			DRealSR [33]			Set5&Set14			
Creterion	SRCC	PLCC	Win Rate	SRCC	PLCC	Win Rate	SRCC	PLCC	Win Rate	SRCC	PLCC	Win Rate
RQI <sub>single distortion</sub>	0.653	0.691	0.58	0.487	0.474	0.47	0.416	0.487	0.52	0.649	0.656	0.33
RQI <sub>single-scale</sub>	0.721	0.758	0.63	0.490	0.479	0.48	0.493	0.550	0.53	-	-	-
RQI <sub>full</sub>	0.744	0.785	0.65	0.504	0.484	0.49	0.529	0.603	0.53	0.664	0.673	0.35

Table 4. Ablation study of the RQI scheme.



Figure 19. NR-IQA metrics PI [5], NIQE [24], Clip-IQA [29] and MANIQA [36] can easily fail on cases where subtle structure of semantics are changed, due to the lack of proper references.



Figure 20. Perception-based FR-IQA metrics LPIPS [44] and DISTS [12] can fail when GT quality is relatively lower. They make contradictory evaluations for models that output perceptually higher results than GTs.

#### References

- Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 8
- [2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 3, 5, 7, 9, 10, 11, 13
- [3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012. 5, 7, 9, 10, 11
- [4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In CVPR, 2018. 4, 6
- [5] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCVW*, 2018. 3, 6, 7, 10, 13
- [6] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 1, 2, 3, 5, 7, 9, 10, 11, 13
- [7] 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 TIP*, 33:2404–2418, 2024. 4, 6, 9
- [8] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 2
- [9] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image superresolution transformer. In *CVPR*, 2023. 2, 3, 4
- [10] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, 2023.
- [11] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 2
- [12] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2020. 2, 4, 6, 7, 8, 10, 13
- [13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2
- [14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2015. 2
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2
- [16] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *ECCV*, 2020. 1, 5, 6
- [17] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021. 3
- [18] 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 *CVPR*, 2022. 4, 6, 9

- [19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 1, 2, 3, 4
- [20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In CVPRW, 2017. 1, 2
- [21] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *ICME*, 2019. 5, 6, 7, 8
- [22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2
- [23] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 7, 8
- [24] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Sign. Process. Letters*, 20(3):209–212, 2012. 3, 6, 7, 10, 13
- [25] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In CVPR, 2017. 8
- [26] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *CVPR*, 2018. 3, 5, 6
- [27] Shaolin Su, Vlad Hosu, Hanhe Lin, Yanning Zhang, and Dietmar Saupe. Koniq++: Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects. In *BMVC*, 2021. 3, 9
- [28] Louis L Thurstone. Psychophysical analysis. *The American journal of psychology*, 38(3):368–389, 1927. 5
- [29] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In AAAI, 2023. 3, 6, 7, 8, 10, 13
- [30] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 132(12):5929– 5949, 2024. 1, 2, 3, 5, 9, 12
- [31] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 1, 2, 3
- [32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 2, 4, 6, 7, 8, 10, 12
- [33] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, 2020. 1, 2, 3, 5, 7, 9, 10, 11, 13
- [34] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 1, 2, 3, 4, 5, 9, 12
- [35] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang.

One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, 2025. 1

- [36] 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 *CVPR*, 2022. 3, 4, 5, 6, 7, 9, 10, 13
- [37] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, 2024. 1, 2, 3, 4, 5, 9, 12
- [38] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, 2017. 8
- [39] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference*, 2012. 5, 7, 9, 10, 11
- [40] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. 1, 2, 3
- [41] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE TIP*, 20(8):2378–2386, 2011. 2
- [42] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched

completely blind image quality evaluator. *IEEE TIP*, 24(8): 2579–2591, 2015. 3

- [43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3, 5
- [44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 4, 6, 7, 8, 10, 13
- [45] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image superresolution. In ECCV, 2022. 2
- [46] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2
- [47] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In CVPR, 2018. 1, 2
- [48] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In CVPR, 2018. 2
- [49] Wei Zhou and Zhou Wang. Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity. In ACM MM, 2022. 7, 8