Bridging the Perception Gap in Image Super-Resolution Evaluation

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

In this supplementary, we provide additional details and results to complement our main submission, including:

1. Details and more analysis of the user experiment.
2. Implementation details of RQI.
3. Ablations on different training and testing settings of the proposed RQI scheme.
4. Additional visual comparisons for RQI evaluation.
5. Additional details and results when training with RQI.
6. Limitations and Discussions.

1. User Experiment

1.1. User Study Details

Since image quality comparison requires fine-grained perceptual judgment, we conduct the user study under a strictly controlled environment, unlike some prior user studies [15, 16]. Specifically, the study was performed in a matte dark room, where the display served as the only light source. All images were shown on a 3K-resolution monitor that had been calibrated to the sRGB color space. Before the experiment, all participants received detailed instructions to the visual comparison task. During each trial, two images from the same LR source were presented side-by-side, and participants were asked to choose the one with better perceptual quality. The order and left–right placement of each pair were randomized to minimize positional bias. For the DRealSR dataset, because the HR images exceed 4K resolution, only center-cropped images are selected for all the evaluations to prevent scaling artifacts. Each image pair received ratings from at least 15 participants, all of whom passed the Ishihara color-vision test.

1.2. Analysis of User Study Results

In this subsection, we conduct a brief analysis of the user results. Specifically, we analyze the overall perceptual preferences among the evaluated SR models and the distribution over user preferences. Figure 1 presents the average Thurstone scores across all datasets, and Figure 2 presents the proportion of best-ranked images for each evaluated model. Since Set5 [2] and Set14 [19] share similar content and contain only a limited number of images, we combine them for the analysis. Several noteworthy observations can be drawn from the statistics.

First, diffusion-based models are generally favored by human observers due to their strong ability to synthesize visually appealing and richly detailed textures. In most cases, these models deliver superior perceptual quality. Second, although ground-truth (GT) images achieve the highest average scores overall, Figure 2 reveals that outputs from certain models can perceptually surpass their corresponding GTs. The proportion of such cases varies with dataset quality — for example, only a small fraction in DIV2K [1], but more than half in Set5 and Set14 [2, 19]. Third, during the user study, we observe that localized hallucination artifacts produced by diffusion-based SR models can strongly influence human preference. Figure 3 illustrates an example from the RealSR [4] dataset, where a white cloud is incorrectly synthesized as a seagull wing by the advanced SeeSR [14] model, partially driven by its strong generation capability inherited from tag-style prompts. Although the hallucination occurs in a small region, it substantially affects users' subjective judgments. Such novel hallucination phenomena also introduce additional challenges for existing image quality metrics.

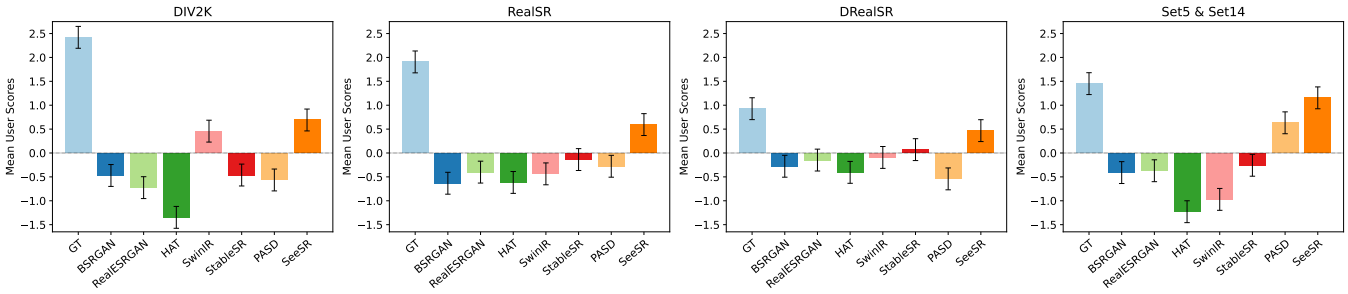

Figure 1. Average user scores on different SR models (including GT) in four SR testing datasets. Error bars correspond to a 95% confidence interval.

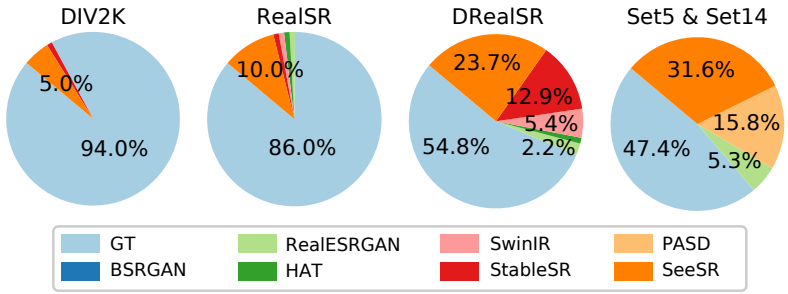

Figure 2. User statistics of the best quality HR image in four SR datasets.

(a) GT (b) SeeSR (c) HAT

Figure 3. We show that sometimes SeeSR [14] produce hallucinations due to its strong generative capability inherited from tag-style prompts. Such degradation may exists in a small area, yet significantly affect the user perception of image quality.

1.3. More Analysis of Table 1 (main paper)

Here we provide more analysis of the evaluation results for different metrics, shown in Table 1 from the original paper.

- Pixel-based metrics such as PSNR and SSIM [12] exhibit negative correlations with human perceptual preference. This is primarily because these distortion-oriented measures tend to regress toward the average solution across multiple plausible reconstructions, a behavior inherent to the ill-posed nature of the SR problem. This mismatch reflects a fundamental contradiction in SR evaluation: metrics designed to favor pixel-wise fidelity fail to capture the perceptual realism that modern SR models aim to produce. Since this mismatch fundamentally affects the evaluation of almost all images, the two method performances appears consistently poor across all metrics.
- Although we identify the limitation of lacking reference for NR-IQA methods in making fair evaluations, such alteration of textures/structures may exist only in some of the images. This partly explains the overall good performance of NR metrics (NIQE [9], PI [3], Clip-IQA [11] and MANIQA [17], *etc*).
- We observe a substantial consistency improvement of RQI on the DIV2K dataset [1]. Since most GT images in DIV2K are of high quality, this improvement highlights RQI's strong ability to distinguish fine-grained perceptual differences. Meanwhile, the strong consistency achieved on the DRealSR dataset [13] can be attributed to RQI's robustness to imperfect

GTs, since in this dataset SR models more frequently produce results that surpass the GTs in perceptual quality (see Figure 2).

- RQI performs slightly worse than DeQA-Score [18] on Set5&Set14, we attribute this to the limited resolutions of Set5 and Set14. Benefited from large-scale pretraining, LLM-based method DeQA-Score is less affected. Nevertheless, RQI still achieves better consistency than the other non-LLM metrics.

2. Implementation Details

For the two FR-IQA models AHIQ [7] and TOPIQ [5], we simply remove the last activation layer to ensure the models produce bidirectional outputs. Fot the NR-IQA model MANIQA [17], we modify it to receive two images as input. Specifically, features extracted from the two inputs are concatenated before the transposed attention block to enable effective cross-image fusion.

During training, all the selected models are trained following their official implementations. 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 [7], and 8 for MANIQA [17] and TOPIQ [5]. AHIQ [7] and MANIQA [17] randomly crop image patches with size 224, while TOPIQ [5] randomly crop image patches with size 384. The crops are randomly flipped during training for augmentation. We split the datasets into training and validation subsets (8:2) with non-overlapping scenes, and select the best-performing models upon their performances on the validation set.

During testing, we randomly crop patches from the inputs, ensuring that each patch pair is taken from the same spatial region of the SR output and its corresponding GT. Since SR evaluation often involves high-resolution inputs, we assess perceptual quality at three different spatial scales. Specifically, we use the original resolution as well as images downsampled by factors of $\times 2$ and $\times 3$. At each scale, we randomly crop 20 patches and compute the final score by averaging across all patches. The downsampling and patch cropping are applied only to images whose resolutions exceed the required input size of the IQA models (i.e., 224 for AHIQ [7] and MANIQA [17], and 384 for TOPIQ [5]).

3. 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 $\text{RQI}_{\text{single distortion}}$. We also compare testing RQI on single-scale images, instead of cropping multi-scale patches, denoted as $\text{RQI}_{\text{single-scale}}$. The models are compared with our full model on user opinions collected from four datasets DIV2K [1], RealSR [4], DRealSR [13], and Set5&Set14 [2, 19]. $\text{RQI}_{\text{single-scale}}$ is not tested on Set5&Set14, since the image resolutions are low and cannot be down-scaled. The results are shown in Table 1.

Table 1. Ablation study of the RQI scheme.

Dataset	DIV2K [1]			RealSR [4]			DRealSR [13]			Set5 [2]&Set14 [19]		
Creterion	SRCC	PLCC	Win Rate	SRCC	PLCC	Win Rate	SRCC	PLCC	Win Rate	SRCC	PLCC	Win Rate
$\text{RQI}_{\text{single distortion}}$	0.653	0.691	0.58	0.487	0.474	0.47	0.416	0.487	0.52	0.649	0.656	0.33
$\text{RQI}_{\text{single-scale}}$	0.721	0.758	0.63	0.490	0.479	0.48	0.493	0.550	**0.53**	-	-	-
RQI_{full}	**0.744**	**0.785**	**0.65**	**0.504**	**0.484**	**0.49**	**0.529**	**0.603**	**0.53**	**0.664**	**0.673**	**0.35**

From Table 1, several observations can be drawn. First, training RQI exclusively on same-distortion image pairs leads to a clear performance drop. We attribute this to two main factors. (1) GT and SR images typically exhibit different types of distortions. GT images often contain natural distortions such as noise or blur, whereas SR outputs introduce algorithm-induced artifacts with different characteristics. Training RQI across heterogeneous distortion types allows the model to learn a broader representation that better captures complex and diverse degradations. (2) Image pairs constructed from the same distortion type often carry large and obvious quality differences, providing limited fine-grained supervision. As a result, the model becomes less capable of handling subtle perceptual distinctions. Second, when applying multi-scale patch evaluation, we observe slight performance gains on RealSR [4], but notably larger improvements on DIV2K [1] and DRealSR [13]. Since images in DIV2K and DRealSR are of higher resolution, 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.

In Table 2, we train under RQI using different losses and report the SRCC results on varying datasets. As can be seen, both L1 and L2 losses perform slightly worse than the Huber loss.

Table 2. SRCCs results for selecting different losses under RQI.

Metric	DIV2K	RealSR	DRealSR	Set5Set14	BSD	QADS	SRIQA
RQI_{L1}	0.704	0.473	0.496	0.655	0.862	0.910	0.691
RQI_{L2}	0.712	0.479	0.514	0.649	0.881	0.905	0.725
RQI	0.744	0.504	0.529	0.664	0.901	0.912	0.733

In Table 3, we show how RQI variants (different models trained on varying datasets under the RQI scheme) perform on the four public IQA benchmarks, where $SRCC_{mean}$ are reported. Comparing with other models in Table 3 of the main paper, all variants generally perform well across benchmarks.

Table 3. Consistency evaluations of RQI variants on four IQA benchmarks.

Model/Dataset	BSD-SR	QADS	SRIQA-Bench	Kadid-10K
AHIQ/Kadid-10K	0.838	0.855	0.582	-
AHIQ/PieAPP	0.875	0.903	0.592	0.706
AHIQ/PIPAL	0.896	0.908	0.752	0.592
MANIQA/Kadid-10K	0.842	0.866	0.571	-
MANIQA/PieAPP	0.880	0.936	0.615	0.870
MANIQA/PIPAL	0.901	0.912	0.733	0.669
TOPIQ/Kadid-10K	0.844	0.896	0.665	-
TOPIQ/PieAPP	0.838	0.885	0.635	0.763
TOPIQ/PIPAL	0.813	0.867	0.687	0.520

4. More Qualitative Comparisons

In this section, we provide more qualitative comparisons to show how RQI outperforms different types of metrics. For easier comparison, all scores are normalized to $[0, 1]$, and a higher score indicates better visual quality. Figure 4 compares RQI with two distortion-based FR-IQA metrics (SSIM [12] and PSNR) in assessing detailed textures, Figure 5 compares RQI with four NR-IQA metrics (PI [3], NIQE [9], Clip-IQA [11] and MANIQA [17]) in evaluating subtle structure or semantic changes. Figure 6 compares RQI with two perception-based FR-IQA metrics (LPIPS [20] and DISTS [6]) in poor GT quality cases. As can be seen, RQI makes correct evaluations on all the cases, showing its superiority as a reliable image metric for SR evaluations.

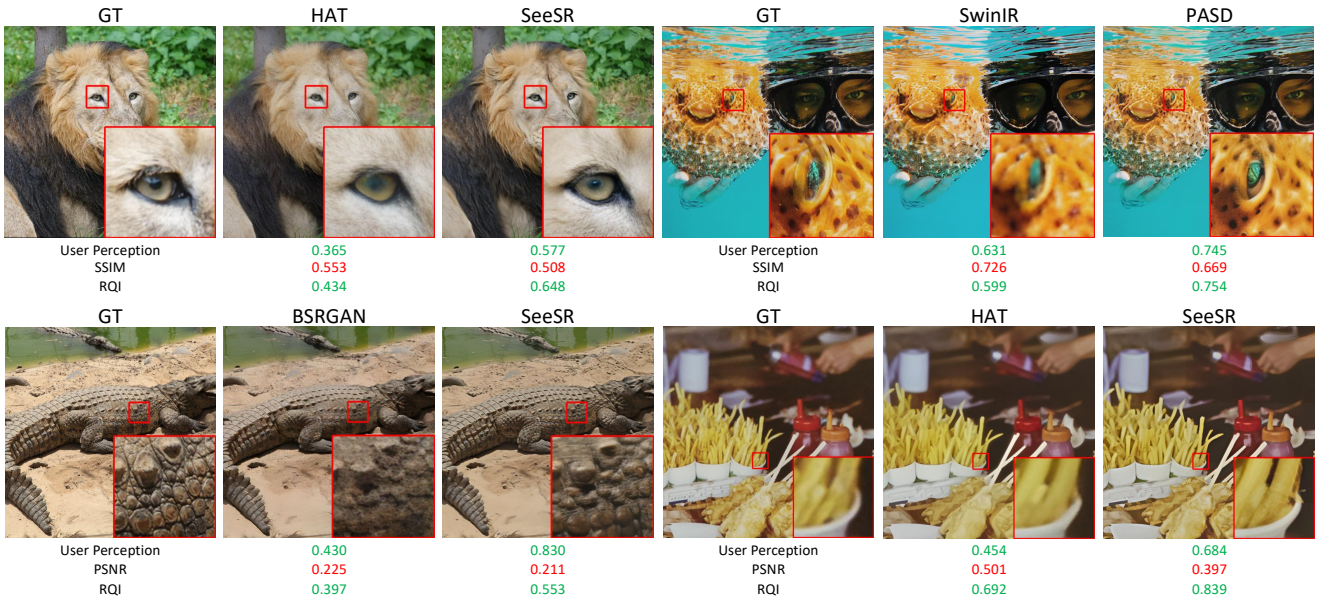

Figure 4. Distortion-based FR-IQA metrics SSIM [12] and PSNR tend to favor blurry regions over textures, leading to contradictory predictions with human perception.

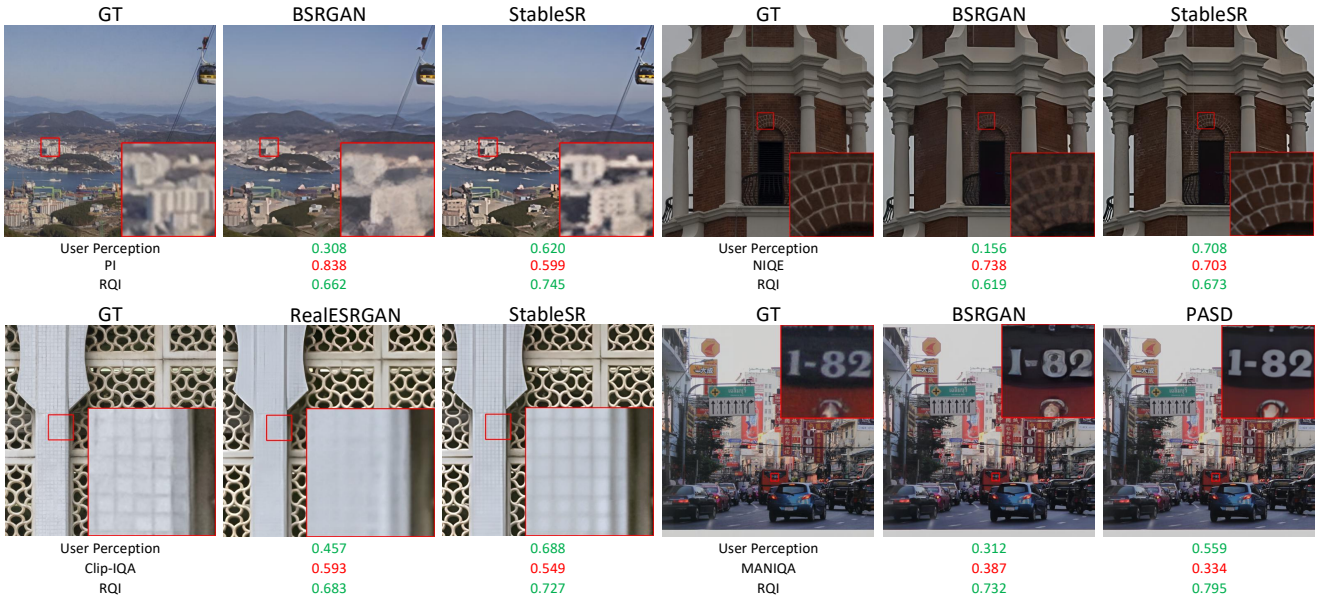

Figure 5. NR-IQA metrics PI [3], NIQE [9], Clip-IQA [11] and MANIQA [17] can fail on cases where subtle structure of semantics are changed, due to the lack of proper references.

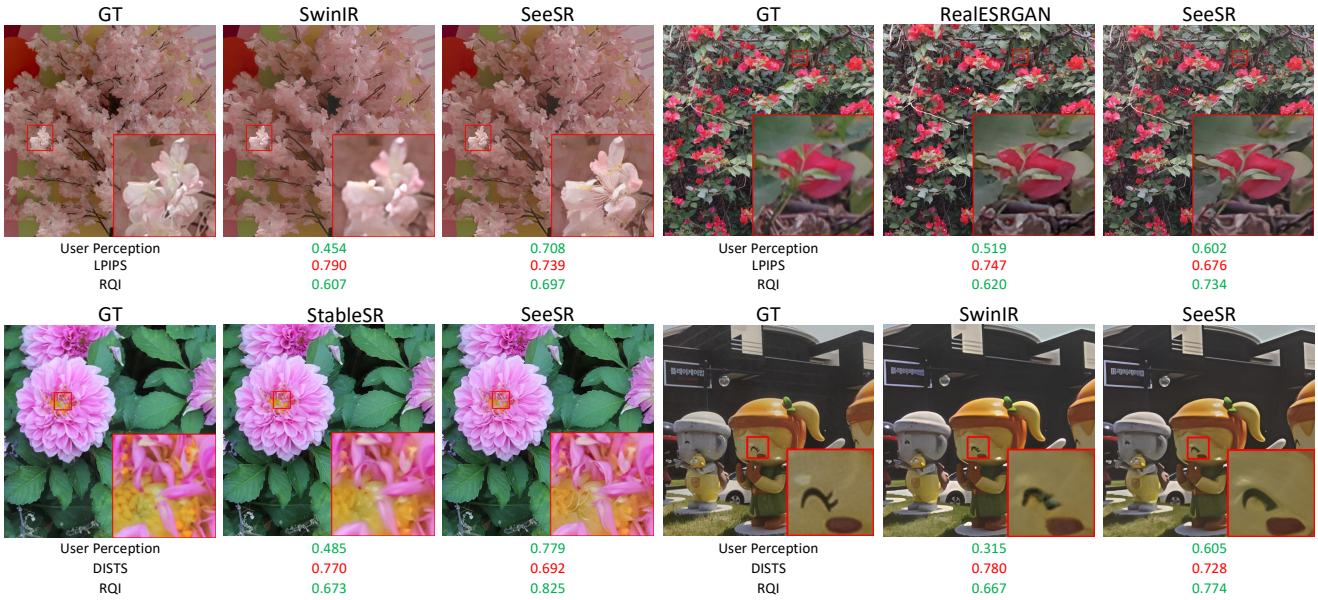

Figure 6. Perception-based FR-IQA metrics LPIPS [20] and DISTS [6] can fail when GT quality is relatively lower. They make contradictory evaluations for models that output perceptually higher results than GTs.

5. Training SR Models with RQI Loss

5.1. Training Details

We train three SR models, SwinIR [8], SeeSR [14] and PiSA-SR [10], following their official training configurations. For SwinIR, the model is first trained with MSE loss for 100K epochs, then trained together with the perceptual loss, GAN loss and RQI loss for another 100K epochs. For PiSA-SR, we adopt the same two-stage strategy as in the original paper: we first train the pixel-level LoRA using the standard pixel-wise loss for 4K iterations, and then train the semantic-level LoRA by introducing the RQI loss for 8.5K iterations, as this component is responsible for the reconstruction of perceptual details. For

SeeSR [14], because its loss is defined in the latent feature space, we follow its pipeline by decoding the latent representations using the frozen decoder and computing the RQI loss on the reconstructed images, where we train 150K iterations for the whole model.

5.2. More Qualitative Comparisons between w/ and w/o RQI Loss

In Figure 7 and Figure 8, we show more visual comparisons when training SR models with the RQI loss. Since the generation ability is limited for SwinIR [8] due to its non-diffusion architecture, we compare with SeeSR [14] or PiSA-SR [10] as baseline models in the figures. RQI is shown particularly effective in preserving structural fidelity (Figure 7) and producing more visually appealing textures (Figure 8). These results further demonstrate the advantages of the proposed approach.

6. Limitations and Discussions

Although RQI has demonstrated strong effectiveness in both evaluating and optimizing SR models, it also has a couple of limitations, as discussed below.

First, the RQI score is meaningful when comparing images that share the same reference. It cannot be used for cross-content quality comparison, because different contents may correspond to references with varying quality levels, making the scores incomparable across images.

Second, RQI primarily measures perceptual quality. In SR problem, fine textures are often irreversibly lost during the degradation process, making pixel-accurate fidelity evaluation fundamentally impossible. As a result, RQI assesses only the perceptual quality of these reconstructed textures. Designing a more principled fidelity measure for such texture-level comparisons remains an important direction for future work.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 1, 2, 3

[2] 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. 1, 3

[3] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCVW*, 2018. 2, 4, 5

[4] 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, 3

[5] 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. 3

[6] 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. 4, 5

[7] 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. 3

[8] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 5, 6

[9] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Sign. Process. Letters*, 20(3):209–212, 2012. 2, 4, 5

[10] 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. In *CVPR*, 2025. 5, 6

[11] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 2, 4, 5

[12] 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

[13] 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. 2, 3

[14] 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, 5, 6

[15] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 1

[16] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 1

[17] 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. 2, 3, 4, 5

[18] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *CVPR*, 2025. 3

[19] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference*, 2012. 1, 3

[20] 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. 4, 5

Figure 7. More visual comparison of training advanced SR models with RQI as an auxiliary loss. RQI is effective in preserving structural fidelity. Please zoom in for a better view.

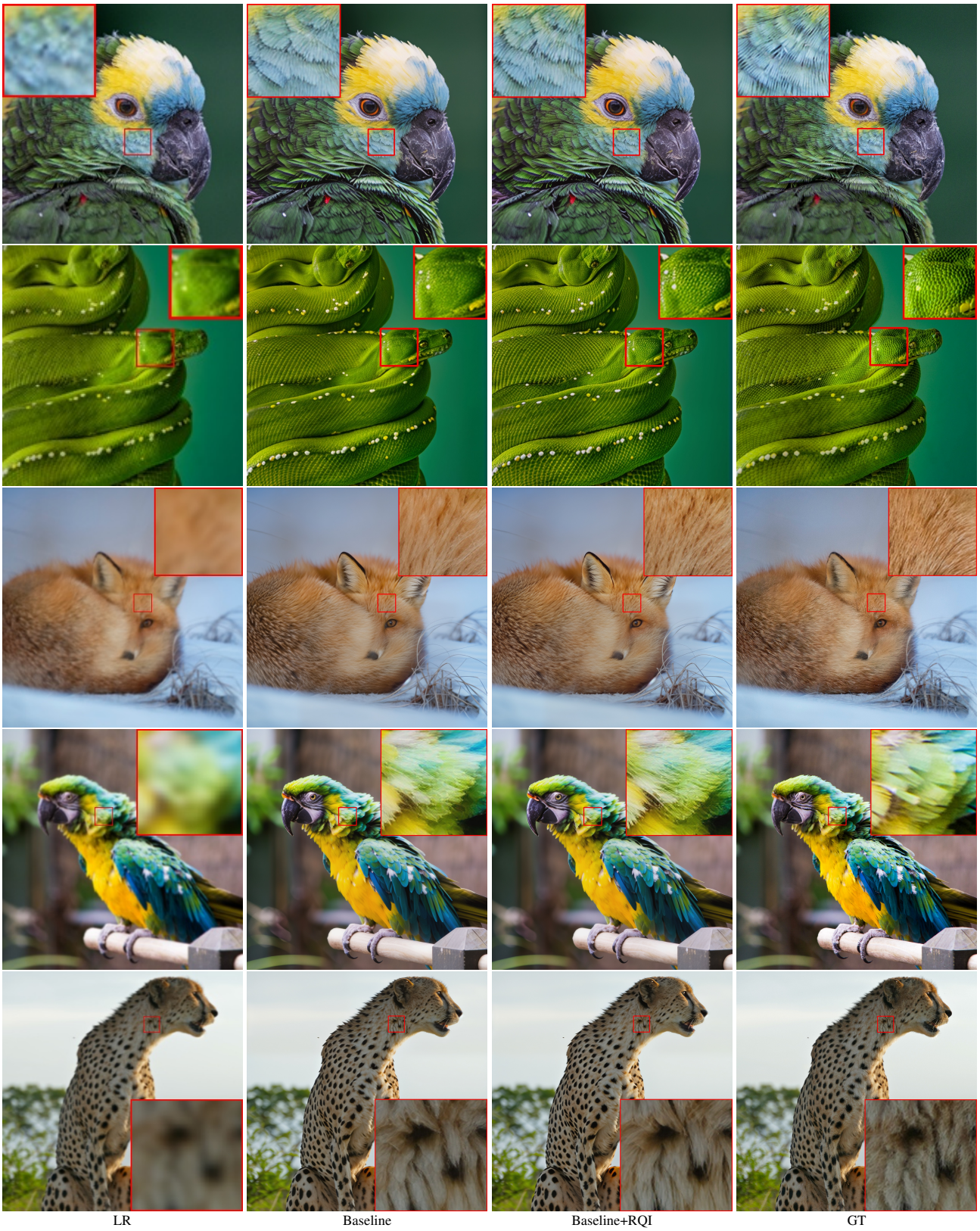

Figure 8. More visual comparison of training advanced SR models with RQI as an auxiliary loss. RQI is effective in generating more visually appealing details. Please zoom in for a better view.