Exploring Semantic Feature Discrimination for Perceptual Image Super-Resolution and Opinion-Unaware No-Reference Image Quality Assessment

Supplementary Material

1. Appendix Section

The supplementary material mainly includes the following contents:

- The specific structure of certain used networks;
- More detailed explanations of our proposed methods;
- More detailed explanations of the experimental settings;
- Additional experimental results and detailed analysis;
- Limitations and future work of our method.

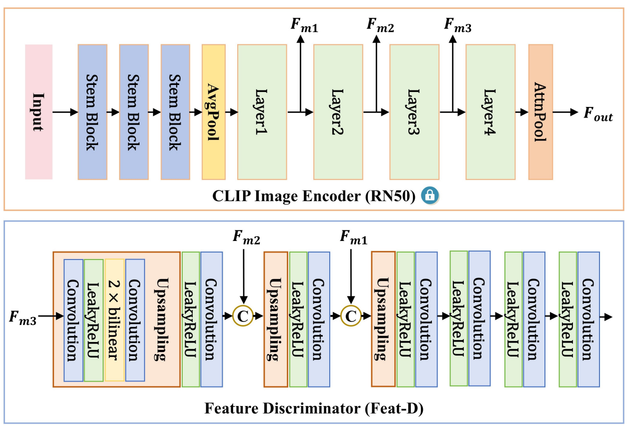


Figure 1. Detailed network structures of the CLIP image encoder and the proposed Feat-D.

2. Detailed Network Structures.

As described in the main document, considering that CLIP [14] can extract more interpretable, semantic-aware and quality-relevant features [4, 5, 9, 16], we select CLIP's image encoder as the semantic feature extractor and employ the proposed feature discriminator (Feat-D) and text-guided discrimination (TG-D) to perform discrimination on the semantic features. As shown in Fig. 1, we present detailed network structures of the CLIP image encoder and the proposed feature discriminator (Feat-D). We select the pixel-wise semantic features F_{m1}, F_{m2}, F_{m3} after Layer1, Layer2, and Layer3 of the CLIP image encoder, as well as the more abstract final output features F_{out}, then Feat-D and TG-D is used to perform discrimination on them respectively.

As for Feat-D, it begins with a combination of an upsampling layer, a LeakyReLU activation function, and a convolutional layer. The upsampling layer consists of a convolutional layer, a LeakyReLU activation function, a 2 $\times$ bilinear interpolation, and another convolutional layer. F_{m3} is processed through this combination and then concatenated with F_{m2}. The concatenated features are processed through another combination of upsampling, LeakyReLU, and convolution and subsequently stacked with F_{m1}. Finally, Feat-D ends with an upsampling layer followed by four combinations of LeakyReLU and convolutional layers. Notably, each convolutional layer in the Feat-D structure is followed by a spectral normalization. Our Feat-D is able to perform fine-grained discrimination on the 3 pixel-wise middle semantic features F_{m1}, F_{m2}, F_{m3} extracted from CLIP image encoder, encouraging the SR network to learn more accurate distributions of high-quality image semantic features.

3. More Analysis for Feat-D

To demonstrate the effectiveness of our Feat-D in semantic awareness, we have used t-SNE [15] to visualize the middle features of the VGG-style vanilla discriminator and our Feat-D in main text. Furthermore, we also visualize some features of two convolutional layers in SRN: one before upsampling and one after upsampling. As shown in Fig. 2, the intermediate semantic features of the SRN trained with our method are richer and clearer than ESRGAN which directly discriminates on images. In addition, we also use t-SNE to further explore the effectiveness of Feat-D in image quality assessment. We visualize the middle features of the vanilla discriminator, the CLIP image encoder, and our Feat-D. We classify the IQA dataset KonIQ10K [6] according to the label scores of the images, divide them into 5 categories every 20 points, and then randomly select 100 images in each category. We then fed all the images into the VGG-style vanilla discriminator, the CLIP Image encoder (RN50) and our Feat-D. We visualize the features after the 2nd BN layer of the vanilla discriminator, the features after Layer1 of CLIP image encoder, and the features after the 3rd upsampling layer of Feat-D, respectively. As shown in Fig. 3, the features of vanilla discriminator and CLIP image encoder are almost completely chaotic, while the features of Feat-D are more orderly and can distinguish images with different quality fractions more clearly, which proves that the image quality correlation of Feat-D features is much stronger. Our Feat-D takes the features of CLIP image encoder as input, so this also shows from another point that Feat-D can strengthen the quality correlation of the semantic features of CLIP.

4. More Explanation for TG-D

Apart from the middle features F_{m1}, F_{m2}, F_{m3} of CLIP image encoder, discriminating the final output feature F_{sout}

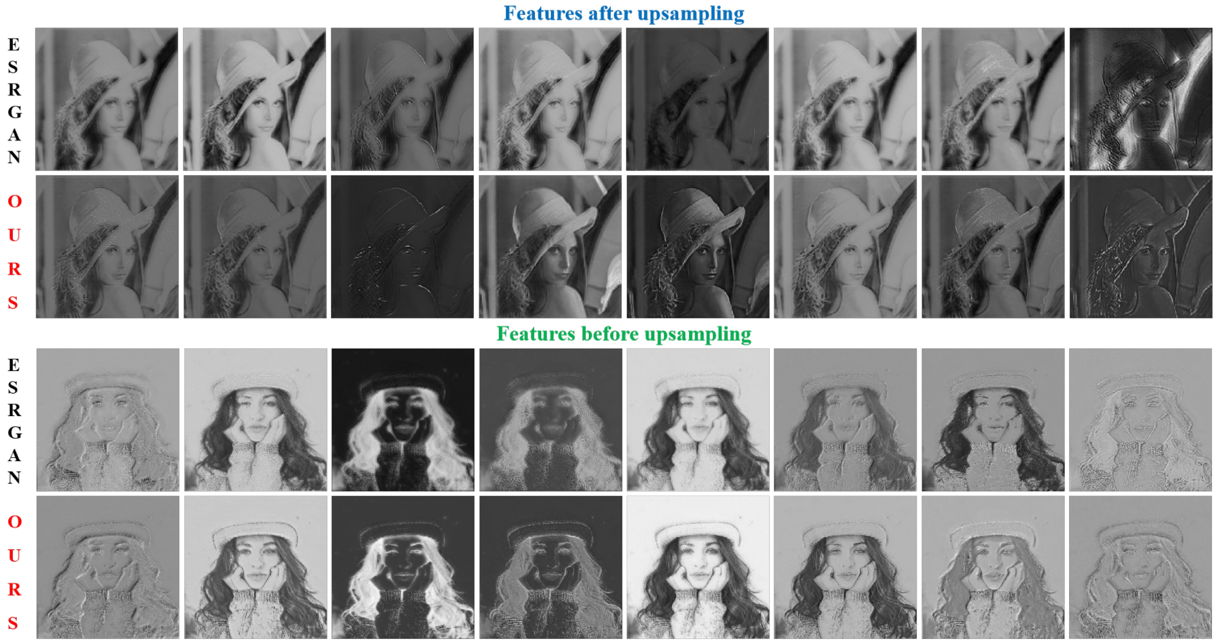
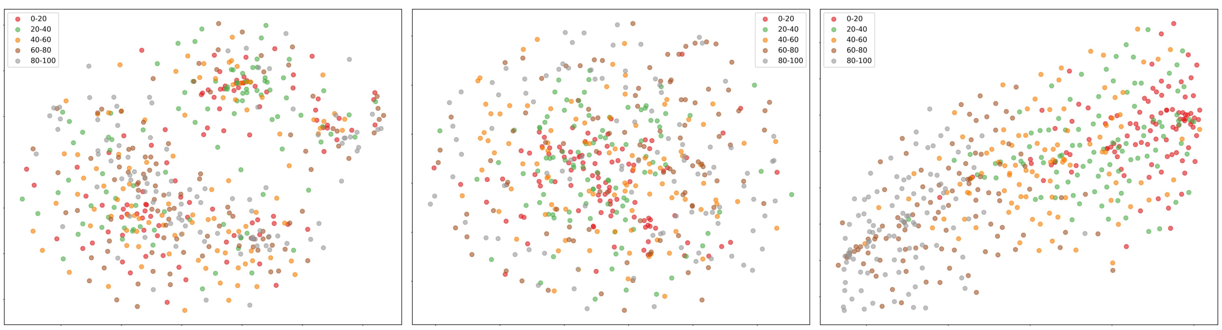


Figure 2. The feature visualization of two convolutional layers in SRN trained with ESRGAN and our SFD, respectively.



t-SNE of vanilla discriminator features t-SNE of CLIP image encoder features t-SNE of our Feat-D features

Figure 3. The t-SNE visualization of vanilla discriminator features, CLIP features and our Feat-D features. We divide the images from IQA dataset KonIQ-10K [6] into 5 categories according to their label scores, and randomly select 100 images for each category. The 5 categories of scores are "0-20", "20-40", "40-60", "60-80", "80-100", respectively.

which is more global and abstract is expected to further enhance the overall performance of our method. Before utilizing the learnable prompt pairs (LPP) to discriminate F_{out}, we have also considered other approaches. A simpler and more straightforward method is to adopt fixed antonymic text prompts (e.g., "Good photo" and "Bad photo" used in CLIPIQA [16]) to calculate the similarity scores between the image features of F_{out} and text features of antonymic text prompts. Then, two different approaches can be used to constrain the SR network's training process: maximizing the similarity scores of the SR images or making the similarity scores of the SR images closer to that of the HR images. We temporarily name the above method as CLIPIQA Loss. However, as illustrated in Fig. 4, we observe that the model trained with CLIPIQA loss will exhibit "mode collapse" when directly applied to some images, SRN may output SR images with anomalous pixel regions during testing. This is probably because that the IQA performance of CLIPIQA is limited due to the ambiguity of human language and CLIP's sensitivity to prompt selection, and a higher CLIPIQA score can't always represent higher perceptual image quality. In contrast, we introduce the learnable LPP in an adversarial learning manner to avoid the issues caused by text selection, and SRN trained with our method do not exhibit the "mode collapse" phenomenon.

5. More Experiment Results

More implementation details. During training, HR images is randomly cropped into 128×128 patches with batch size

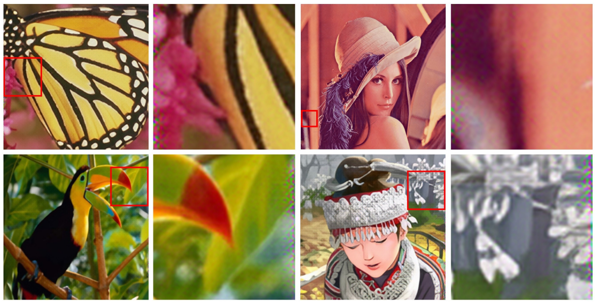


Figure 4. The "mode collapse" in SR images: SR networks trained with CLIPIQA loss may output anomalous pixel regions during testing on some datasets.

Table 1. Quantitative comparison of our method vs. other SOTA methods for $\times 4$ SR task. The best and the second-best are marked in red and blue, respectively.

Benchmark	Metric	ESRGAN [17]	LDL [10]	DualFormer [11]	SeD-P [9]	RRDB +Ours	SwinIR $+L_{GAN}$	SwinIR +Ours
BSDS100	PSNR ↑	25.33	25.97	26.59	26.38	26.90	25.58	27.06
	SSIM ↑	0.653	0.682	0.696	0.692	0.710	0.676	0.717
	LPIPS ↓	0.161	0.153	0.161	0.150	0.161	0.157	0.160
	DISTS ↓	0.116	0.118	0.120	0.117	0.121	0.122	0.118
Manga109	PSNR ↑	28.41	29.62	29.90	29.99	30.36	29.35	30.91
	SSIM ↑	0.860	0.873	0.886	0.888	0.893	0.880	0.902
	LPIPS ↓	0.065	0.054	0.053	0.048	0.048	0.054	0.045
	DISTS ↓	0.047	0.036	0.038	0.036	0.035	0.037	0.032

of 32 for classical SISR, and 256 × 256 patches for real-world SISR and OU NR-IQA. We initialize the parameters of RRDB with the pre-trained fidelity-oriented model. We use Adam [8] optimizer to train the network with a initial learning rate of 10^{-4}. The hyperparameters λ_1, λ_2, λ_3, λ_4 in total loss are set to 0.01, 1, 0.01, 0.005 for classic SISR and 1, 1, 0.1, 0.005 for real-world SISR, respectively. The length and number of learnable prompt pairs are set to 32 and 5 for classic SISR, as well as 64 and 1 for real-world SISR and OU NR-IQA. The weight coefficients α_1 and α_2 in the SFD-IQA process are set to 0.9 and 0.1, respectively.

More quantitative results for perceptual SISR. In Tab. 1, we present more quantitative results that can not be included in the main document due to space limitations, including results on BSDS100 [12] and Manga109 [13] datasets for classical SISR. Our method achieves the best PSNR and SSIM scores on all datasets while maintaining highly competitive perceptual metrics. This indicates that our method can achieve better PD trade-off, sacrificing less fidelity in exchange for improved perceptual image quality.

More qualitative results for perceptual SISR. We provide more qualitative comparisons with state-of-the-art (SOTA) GAN-based SR methods on both classical and real-world SISR tasks. As shown in Fig. 5, Fig. 6, Fig. 7, and Fig. 8, our method outperforms others by more accurately recovering fine-grained textures, especially in challenging details such as fur, buildings, and text, while producing fewer artifacts. This strongly demonstrates that our method effectively encourages the SR network to learn more fine-grained semantic feature distributions, leading to the generation of more realistic SR images.

Table 2. Ablation studies on different semantic feature extractors.

Extractors	Set5 [2]			DIV2K100 [1]		
	PSNR ↑	LPIPS ↓	DISTS ↓	PSNR ↑	LPIPS ↓	DISTS ↓
RN50	31.63	0.063	0.098	29.75	0.097	0.053
ViT-B/16	31.32	0.068	0.103	29.21	0.115	0.061

The effects of the different semantic feature extractors. To investigate the impact of different semantic feature extractors on our method, we conduct ablation experiments with 2 different semantic feature extractors of CLIP. Notably, since the feature scale of ViT-based extractior is different from that of ResNet-based extractiors, we only conduct ablation experiments on TG-D for ViT-based extractor. As shown in Tab. 2, our methods based on RN50 extractor outperform that of ViT-based extractor in terms of both perceptual quality and fidelity. This is because CNN-based extractor can extract semantic features with more positional information, which is more beneficial for low-level vision.

More results and analysis for OU NR-IQA. In the main document, we present a detailed comparisons of our SFD-IQA with other OU NR-IQA methods across both SR IQA datasets and authentically distorted IQA datasets. Our SFD-IQA achieves the best results on all metrics across all datasets. As explained in the main document, our SFD-IQA benefits from the dual advantages of CLIP and super-resolution discriminators, which explains its remarkable performance in OU NR-IQA tasks. The effectiveness of our method in the OU NR-IQA tasks also proves the discriminative ability of the proposed Feat-D and TG-D, which can encourage the SR network to learn more fine-grained semantic feature distributions.

Based on the analysis in Sec. 4, we further discuss the advantages of our SFD-IQA compared to CLIP-IQA. Due to the limitations of CLIP-IQA, CLIPIQA's ability to evaluate the quality of certain images is inadequate. As shown in Fig. 9, CLIPIQA will assign a wrong score for a SR image with "mode collapse", which is even higher than the better-look GT image without "mode collapse", this is obviously unreasonable. In contrast, our SFD-IQA can better distinguish between the high-quality and low-quality images and assign more reasonable scores for them. Since the difference between the top row of images and the bottom row of GT images is minimal except for small areas with "mode collapse", our SFD-IQA reasonably assigns similar scores to corresponding images while accurately distinguishing images with pixel anomaly regions using a certain score difference. Moreover, compared to "Lenna" on the left and "Pepper" on the right with lower overall perceptual quality, our SFD-IQA assigns relatively higher scores to the more

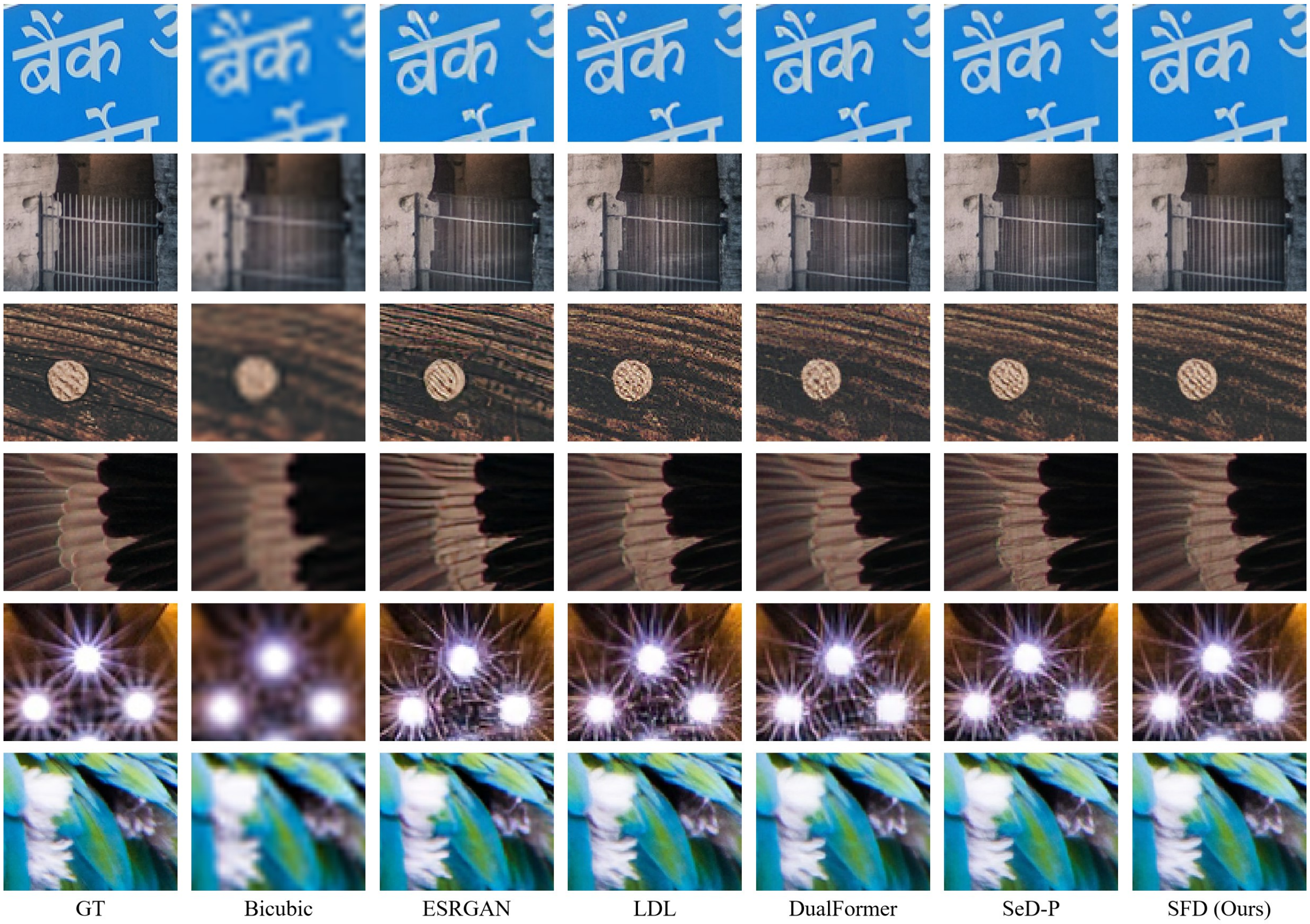


Figure 5. More visual comparisons of different GAN-based SR methods on DIV2K [1] validation set for ×4 classic super-resolution.

natural and realistic image “Head” in the middle, which better aligns with human visual perception. The above analysis further demonstrates the robustness of our SFD-IQA, showcasing its superior OU NR-IQA ability.

6. Limitations and Future Work

By introducing Feat-D and TG-D to perform discrimination on the semantic features from CLIP, we enable the SR network to generate more fine-grained and more realistic texture details, thereby achieving better perception-distortion trade-off. Despite these benefits, there is still room for further improving our method in balancing fidelity and perceptual quality, and our approach increases the computational and storage overhead during the training process. Additionally, exploring more efficient ways to integrate the trained Feat-D and LPP into the OU NR-IQA method is a worthwhile direction for further research.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 3, 4

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference (BMVC)*, 2012. 3

[3] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3086–3095, 2019. 5

[4] Jun Cheng, Dong Liang, and Shan Tan. Transfer clip for generalizable image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25974–25984, 2024. 1

[5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 1

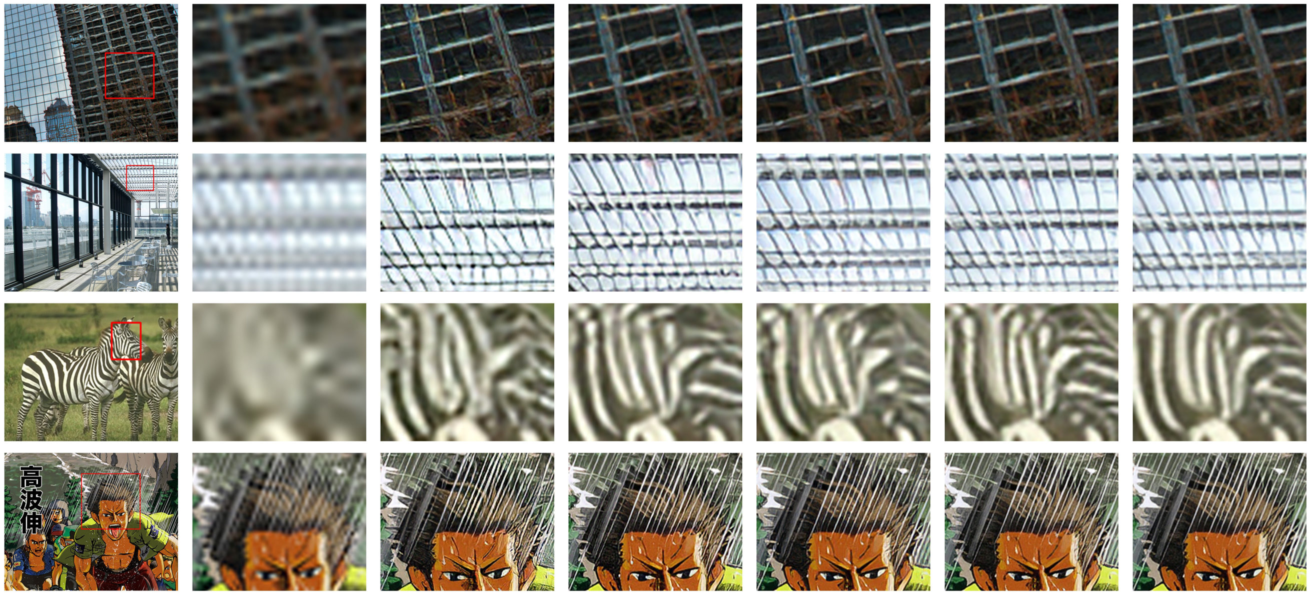


Figure 6. More visual comparisons of different GAN-based SR methods on Urban100 [7], BSDS100 [12], and Manga109 [13] datasets for ×4 classic super-resolution.

Figure 7. More visual comparisons of different real-world SR methods on RealSR [3] for ×4 real-world super-resolution.

[6] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 1, 2

[7] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 5

[8] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[9] Bingchen Li, Xin Li, Hanxin Zhu, Yeying Jin, Ruoyu Feng, Zhizheng Zhang, and Zhibo Chen. Sed: Semantic-aware discriminator for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25784–25795, 2024. 1, 3

[10] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 3

[11] Xin Luo, Yunan Zhu, Shunxin Xu, and Dong Liu. On the effectiveness of spectral discriminators for perceptual quality improvement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13243–13253, 2023. 3

[12] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, pages 416–423. IEEE, 2001. 3, 5

[13] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto,

Figure 8. More visual comparisons of different real-world SR methods on DrealSR [18] for ×4 real-world super-resolution.

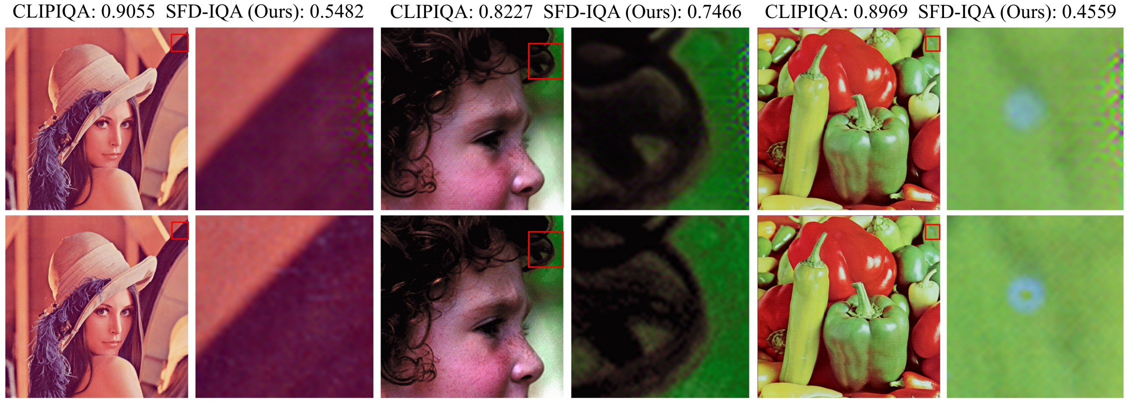


Figure 9. Comparisons between CLIPIQA and our SFD-IQA for OU NR-IQA. The top row consists of SR images with "mode collapse", while the bottom row contains the GT images. Compared to CLIPIQA, our SFD-IQA can more accurately distinguish the quality of images, even for highly similar ones, and assign them more reasonable scores.

Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia tools and applications*, 76:21811–21838, 2017. 3, 5

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[15] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[16] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 1, 2

[17] 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 *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3

[18] 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 *Computer Vision – ECCV 2020*, pages 101–117. Springer International Publishing, 2020. 6