This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Guanglu Dong Xiangyu Liao Mingyang Li Guihuan Guo Chao Ren*
College of Electronics and Information Engineering, Sichuan University, China
{dongguanglu, liaoxiangyu1, 2024222055110, guoguihuan}@stu.scu.edu.cn, chaoren@scu.edu.cn

Abstract

Generative Adversarial Networks (GANs) have been widely applied to image super-resolution (SR) to enhance the perceptual quality. However, most existing GAN-based SR methods typically perform coarse-grained discrimination directly on images and ignore the semantic information of images, making it challenging for the super resolution networks (SRN) to learn fine-grained and semantic-related texture details. To alleviate this issue, we propose a semantic feature discrimination method, SFD, for perceptual SR. Specifically, we first design a feature discriminator (Feat-D), to discriminate the pixel-wise middle semantic features from CLIP, aligning the feature distributions of SR images with that of high-quality images. Additionally, we propose a text-guided discrimination method (TG-D) by introducing learnable prompt pairs (LPP) in an adversarial manner to perform discrimination on the more abstract output feature of CLIP, further enhancing the discriminative ability of our method. With both Feat-D and TG-D, our SFD can effectively distinguish between the semantic feature distributions of low-quality and high-quality images, encouraging SRN to generate more realistic and semantic-relevant textures. Furthermore, based on the trained Feat-D and LPP, we propose a novel opinion-unaware no-reference image quality assessment (OU NR-IQA) method, SFD-IQA, greatly improving OU NR-IQA performance without any additional targeted training. Extensive experiments on classical SISR, real-world SISR, and OU NR-IQA tasks demonstrate the effectiveness of our proposed methods. Code is available at https://github.com/GuangluDong0728/SFD.

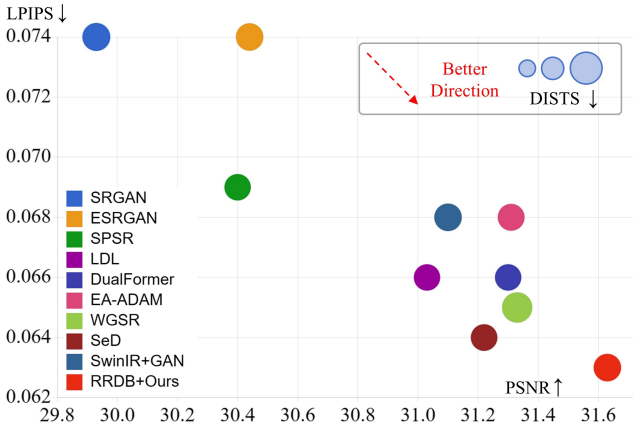


Figure 1. Perception-distortion trade-off comparison between our proposed SFD and other SOTA GAN-based SISR methods.

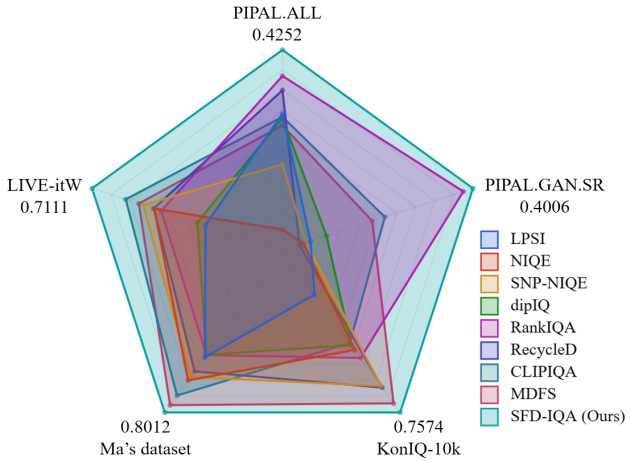
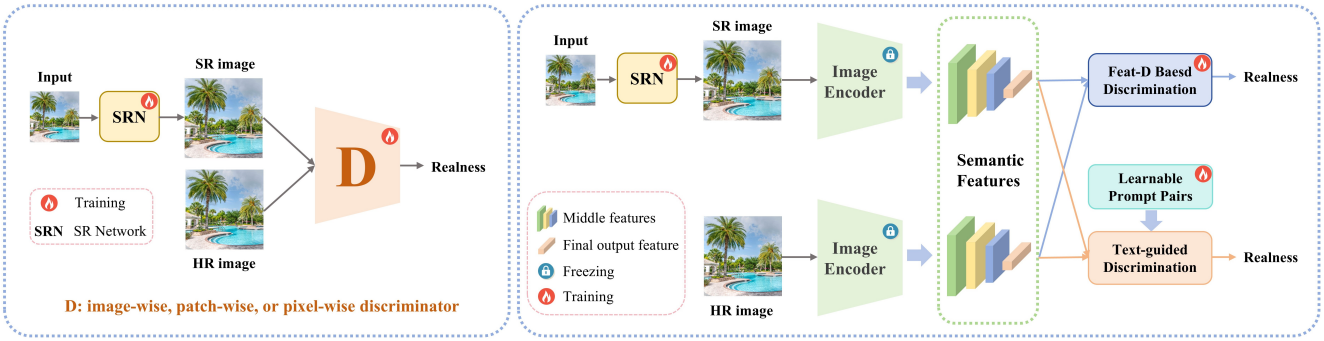


Figure 2. Relative PLCC comparation between the proposed SFD-IQA and other OU NR-IQA methods.

1. Introduction

As one of the most important tasks in low-level computer vision [14, 19, 33, 36, 50, 52, 60, 67], single image super-resolution (SISR) [11, 18, 28, 34, 56, 57, 62, 80] aims to accurately reconstruct high-resolution (HR) images from the given low-resolution (LR) counterparts. In recent years, with advancements in deep learning, SISR has achieved rapid and substantial progress. Since Dong et al. [11] first introduce deep learning to SISR task, numerous studies have emerged to explore more advanced network architectures, including

*Corresponding author



(a) Vanilla GAN-based SISR methods **(b) Our proposed SFD which performs discrimination on the semantic features**

Figure 3. (a) Vanilla GAN-based SISR methods: performing discrimination on images; (b) Our proposed SFD: performing discrimination on the semantic features. Vanilla GAN-based SISR typically uses an image-wise, patch-wise, or pixel-wise discriminator to determine if the distribution of SR images align with that of high-quality real-world images, ignoring the semantic information of images. Such coarse-grained discrimination makes it challenging for the SR network to reconstruct fine-grained, semantic-relevant textures. In contrast, we propose to perform discrimination on the semantic features, encouraging the SR network to generate more realistic semantic textures.

CNN-based models [28, 34, 56, 57, 81, 82] and Transformer-based models [7, 13, 31, 38, 72, 80], achieving continuous improvements in fidelity-oriented performance. However, these works primarily employ pixel-wise L1 or MSE losses for the training process, which often produce overly smooth results and fail to generate realistic texture details [4, 29].

To overcome the above limitations, a series of generative adversarial networks (GANs) [17] based SISR models [6, 15, 29, 30, 32, 40, 47, 48, 63, 65] have been proposed to generate more realistic images that align more closely with human visual perception by effectively balancing perceptual quality and fidelity. As shown in Figure 3 (a), most existing GAN-based methods perform discrimination on the images using an image-wise [54], patch-wise [23], or pixel-wise [53] discriminator to distinguish between the distributions of SR images and HR images. However, directly discriminating on the images is coarse-grained and lacks semantic awareness. This leads to that the textures of generated images do not match its semantic and often come with artificial artifacts, making it challenging to produce realistic semantic textures.

In this paper, different from existing works, we propose to perform discrimination on the semantic features, aiming to align the semantic feature distributions of SR images more closely with that of HR images. Considering that CLIP [51] can extract more interpretable, semantic-aware and quality-relevant features [9, 20, 30, 61], we select CLIP's image encoder as the semantic feature extractor and propose two methods to perform joint discrimination on the semantic features. Specifically, as shown in Figure 3 (b), as for the pixel-wise middle semantic features, we first design a feature discriminator, Feat-D, to distinguish between the feature distributions of SR images and HR images, encouraging the semantic feature distributions of SR images to match that of HR images. As for the final output feature which is more abstract and global, we propose a text-guided discrimination method, TG-D, by introducing learnable prompt pairs (LPP) into an adversarial training process to make them sensitive to global image quality and thus able to discriminate the abstract semantic features, further enhancing our method's overall discriminative capability. With both Feat-D and TG-D, the SR network is able to learn more fine-grained semantic feature distributions, ultimately producing more realistic and semantic-related texture details.

Additionally, existing researches [39, 40, 83] have shown that during the training process where the discriminators in GAN-based SISR learn to distinguish the realness of image distributions, they simultaneously and implicitly learn to assess image quality, making them highly applicable to opinion-unaware no-reference image quality assessment (OU NR-IQA). Simultaneously, recent studies [20, 44] have also shown the advantage of CLIP for OU NR-IQA. In this paper, to further improve OU NR-IQA performance, we propose a novel OU NR-IQA method, SFD-IQA, by integrating the well trained Feat-D and TG-D which conveniently combines both of the above advantages. Thanks to the superior semantic feature discrimination capability of Feat-D and TG-D and the inherent image quality correlation of CLIP, we greatly improve the OU NR-IQA performance without any additional task-specific training on IQA datasets.

As shown in Fig. 1, our SFD demonstrates a better perception-distortion (PD) trade-off on the PSNR-LPIPS-DISTS plane. Compared to other state-of-the-art (SOTA) GAN-based SISR methods, our LPIPS and PSNR scores are the best among all methods with competitive DISTS scores. Additionally, Fig. 2 illustrates a comparison between our SFD-IQA and other SOTA OU NR-IQA methods, where SFD-IQA outperforms other methods across both SR IQA datasets and authentically distorted IQA datasets. The main contributions of this work are summarized below:

- We propose a semantic feature discrimination (SFD)

method for perceptual SISR, which consists of a feature discriminator (Feat-D) based discrimination and a text-guided discrimination (TG-D). Our SFD can effectively distinguish between the CLIP semantic feature distributions of SR and HR images, encouraging SR networks to learn more fine-grained and semantic-related textures, and achieving better perception-distortion (PD) trade-off.
- We propose a novel OU NR-IQA method, SFD-IQA, by effectively integrating the two well trained feature discrimination methods. Without any additional task-specific training on IQA datasets, SFD-IQA can significantly enhance the OU NR-IQA performance for both SR images and other authentically distorted images.
- Extensive experiments on classical SISR, real-world SISR, and OU NR-IQA tasks demonstrate the effectiveness of our method. SFD can help SR networks achieve better PD trade-off and generate more realistic semantic textures, and SFD-IQA is able to greatly improve OU NR-IQA performance. Moreover, our approach can be easily integrated into other GAN-based image restoration methods.

2. Related Works

2.1. Fidelity-Oriented SISR

Recently, deep learning based single image super-resolution (SISR) [11, 18, 28, 29, 62, 70, 81, 82] has made tremendous advancements. Since SRCNN [11] first introduces CNN into SISR, numerous SISR backbones [18, 28, 34, 56, 57, 81, 82] have continued to emerge to enhance the performance of fidelity-oriented SISR. Specifically, RCAN [81] introduces channel attention into SISR network, and RDN [82] leverages dense connection to fully utilize hierarchical features for SISR. Furthermore, based on the success of the vision transformer [12] framework, [13, 31, 72, 80] further improve the SISR performance. However, in order to maximize fidelity, most of the aforementioned methods only employ pixel-wise loss functions to optimize the training process and ignore the perceptual quality, which often leads to overly smooth results and fails to achieve visually appealing results.

2.2. Perception-Oriented GAN-based SISR

To improve perceptual quality, GAN-based SISR [15, 27, 30, 32, 39, 47, 64, 65, 75, 79] has received constant attention in recent years. Specifically, SRGAN [29] introduces GANs into SR tasks to obtain SR images more in line with human perception. DualFormer [39] applies an improved spectral discriminator to improve perceptual quality. Sed [30] incorporates the semantic features into the vanilla discriminator to finely distinguish the image distributions. EA-ADAM [55] develops a new optimizer from a multi-objective optimization perspective to achieve a better perception-distortion tradeoff. However, due to the lack of semantic-relevant perception, most GAN-based SR networks are difficult to produce fine-grained semantic textures, result in exhibiting artificial artifacts. In this paper, we propose to discriminate the semantic features of CLIP to guide the SR network to learn more fine-grained semantic feature distributions and generate more realistic semantic textures.

2.3. Opinion-Unaware NR-IQA

OU NR-IQA [2, 43, 46, 61, 69, 71, 77, 83] aims to evaluate the quality of images without human-labeled data. Traditional OU NR-IQA methods [69, 76, 77] are typically based on natural scene statistics (NSS), which estimate distribution parameters to assess quality. Researches on deep learning-based OU NR-IQA are relatively sparse, with most methods centered around learning-to-rank [35, 43] and contrastive learning [44]. Recently, with the emergence of the pre-trained vision-language model CLIP [51], [20, 61] leverage the rich visual-linguistic priors encapsulated in CLIP to evaluate image perceptual quality. Additionally, RecycleD [83] proposes to recycle the trained discriminator for GAN-based SISR to predict perceptual quality. Following [83], [39] and [40] further enhance the IQA ability of the discriminators in GAN-based SISR. However, these discriminator-based methods only rely on the trained semantic-unaware discriminator and ignore the assessment for semantic, which is difficult to achieve satisfactory performance.

3. Method and Analyses

In this section, we provide a detailed explanation and analysis of our method. Firstly, we describe the overall framework of our method in Sec. 3.1. Next, we provide a detailed introduction and analysis of the two proposed semantic feature discrimination methods in Sec. 3.2. Lastly, we discuss how to extend our method to OU NR-IQA in Sec. 3.3.

3.1. Overall Framework

We propose a noval semantic feature discrimination (SFD) method for perceptual SISR, aiming to achieve better perception-distortion (PD) trade-off. Differing from vanilla GAN-based SR methods that only consider coarse-grained image distributions with poor semantic relevance, our SFD focuses on the distributions of semantic features. As illustrated in Fig. 4, given the input low-resolution image I_{lr}, we can obtain its corresponding SR image I_{sr} through the SRN. Then, through the pre-trained CLIP [51] image encoder E_i, we can obtain rich semantic features of I_{sr} with different scales and characteristics, including the multi-scale pixel-wise middle features F_{smi}, F_{smi}, $F_{smi} \in \mathbb{R}^{C_i \times H_i \times W_i}$ (i = 1,2,3), and the more abstract final output feature $F_{sout} \in \mathbb{R}^{1 \times L}$. To better leverage these semantic features to improve the robustness of our approach, we propose two methods to separately perform discrimination on them.

On one hand, as for the pixel-wise middle features, we design a feature discriminator (Feat-D) to distinguish be-

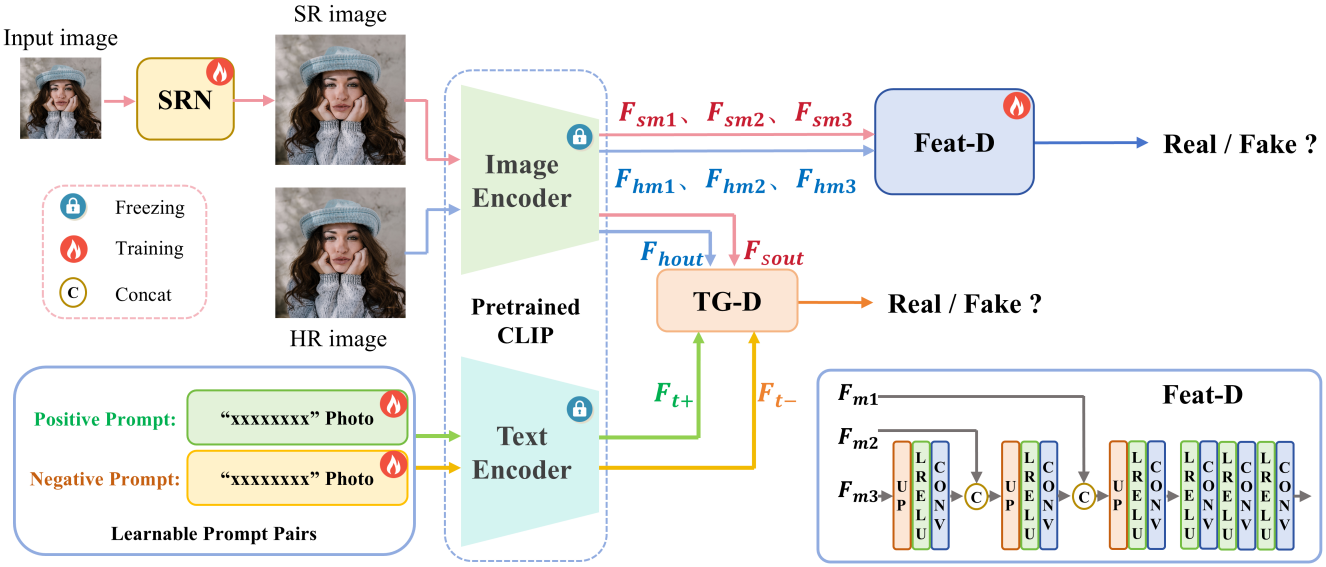


Figure 4. Framework of the proposed SFD. SFD consists of a feature discriminator (Feat-D) and a text-guided discrimination (TG-D) method, Feat-D and TG-D are used to perform discrimination on the middle features and final output features from CLIP, respectively.

tween the feature distributions of SR and HR images. On the other hand, we propose a text-guided discrimination (TG-D) method to perform discrimination on F_{sout}, further enhancing the overall discriminative ability. With both Feat-D and TG-D, SFD is able to facilitate the SR network (SRN) to learn more fine-grained semantic feature distributions and generate more realistic and semantic-related textures.

3.2. Discrimination on Semantic Features

Semantic Feature Discriminator. As for the middle features of CLIP which contain multi-scale pixel-wise semantic information [30], we first design a feature discriminator, Feat-D, to distinguish between the feature distributions of SR and HR images, encouraging the semantic feature distribution of SR images to match that of HR images. Specifically, as shown in Figure 4, we can obtain the middle features $F_{sm1}, F_{sm2}, F_{sm3}$ of SR image I_{sr}, and $F_{hm1}, F_{hm2}, F_{hm3}$ of HR image I_{hr} through the freezing CLIP image encoder E_i. Then, the Feat-D is used to perform discrimination on those pixel-wise semantic features. During the adversarial training process, the Feat-D and the SRN will be optimized respectively using the following adversarial loss functions:

$$\begin{aligned} L_{D_{fea}} = &\mathbb{E}_{F_{hr}\sim P_{F_{hr}}} log(1 - D(F_{hm1}, F_{hm2}, F_{hm3})) \\ &+ \mathbb{E}_{F_{sr}\sim P_{F_{sr}}} log(D(F_{sm1}, F_{sm2}, F_{sm3})), \end{aligned} \tag{1}$$

$$\begin{aligned} L_{G_{fea}} = &\mathbb{E}_{F_{hr}\sim P_{F_{hr}}} log(D(F_{hm1}, F_{hm2}, F_{hm3})) \\ &+ \mathbb{E}_{F_{sr}\sim P_{F_{sr}}} log(1 - D(F_{sm1}, F_{sm2}, F_{sm3})), \end{aligned} \tag{2}$$

where $P_{F_{hr}}$ and $P_{F_{sr}}$ represent the feature distributions of HR and SR images, respectively. Benefiting from the excellent discriminative ability of Feat-D, the SRN can learn more fine-grained semantic feature distributions aligned with HR images, generating more semantically relevant textures.

Learnable Prompt Pairs Driven Text-Guided Discrimination. Discriminating the final output feature F_{sout} which is more global and abstract is expected to further enhance the overall performance of our method. However, differing from the middle features, F_{sout} lacks the pixel-wise semantic information due to its more abstract nature, making it less suitable for Feat-D's pixel-wise discrimination process. Therefore, leveraging the inherent visual-language connection of CLIP, we propose a text-guided discrimination method, TG-D, to perform discrimination on F_{sout}. Specifically, inspired by [20, 61], we introduce learnable prompt pairs (LPP) into our method. As shown in Fig. 4, given the positive prompt and negative prompt, we can obtain the text features F_{t+} and F_{t-} through the CLIP text encoder E_t. Additionally, the final output features of SR image F_{sout} and HR image F_{hout} can be got from E_i. Then, we adopt the same computation method as in [61] to calculate the relative cosine similarity score between F_{sout} and the text features. This takes two steps: firstly, the cosine similarities s_+ between F_{sout} and F_{t+}, s_- between F_{sout} and f_{t-} can be respectively calculated as following:

$$s_+ = \frac{F_{sout} \odot F_{t+}}{||F_{sout}|| \cdot ||F_{t+}||}, \; s_- = \frac{F_{sout} \odot F_{t-}}{||F_{sout}|| \cdot ||F_{t-}||}, \tag{3}$$

where $\odot$ denotes vector dot-product and $||\cdot||$ represents L_2 norm. Then, Softmax is used to calculate the final score:

$$s_{sr} = \frac{e^{s_+}}{e^{s_+} + e^{s_-}}. \tag{4}$$

As for F_{hout} of HR image, we can also adopt Eq. (3) and Eq. (4) to calculate the cosine similarity score s_{hr}. Then, we

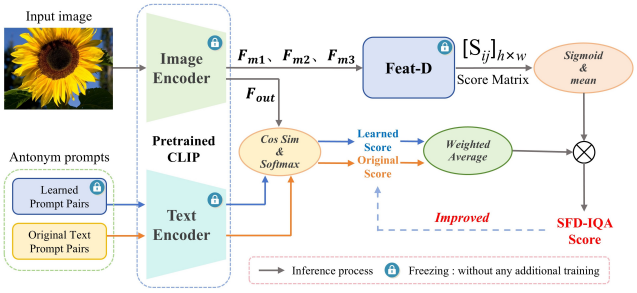


Figure 5. Framework of the proposed SFD-IQA. SFD-IQA is directly based on the well trained Feat-D and LPP, and it doesn't require any additional task-specific training on IQA datasets.

utilize s_{sr} and s_{hr} to train the LPP and SRN in an adversarial manner:

$$L_{D_{tg}} = \mathbb{E}_{F_{hr}\sim P_{F_{hr}}} log(1 - s_{hr}) + \mathbb{E}_{F_{sr}\sim P_{F_{sr}}} log(s_{sr}), \tag{5}$$

$$L_{G_{tg}} = \mathbb{E}_{F_{hr}\sim P_{F_{hr}}} log(s_{hr}) + \mathbb{E}_{F_{sr}\sim P_{F_{sr}}} log(1 - s_{sr}), \tag{6}$$

where $L_{D_{tg}}$ and $L_{G_{tg}}$ represent the adversarial loss functions for LPP and SRN, respectively. During the adversarial training process, LPP will be more and more sensitive to image quality and TG-D will be able to discriminate F_{sout} in the form of assigning a global score. Finally, the SRN is optimized by the combination of the following losses as:

$$\begin{aligned} L_{total} = \lambda_1 ||I_{hr}, I_{sr}||_1 + \lambda_2 ||\phi_i(I_{hr}), \phi_i(I_{sr})||_1 \\ + \lambda_3 L_{G_{fea}} + \lambda_4 L_{G_{tg}}, \end{aligned} \tag{7}$$

where $||\cdot||_1$ represents the L_1 norm, $\phi_i(\cdot)$ is the layers of pre-trained VGG-19 network[54], λ_1, λ_2, λ_3, λ_4 represent the hyperparameters for the above losses.

3.3. Extension to Opinion-Unaware NR-IQA

During the process where the Feat-D and TG-D learns to distinguish the distributions of semantic features in an adversarial manner, they simultaneously and implicitly learn to assess image quality. By integrating the well trained Feat-D and LPP, we design a novel OU NR-IQA method, SFD-IQA. As shown in Fig. 5, given an image x, we can obtain its middle semantic features F_{m1}, F_{m2}, F_{m3}, and the final output feature F_{out} through the CLIP image encoder E_i. Then, Feat-D will convert F_{m1}, F_{m2}, F_{m3} into a score matrix:

$$[s_{ij}]_{h\times w} = D(F_{m1}, F_{m2}, F_{m3}), \tag{8}$$

where $[s_{ij}]_{h\times w}$ denotes the score matrix. Additionally, according to Eq. 2 and Eq. 3, we adopt F_{out} and LPP to calculate the similarity score s_{lp}. Due to CLIP's outstanding contrastive learning training scheme, F_{out} is already sensitive to quality-related text prompts. To further enhance the robustness of SFD-IQA, we also adopt the original antonymic text prompts used in CLIP-IQA (e.g., "Good photo" and "Bad photo") to calculate the similarity score s_o with F_{out}. Then we perform a weighted average on s_o and s_{lp}:

$$s_{aver} = \alpha_1 s_o + \alpha_2 s_{lp}, \tag{9}$$

where α_1 and α_2 denotes the weight coefficients. Finally, following [83], we adopt the product of s_{aver} and the average of $[s_{ij}]_{h\times w}$ as the final IQA score s_d:

$$s_d = (s_{aver}) \times \frac{\sum_{i=0}^{h}\sum_{j=0}^{w}(sigmoid([s_{ij}]_{h\times w}))}{h \times w}, \tag{10}$$

where $sigmoid$ denotes the Sigmoid function to adjust the score range to [0, 1]. Notably, we directly apply the Feat-D and TG-D to the above IQA process without any additional task-specific training. Benefiting from the semantic-related discriminative abilities of Feat-D and TG-D, SFD-IQA can significantly improve the OU NR-IQA performance.

4. Experiments

4.1. Experimental Setup

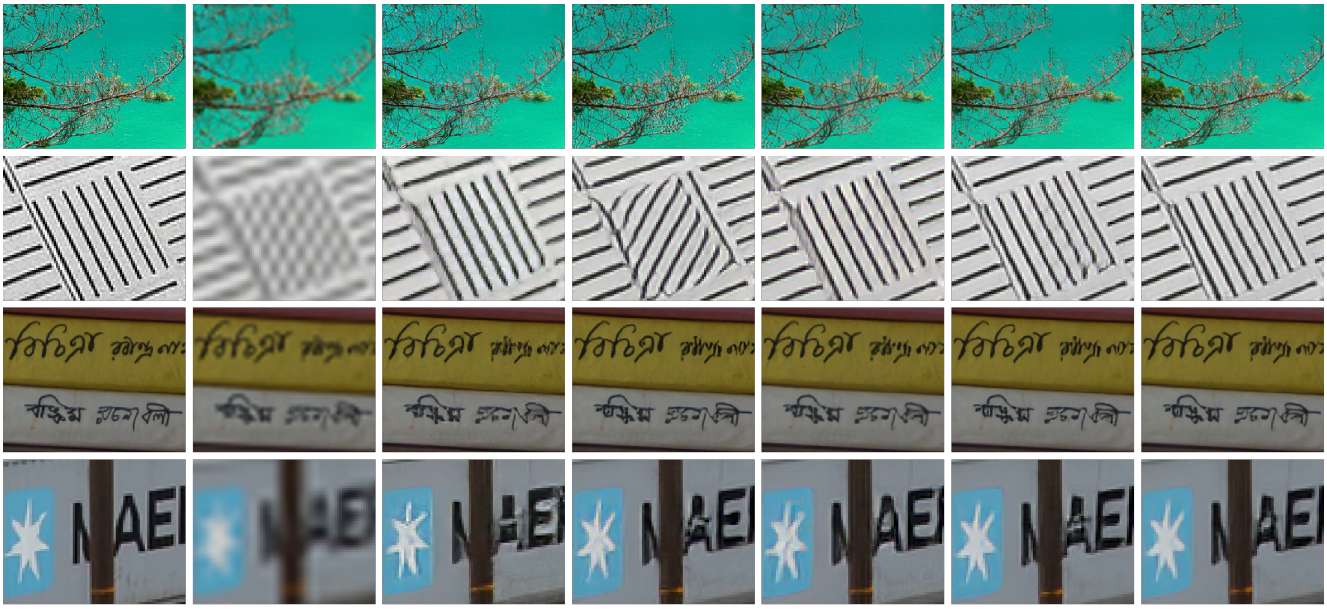
Training Datasets. For classical SISR, following previous works [15, 30, 39], we train our network using 3450 images from DF2K [1, 58], and the bicubic downsampling is used to synthesize LR-HR training pairs. For real-world SISR and OU NR-IQA, following [39, 65, 83], we adopt DF2K and OutdoorSceneTraining [63] as the training dataset, and the LR-HR training pairs are generated by the degradation pipeline of Real-ESRGAN [65].

Test Datasets and Evaluation Metrics. For classical SISR, we test on 4 commonly used benchmarks: Set5 [3], Set14 [74], Urban100 [22], and DIV2K validation set [1], using 4 metrics to comprehensively evaluate the performance of different methods, including fidelity metrics: PSNR and SSIM [66] (calculated on the Y channel in YCbCr space), and perceptual quality metrics: LPIPS [78], DISTS [10]. For real-world SISR, following [62, 70], we employ 2 center-cropped real-world datasets, RealSR [5] and DRealSR [68]. We also create another dataset, RealD2K, by degrading DIV2K validation set under the same pipeline as that in training. We add another metric NIQE [77] for real-world SISR. For OU NR-IQA, following [39, 83], we use two categories of IQA datasets, includinng SR IQA datasets: PIPAL [24], Ma's dataset [41], and authentically distorted IQA datasets: KonIQ-10k [21], LIVE-itW [16]. To evaluate the IQA performance, we adopt three widely used metrics: PLCC, SRCC and KRCC [25].

Implementation Details. For the generator, following [30, 65], we use the competitive RRDB [64] as the backbone. We implement the experiments on NVIDIA 4090 GPUs with PyTorch [49]. We initialize the parameters with the pre-trained fidelity-oriented model. We use Adam [26] optimizer to train the network with a learning rate of 10^{-4}.

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	RankSR GAN [79]	SRGAN [29]	SFTGAN [63]	ESRGAN [64]	SPSR [42]	LDL [32]	DualFormer [39]	WGSR [27]	SeD [30]	RRDB +Ours	SwinIR $+L_{GAN}$	SwinIR +Ours
Set5	PSNR ↑	29.65	30.52	30.06	30.44	30.40	31.03	31.30	31.33	31.22	31.63	31.10	31.78
	SSIM ↑	0.838	0.863	0.848	0.852	0.844	0.861	0.869	0.866	0.868	0.877	0.869	0.877
	LPIPS ↓	0.070	0.069	0.080	0.074	0.069	0.066	0.066	0.065	0.064	0.063	0.068	0.059
	DISTS ↓	0.104	0.108	0.109	0.097	0.093	0.093	0.093	0.110	0.094	0.098	0.098	0.089
Set14	PSNR ↑	26.45	27.01	26.74	26.28	26.64	27.10	27.39	27.40	27.37	27.79	27.05	27.94
	SSIM ↑	0.703	0.728	0.718	0.699	0.714	0.735	0.739	0.739	0.736	0.752	0.731	0.754
	LPIPS ↓	0.135	0.123	0.131	0.131	0.135	0.131	0.120	0.138	0.117	0.117	0.120	0.118
	DISTS ↓	0.110	0.109	0.113	0.099	0.099	0.098	0.092	0.112	0.093	0.096	0.098	0.091
Urban100	PSNR ↑	24.47	25.02	24.34	24.35	24.80	25.49	25.69	25.78	25.93	26.18	25.83	26.57
	SSIM ↑	0.729	0.778	0.724	0.734	0.747	0.767	0.773	0.781	0.780	0.787	0.785	0.798
	LPIPS ↓	0.138	0.130	0.134	0.123	0.121	0.110	0.115	0.135	0.107	0.109	0.100	0.099
	DISTS ↓	0.104	0.110	0.106	0.088	0.086	0.082	0.085	0.108	0.079	0.082	0.081	0.079
DIV2K100	PSNR ↑	28.03	28.18	28.09	28.20	28.18	28.96	29.25	29.19	29.27	29.75	28.99	29.90
	SSIM ↑	0.767	0.788	0.771	0.777	0.772	0.795	0.802	0.804	0.803	0.815	0.803	0.818
	LPIPS ↓	0.121	0.122	0.133	0.115	0.113	0.101	0.097	0.104	0.094	0.097	0.094	0.099
	DISTS ↓	0.064	0.063	0.074	0.059	0.055	0.053	0.056	0.054	0.054	0.053	0.050	0.050



GT Bicubic ESRGAN LDL DualFormer SeD-P SFD (Ours)

Figure 6. Visual comparison of different GAN-based SR methods on ×4 super-resolution.

4.2. Results on Classical SISR

To validate the effectiveness of our method, we compare it with several SOTA GAN-based SR methods, including [27, 29, 30, 32, 39, 42, 64, 79]. As shown in Tab. 1, compared to other approaches, our method achieves the best PSNR and SSIM scores on all benchmarks, while also holding highly competitive perceptual metrics, meaning that our method can sacrifice less fidelity in exchange for improved perceptual quality. To validate generalizability, we also employ another Transformer-based SR backbone, SwinIR [31]. It's clearly that our method also demonstrates superior performance with SwinIR. Fig. 6 presents the visual comparison, indicating that our method generates more realistic semantic textures with fewer artifacts than other methods.

4.3. Results on Real-World SISR

For real-world SISR, we provide quantitative and visual comparison with representative SOTA real-world SISR models including [31, 32, 65, 75]. The quantitative comparison is demonstrated in Table 2. The results show that our SFD achieves the best SSIM score on all datasets, while also maintaining competitive perceptual metrics, especially LPIPS and DISTS, indicating that our approach can achieve better PD trade-off. As shown in Fig. 7, our method produces SR



Figure 7. Visual comparison of different GAN-based real-world SR methods on ×4 super-resolution.

Table 2. Quantitative comparison of our method vs. other state-of-the-art real-world SISR methods for ×4 SR task. The best and the second-best are marked in red and blue, respectively.

Benchmark	Metrics	BSRGAN [75]	RealESRGAN [65]	LDL [32]	RealSwinIR [31]	SFD (Ours)
RealD2K	PSNR ↑	23.51	23.50	22.94	22.67	23.40
	SSIM ↑	0.585	0.604	0.597	0.581	0.606
	LPIPS ↓	0.379	0.368	0.373	0.368	0.358
	DISTS ↓	0.280	0.278	0.284	0.279	0.278
	NIQE ↓	7.537	8.369	8.384	7.866	8.600
RealSR	PSNR ↑	26.39	25.69	25.28	25.90	26.12
	SSIM ↑	0.765	0.762	0.757	0.768	0.772
	LPIPS ↓	0.267	0.273	0.275	0.262	0.256
	DISTS ↓	0.212	0.206	0.212	0.200	0.200
	NIQE ↓	5.657	5.830	5.992	5.926	5.797
DRealSR	PSNR ↑	28.75	28.64	28.21	28.21	29.20
	SSIM ↑	0.803	0.805	0.813	0.798	0.815
	LPIPS ↓	0.288	0.285	0.279	0.285	0.284
	DISTS ↓	0.214	0.209	0.213	0.210	0.221
	NIQE ↓	6.519	6.693	7.143	6.766	6.878

images with sharp edges and fine textures, accompanied by fewer undesirable artifacts. The above results suggest that our SFD can distinguish complex degraded real world images and facilitate the SR network to learn more fine-grained and semantically accurate textures.

4.4. Results on OU NR-IQA

To verify the effectiveness of our SFD-IQA for OU NR-IQA tasks, we compared SFD-IQA with several representative OU NR-IQA methods, including traditional methods [37, 69, 71, 76, 77], deep learning-based methods [2, 35, 39, 43, 46, 61, 83]. The results are presented in Table 3. Thanks to the excellent ability of Feat-D and TG-D to distinguish between low-quality and high-quality images, learned during the real-world SISR training process, our SFD-IQA achieves the best performance across all four datasets, validating its enhanced ability to predict the perceptual quality of images.

4.5. Ablation Studies

The effects of Feat-D and TG-D on perceptual SISR. To investigate the effectiveness of our Feat-D and TG-D, we conduct ablation experiments on them using the DIV2K100 test set. As shown in Fig. 8, compared to the baseline (ESRGAN), Feat-D and TG-D can both independently improve the perceptual quality with less decrease in fidelity. Furthermore, when both Feat-D and TG-D are introduced into the

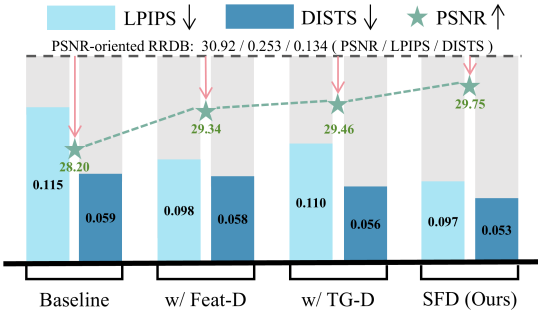


Figure 8. The effects of Feat-D and TG-D on perceptual SISR.

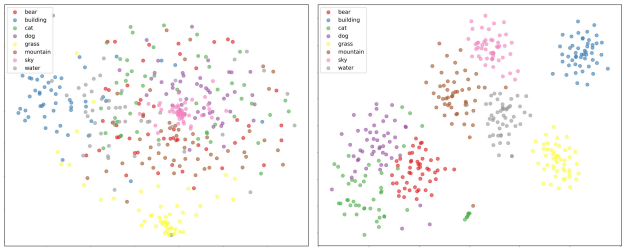


Figure 9. The t-SNE visualization of vanilla discriminator features and our Feat-D features. We select 8 categories of images from OST dataset, including "bear", "building", "cat", "dog", "grass", "mountain", "sky" and "water" and 50 images for each category.

baseline, we can trade the minimum PSNR decrease for the best perceptual quality. The above ablation results strongly verify the superiority of Feat-D and TG-D.

The effects of Feat-D and TG-D in SFD-IQA. To investigate the contribution of our Feat-D and TG-D to the proposed SFD-IQA, we conduct ablation studies by respectively remove Feat-D and TG-D from our SFD-IQA. Tab. 4 presents the comparative results on the 3 SR related sub-types of PIPAL [24] training set and 3 sub-types of KonIQ-10k [21] dataset. Compared to the baseline without Feat-D and TG-D, it's clearly that each of Feat-D and TG-D can independently enhance the OU NR-IQA performance to a certain extent. With both Feat-D and TG-D, SFD-IQA significantly improves the performance in most cases. Especially on the GAN-based SR subset of PIPAL, it boosts the PLCC by 86.0% and SRCC by 28.8%. The above results demonstrate the effectiveness of Feat-D and TG-D.

T-SNE analysis for Feat-D. To further demonstrate the

Table 3. OU NR-IQA comparison between different methods. The best and the second-best are marked in red and blue, respectively.

IQA Methods	PIPAL [24]			KonIQ-10k [21]			Ma's dataset [41]			LIVE-itW [16]		
	PLCC ↑	SRCC ↑	KRCC ↑	PLCC ↑	SRCC ↑	KRCC ↑	PLCC ↑	SRCC ↑	KRCC ↑	PLCC ↑	SRCC ↑	KRCC ↑
QAC [71]	0.2566	0.2220	0.1503	0.3719	0.3397	0.2302	0.5071	0.4387	0.3055	0.2720	0.0457	0.0370
LPSI [69]	0.2897	0.2031	0.1363	0.2066	0.2239	0.1504	0.5293	0.4896	0.3543	0.2877	0.0834	0.0524
NIQE [77]	0.0439	0.0608	0.0401	0.4637	0.5291	0.3666	0.6427	0.6261	0.4520	0.4790	0.4493	0.3062
IL-NIQE [76]	0.2748	0.2412	0.1630	0.5316	0.5057	0.3504	0.7155	0.6548	0.4751	0.5039	0.4394	0.2985
SNP-NIQE [37]	0.1820	0.1801	0.1214	0.6340	0.6285	0.4435	0.6250	0.5621	0.3984	0.5201	0.4657	0.3162
dipIQ [43]	0.2800	0.2304	0.1535	0.4429	0.2375	0.1594	0.5167	0.4866	0.3607	0.3180	0.1774	0.1207
RankIQA [35]	0.3698	0.3715	0.2522	0.5028	0.4983	0.3448	0.5178	0.4913	0.3648	0.4528	0.4307	0.2945
RecycleD [83]	0.3396	0.3331	0.2265	0.6407	0.6364	0.4468	0.5983	0.5555	0.3874	0.4806	0.4661	0.3194
DualFormer [39]	0.2896	0.2787	0.1973	0.6543	0.6321	0.4435	0.6172	0.6054	0.4044	0.5068	0.4897	0.3312
ContentSep [2]	-	-	-	0.6274	0.6401	0.4529	-	-	-	0.5130	0.5060	0.3450
CLIPIQA [61]	0.2831	0.2762	0.1862	0.4330	0.4145	0.2821	0.7175	0.7055	0.5063	0.5853	0.5647	0.4010
MDFS [46]	0.2651	0.2273	0.1532	0.7123	0.7333	0.5344	0.7649	0.7142	0.5205	0.5364	0.4821	0.3274
SFD-IQA (Ours)	0.4252	0.4092	0.2816	0.7574	0.7440	0.5443	0.8012	0.7870	0.5864	0.7111	0.6984	0.5090

Table 4. OU NR-IQA performance comparison on sub-types of PIPAL training set and KonIQ-10k dataset. "Baseline" refers to the model without neither Feat-D nor TG-D. "W/ Feat-D" and "w/. TG-D" refer to the models that only include Feat-D or TG-D.

Metrics	Dataset. Sub-Types	Baseline	w/. Feat-D	w/. TG-D	SFD-IQA (Ours)
PLCC	PIPAL. Trad.SR	0.3787	0.3684	**0.3910**	0.3797 (+0.26%)
	PIPAL. PSNR.SR	0.4290	0.4504	0.4547	**0.4657** (+8.55%)
	PIPAL. GAN.SR	0.2154	0.3958	0.2430	**0.4006** (+86.0%)
	KonIQ-10k. Train	0.7092	0.7484	0.7408	**0.7581** (+6.90%)
	KonIQ-10k. Val	0.7055	0.7452	0.7350	**0.7528** (+6.70%)
	KonIQ-10k. Test	0.6865	0.7493	0.7465	**0.7576** (+10.4%)
SRCC	PIPAL. Trad.SR	0.3545	0.3420	**0.3660**	0.3522 (-0.64%)
	PIPAL. PSNR.SR	0.4432	0.4797	0.4672	**0.4932** (+11.3%)
	PIPAL. GAN.SR	0.2376	0.3048	0.2598	**0.3061** (+28.8%)
	KonIQ-10k. Train	0.6849	0.7325	0.7130	**0.7441** (+4.92%)
	KonIQ-10k. Val	0.6803	0.7264	0.7072	**0.7373** (+4.51%)
	KonIQ-10k. Test	0.6862	0.7359	0.7131	**0.7475** (+8.89%)

effectiveness of our Feat-D, we use t-SNE [59] to visualize the middle features of the VGG-style vanilla discriminator and our Feat-D. We select 8 categories of images from the OST [63] dataset, and 50 images are randomly selected for each category. We visualize features after the 3rd upsampling layer of our Feat-D and after the 2nd BN layer of the vanilla discriminator, respectively. As shown in Fig. 9, the feature clustering of our Feat-D is much better than that of the vanilla discriminator. This strongly proves that the proposed Feat-D is semantic-aware and can perform more fine-grained discrimination, encouraging the SR network to generate more more realistic semantic textures.

The effectiveness of our method on other image restoration tasks. We preliminarily explore the effectiveness of our approach in addressing other image restoration tasks, such as image deraining and image deblurring. We select a commonly used image restoration baselines NAFNet [8], and conduct training and testing on the Rain100L [73] and GoPro [45] datasets for the two tasks. As shown in Tab. 5, compared to fidelity-oriented baseline methods and vanilla discriminator based methods, our method significantly improves perceptual metrics while maintaining a high level of fidelity on both two tasks, demonstrating the potential of our method to be integrated into other image restoration tasks for better perceptual quality.

Table 5. The effectiveness of our method on image deraining and image deblurring. "VD" means vanilla discriminator.

Method	Rain100L [73]			GoPro [45]		
	PSNR ↑	LPIPS ↓	DISTS ↓	PSNR ↑	LPIPS ↓	DISTS ↓
NAFNet	**37.00**	0.022	0.031	**32.87**	0.093	0.072
+ VD	36.39	0.015	0.020	32.01	0.068	0.036
+ Ours	36.49	**0.014**	**0.019**	32.19	**0.064**	**0.034**

5. Conclusion

In this paper, we propose a semantic feature discrimination method, SFD, for perceptual SISR and OU NR-IQA. We first design a feature discriminator (Feat-D) to discriminate the semantic features from CLIP, effectively encouraging the feature distributions of SR images to match that of HR images. Then we propose a text-guided discrimination (TG-D) method by training the learnable prompt pairs (LPP) in an adversarial manner to further enhance the discriminative ability. Our SFD facilitates the SR network to generate more fine-grained and more realistic semantic textures, achieving better perception-distortion trade-off. Furthermore, by effectively integrating the well trained Feat-D and LPP, we significantly improve the OU NR-IQA performance without any additional task-specific training on IQA datasets. Extensive experiments on classical SISR, real-world SISR, and OU NR-IQA tasks validate the effectiveness of our approach.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grant 62171304 and partly by the Natural Science Foundation of Sichuan Province under Grant 2024NSFSC1423, Cooperation Science and Technology Project of Sichuan University and Dazhou City under Grant 2022CDDZ-09, the TCL Science and Technology Innovation Fund under grant 25JZH008, and the Young Faculty Technology Innovation Capacity Enhancement Program of Sichuan University under Grant 2024SCUQJTX025.

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. 5

[2] Nithin C Babu, Vignesh Kannan, and Rajiv Soundararajan. No reference opinion unaware quality assessment of authentically distorted images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2459–2468, 2023. 3, 7, 8

[3] 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. 5

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 2

[5] 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

[6] Du Chen, Zhengqiang Zhang, Jie Liang, and Lei Zhang. Ssl: A self-similarity loss for improving generative image super-resolution. In *ACM Multimedia 2024*, 2024. 2

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 2

[8] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 8

[9] 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. 2

[10] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 5

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 1, 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 3

[13] Jinsheng Fang, Hanjiang Lin, Xinyu Chen, and Kun Zeng. A hybrid network of cnn and transformer for lightweight image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1103–1112, 2022. 2, 3

[14] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6):2944–2956, 2017. 1

[15] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2360–2369, 2021. 2, 3, 5

[16] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 5, 8

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[18] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018. 1, 3

[19] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 1

[20] 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. 2, 3, 4

[21] 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. 5, 7, 8

[22] 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

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[24] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 633–651. Springer, 2020. 5, 7, 8

[25] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. 5

[26] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[27] Cansu Korkmaz, A Murat Tekalp, and Zafer Dogan. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5926–5936, 2024. 3, 6

[28] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 1, 2, 3

[29] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2, 3, 6

[30] 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. 2, 3, 4, 5, 6

[31] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2, 3, 6, 7

[32] 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. 2, 3, 6, 7

[33] Xiangyu Liao, Tianheng Zheng, Jiayu Zhong, Pingping Zhang, and Chao Ren. Asymmetric mask scheme for self-supervised real image denoising. *arXiv preprint arXiv:2407.06514*, 2024. 1

[34] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 2, 3

[35] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pages 1040–1049, 2017. 3, 7, 8

[36] Xiao Liu, Xiangyu Liao, Xiuya Shi, Linbo Qing, and Chao Ren. Efficient information modulation network for image super-resolution. In *ECAI 2023*, pages 1544–1551. IOS Press, 2023. 1

[37] Yutao Liu, Ke Gu, Yongbing Zhang, Xiu Li, Guangtao Zhai, Debin Zhao, and Wen Gao. Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):929–943, 2019. 7, 8

[38] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022. 2

[39] 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. 2, 3, 5, 6, 7, 8

[40] Chenxi Ma. Uncertainty-aware gan for single image super resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4071–4079, 2024. 2, 3

[41] 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. 5, 8

[42] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7769–7778, 2020. 6

[43] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017. 3, 7, 8

[44] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 2, 3

[45] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 8

[46] Zhangkai Ni, Yue Liu, Keyan Ding, Wenhan Yang, Hanli Wang, and Shiqi Wang. Opinion-unaware blind image quality assessment using multi-scale deep feature statistics. *IEEE Transactions on Multimedia*, 2024. 3, 7, 8

[47] JoonKyu Park, Sanghyun Son, and Kyoung Mu Lee. Content-aware local gan for photo-realistic super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10585–10594, 2023. 2, 3

[48] Seung Ho Park, Young Su Moon, and Nam Ik Cho. Perception-oriented single image super-resolution using optimal objective estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1725–1735, 2023. 2

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[50] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36, 2023. 1

[51] 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. 2, 3

[52] Chao Ren, Yizhong Pan, and Jie Huang. Enhanced latent space blind model for real image denoising via alternative optimization. *Advances in Neural Information Processing Systems*, 35:38386–38399, 2022. 1

[53] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8207–8216, 2020. 2

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 5

[55] Lingchen Sun, Jie Liang, Shuaizheng Liu, Hongwei Yong, and Lei Zhang. Perception-distortion balanced super-resolution: A multi-objective optimization perspective. *IEEE Transactions on Image Processing*, 33:4444–4458, 2024. 3

[56] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017. 1, 2, 3

[57] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017. 1, 2, 3

[58] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5

[59] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[60] Cong Wang, Jinshan Pan, Wei Wang, Jiangxin Dong, Mengzhu Wang, Yakun Ju, and Junyang Chen. Promptrestorer: A prompting image restoration method with degradation perception. *Advances in Neural Information Processing Systems*, 36:8898–8912, 2023. 1

[61] 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. 2, 3, 4, 7, 8

[62] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 1, 3, 5

[63] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 2, 5, 6, 8

[64] 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, 5, 6

[65] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 2, 3, 5, 6, 7

[66] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[67] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 1

[68] 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. 5

[69] Qingbo Wu, Zhou Wang, and Hongliang Li. A highly efficient method for blind image quality assessment. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 339–343. IEEE, 2015. 3, 7, 8

[70] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25456–25467, 2024. 3, 5

[71] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 995–1002, 2013. 3, 7, 8

[72] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 2, 3

[73] Wenhan Yang, Robby T Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1377–1393, 2019. 8

[74] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 5

[75] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 3, 6, 7

[76] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 3, 7, 8

[77] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 3, 5, 7, 8

[78] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep

features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[79] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Super resolution generative adversarial networks with learning to rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7149–7166, 2021. 3, 6

[80] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European conference on computer vision*, pages 649–667. Springer, 2022. 1, 2, 3

[81] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2, 3

[82] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2, 3

[83] Yunan Zhu, Haichuan Ma, Jialun Peng, Dong Liu, and Zhiwei Xiong. Recycling discriminator: Towards opinion-unaware image quality assessment using wasserstein gan. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 116–125, 2021. 2, 3, 5, 7, 8