SeD: Semantic-Aware Discriminator for Image Super-Resolution

Bingchen Li[1*], Xin Li[1*], Hanxin Zhu[1], Yeying Jin[2], Ruoyu Feng[1], Zhizheng Zhang[3], Zhibo Chen[1†]

[1]University of Science and Technology of China

[2]National University of Singapore   [3]Galbot

{lbc31415926, lixin666, hanxinzhu}@mail.ustc.edu.cn, jinyeying@u.nus.edu,

ustcfry@mail.ustc.edu.cn, zhangzz@galbot.com, chenzhibo@ustc.edu.cn

Abstract

Generative Adversarial Networks (GANs) have been widely used to recover vivid textures in image super-resolution (SR) tasks. In particular, one discriminator is utilized to enable the SR network to learn the distribution of real-world high-quality images in an adversarial training manner. However, the distribution learning is overly coarse-grained, which is susceptible to virtual textures and causes counter-intuitive generation results. To mitigate this, we propose the simple and effective Semantic-aware Discriminator (denoted as SeD), which encourages the SR network to learn the fine-grained distributions by introducing the semantics of images as a condition. Concretely, we aim to excavate the semantics of images from a well-trained semantic extractor. Under different semantics, the discriminator is able to distinguish the real-fake images individually and adaptively, which guides the SR network to learn the more fine-grained semantic-aware textures. To obtain accurate and abundant semantics, we take full advantage of recently popular pretrained vision models (PVMs) with extensive datasets, and then incorporate its semantic features into the discriminator through a well-designed spatial cross-attention module. In this way, our proposed semantic-aware discriminator empowered the SR network to produce more photo-realistic and pleasing images. Extensive experiments on two typical tasks, i.e.*, SR and Real SR have demonstrated the effectiveness of our proposed methods. The code will be available at* https://github.com/lbc12345/SeD.

1. Introduction

Deep learning has accelerated the great development of Single Image Super-resolution (SISR) [12, 34, 36, 50, 59, 60], which aims to recover the vivid high-resolution (HR) im-
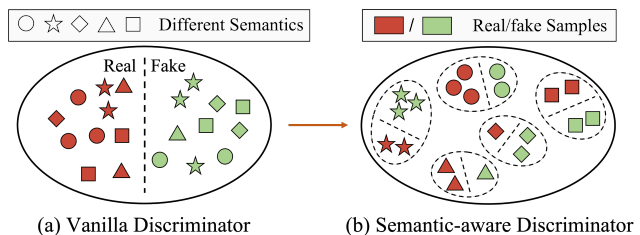

Figure 1. Comparison between the vanilla discriminator and our proposed semantic-aware discriminator.

age from its degraded low-resolution (LR) counterpart. In general, there are two typical optimization objectives for existing SISR methods, i.e., objective and subjective quality. To pursue the higher objective quality, a series of elaborate frameworks based on convolutional neural networks (CNN) [9, 31, 36, 72, 73], Transformer [5, 15, 34, 39, 63] have been proposed to improve the representation ability of the SR network through the constrain of the pixel-wise loss function (e.g., L1 loss and MSE loss). However, the pixel-wise loss function cannot enable the SR network to have a promising hallucination capability, which performs poorly on texture generation and results in an unsatisfied subjective quality [27, 32].

To improve the subjective quality, it is necessary to generate human-pleasing textures for distorted images. Inspired by the generative adversarial network (GAN) [17], a series of pioneering works [27, 59, 60, 70, 71] regard the SR network as the generator, and then introduce the discriminator to enable the SR networks with realistic texture generation capability. There are three typical discriminators for SR networks, i.e., image-wise discriminator [24, 54], patch-wise discriminator [22], and pixel-wise discriminator [52]. In particular, the image-wise discriminator aims to distinguish the real/fake images from the global distributions, e.g., VGG-like discriminator in [27, 59, 71]. However, the image-wise distribution is so coarse-grained that causes the network to produce non-ideal local textures. To

---

[*]Equal contribution

[†]Corresponding Author

enhance the local textures, [42, 61, 62] utilize the patch-wise discriminator to determine the patch-wise distribution and adjust the patch size with different sizes of receptive fields. Furthermore, the pixel-wise discriminator [52] distinguishes the real/fake distribution in a per-pixel manner, while bringing the large computational cost. However, the above works ignore the fact that the textures of one image should meet the distribution of its semantics. It is necessary to achieve the fine-grained semantic-ware texture generation for SR.

To generate the semantic-aware textures, one intuitive method is to integrate the semantics of images into the generator, which enables the generator the semantic adaptive capability. Actually, this brings two obvious drawbacks to semantic-aware texture generation: 1) The low-quality image may yield worse and even error semantic extraction, which prevents reasonable texture generation. 2) In the inference stage, the complex semantic extractor will cause the catastrophic growth of computation and model complexity for the SR network. In contrast, we aim to achieve the semantic-aware texture generation for SR from another perspective, *i.e.*, Discriminator.

In this paper, we propose the first Semantic-aware Discriminator for Image Super-resolution (SR), dubbed SeD, which is inspired by [74]. It is noteworthy that the discriminator works by distinguishing whether the image/patch/pixel is real or fake. With adversarial training, it is enabled to measure the distribution distance between the generated image and its reference image. As shown in Fig. 1, the vanilla discriminator only measures the distribution distance while ignoring the semantics, which is susceptible to the coarse-grained average textures (*e.g.* noise). To mitigate this, we introduce the semantics extracted from recent popular pretrained vision models (PVMs) (*e.g.* ResNet-50 or CLIP) as the condition of the discriminator, which lets the discriminator measure the distribution distance individually and adaptively for different semantics (See Fig. 1(b)). With the constraint of the semantic-aware discriminator, the SR networks (*i.e.*, generator) can achieve more fine-grained semantic-aware texture generation.

There is one crucial step to achieve a semantic-aware discriminator, *i.e.*, how to extract the semantics and incorporate them into a discriminator. A naïve strategy is to directly introduce the features of the last layer from a pretrained classification network as the semantics and concatenate them in a discriminator as a condition. However, it will prevent effective semantic guidance for discriminators, since it does not consider the characteristics of SR. To achieve fine-grained semantic-aware texture generation, one intuition is that the semantics is required to be pixel-wise due to semantics in different regions might be different. Considering this, we extract the semantics from the middle features of the PVMs. Depart from this, we also de-

sign one semantic-aware fusion block (SeFB) to better incorporate the semantics into the discriminator. Concretely, in SeFB, we regard the semantics extracted from PVMs as the query, to warp the semantic-aware image feature to the discriminator through cross-attention, which brings better semantic guidance. We validate the effectiveness of our proposed SeD on two typical SR tasks, *i.e.*, classical and Real-world image SR tasks. Our SeD is general and applicable to different benchmarks for GAN-based SR, *e.g.*, ESRGAN [59], RealESRGAN [60], and BSRGAN [70].

The contributions of this paper can be summarized as follows:

- We pinpoint the importance of fine-grained semantic-aware texture generation for SR, and for the first time propose a semantic-aware discriminator (SeD) for the SR task, by incorporating the semantics from pretrained vision models (PVMs) into the discriminator.
- To better incorporate the guidance of semantics for the discriminator, we propose the semantic-aware fusion block (SeFB) for SeD, which extracts the pixel-wise semantics and warped the semantic-aware image features into the discriminator through the cross-attention manner.
- Extensive experiments on two typical SR tasks, *i.e.*, classical and Real-world image SR have revealed the effectiveness of our proposed SeD. Moreover, our SeD can be easily integrated into many benchmarks for the GAN-based SR methods in a plug-and-play manner.

2. Related Works

2.1. Single Image Super-resolution

In recent years, deep learning has posed great progress in Single Image Super-resolution (SISR) [12, 73]. Typically, most SR methods focus on exploring better backbones [9, 28, 34, 36, 44, 64, 72] to improve objective quality.

As the pioneering work, [12] firstly propose the SR-CNN, which introduces the CNN to SISR and optimizes the network by minimizing the mean square error (MSE) between the super-resolved images (SR) and their corresponding high-resolution (HR) counterparts. Subsequently, numerous SR works have been explored to further improve the representation capability of the network from different perspectives, such as incorporating the residual connection [25, 27, 36], dense connection [59, 73], multi-scale representation [16, 29], etc. Specifically, some works [4, 9, 72] employ attention mechanisms to further enhance high-frequency textures. However, these works lack the utilization of long-range dependencies present in images, thereby limiting their performance on SR. With the emergence of vision transformers, some studies [34, 39, 63] regard pixels as tokens and make full use of long-range dependencies with window-based self-attention, resulting in visually pleasing outputs for both classical (*i.e.*, bicubic) and real-world im-
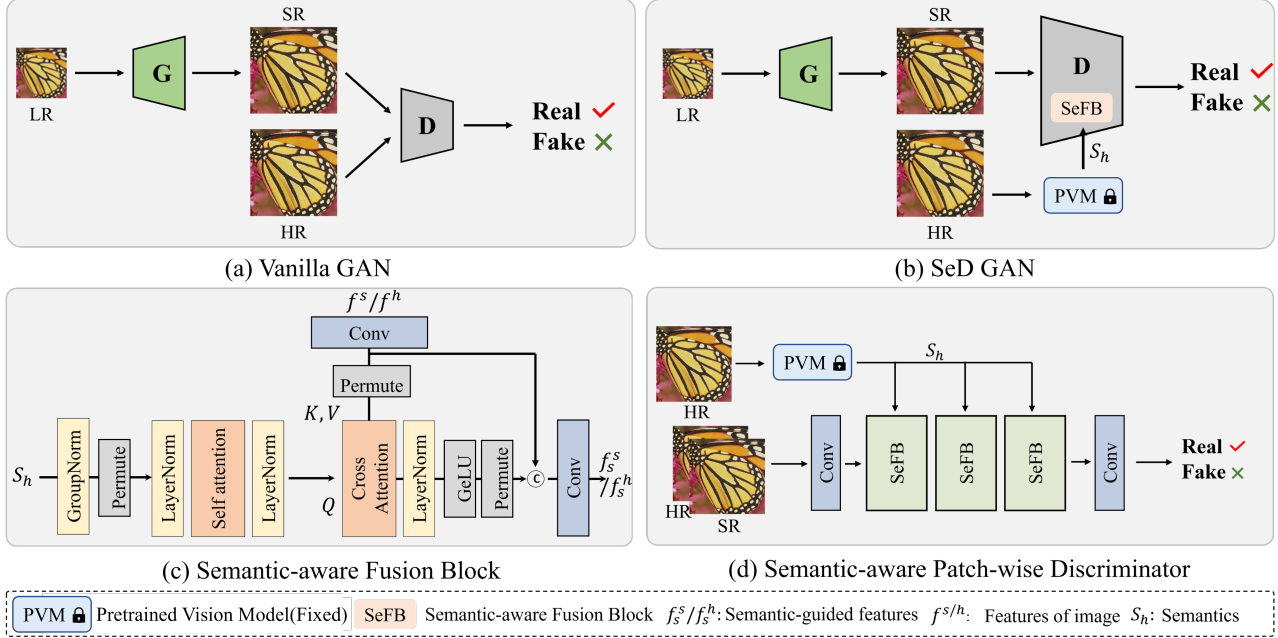
Figure 2. Illustration of (a) GAN-based SR with the vanilla discriminator. (b) Our proposed semantic-aware discriminator (SeD). (c) The network structure of SeFB. (d) The network structure of P+SeD. The vanilla discriminator measures the distributions of images regardless of the semantics, which causes the SR network to learn the average textures (*i.e.*, noise) or generate textures not related to the semantics. In contrast, our proposed semantic-aware discriminator exploits the fine-grained semantics as the condition of the discriminator, which poses the SR network to learn more fine-grained semantic-aware textures for SR.

age SR tasks.

2.2. GAN-based Image Super-resolution

To generate more appealing textures, some works aim to leverage the excellent hallucination capability of the generative models to enhance the subjective quality of SR Networks. In contrast to commonly-used pixel-wise constraints, *e.g.*, L1 and L2 losses, the adversarial loss is introduced to enable the SR network to learn the distribution of real-world high-quality images. SRGAN [27], as the first work for GAN-based SR, systematically investigates the advantages of GAN for SR. Following this, a series of works [26, 59, 66, 71] began to refine the architecture of the SR network (*i.e.*, the generator), achieving stable training and more realistic textures. Despite this, few works take into account optimizing another core component in GAN, *i.e.*, the discriminator, which determines whether the distribution learned by the SR network aligns with real-world high-quality images.

Starting from this point, our proposed Semantic-aware Discriminator (SeD) is expected to enable the SR network to achieve more fine-grained semantic-aware texture generation.

2.3. Pretrained Vision Models (PVMs)

Benefiting from large-scale datasets, such as ImageNet-22K [10], JFT-300M [55], LAION-5B [53], and the elaborate pretraining strategies [11, 20], Pretrained Vision Models (PVMs), *e.g.*, ViT [6, 13], Swin Transformer [37, 38], have demonstrated their immense potential in various fields, *e.g.*, classification [7, 65], vision-language task [3, 33, 46, 51], generation tasks [14, 47, 48]. Compared to smaller models [19], PVMs are friendly to most downstreaming tasks, thanks to their strong and robust representation capability. Recently, the large vision-language model CLIP [46] garnered significant interest for leveraging extensive amounts of collected vision-language pairs to learn a more fine-grained semantic space. As a result, a series of works incorporate this powerful feature extractor of CLIP into various tasks, including prompt learning [67], generation, etc. Inspired by this, we aim to extract more fine-grained semantics from CLIP to guide our SeD.

3. Method

In this section, we first review the typical GAN-based SR method and then describe the difference between our method with it in the overall framework. Next, we clarify our proposed semantic-aware discriminator in detail. Finally, we describe how to integrate our SeD into the existing

patch-wise discriminator and pixel-wise discriminator.

3.1. Preliminary

GAN-based SR [26, 27, 45, 59, 60, 70] aims to improve the perceptual quality of super-resolved images. With a simple discriminator, GAN-based SR introduces adversarial training to enable the SR network to generate vivid textures. The objectives of GAN-based SR are composed of three losses, including pixel-wise supervised loss $\mathcal{L}_s$, *e.g.*, L1 Loss and MSE loss, a perceptual loss $\mathcal{L}_p$, and an adversarial loss $\mathcal{L}_{adv}$. Among them, pixel-wise supervised loss is used to constrain the pixel-wise consistency and ensure the realness of pixel values. Perceptual loss [23, 27] $\mathcal{L}_p$ exploits the features of VGG [54] to give coarse perception constraints between generated and ground-truth images. In the adversarial training process, a discriminator D is exploited to distinguish whether the image is real or fake, which is optimized with the loss function as:

$$\mathcal{L}_D = \mathbb{E}_{I_h \sim P_{I_h}} log(1 - D(I_h)) + \mathbb{E}_{I_s \sim P_{I_s}}(D(I_s)), \quad (1)$$

where the P_{I_h} and P_{I_s} are the distribution of the high-quality images I_{I_h} and the super-resolved images I_s, respectively. The SR network (*i.e.*, the generator) is optimized by the combination of three losses as:

$$\mathcal{L}_G = \mathcal{L}_s + \lambda_p \mathcal{L}_p + \lambda_a \mathcal{L}_{adv}, \quad (2)$$

where adversarial loss $\mathcal{L}_{adv}$ enables the generator to cheat the discriminator D with the contrary purpose of $\mathcal{L}_D$ as:

$$\mathcal{L}_{adv} = \mathbb{E}_{I_h \sim P_{I_h}} log(D(I_h)) + \mathbb{E}_{I_s \sim P_{I_s}}(1 - D(I_s)). \quad (3)$$

3.2. Overall Framework

The overall framework of our proposed Semantic-aware Discriminator (SeD) is shown in Fig. 2. Given the low-resolution image I_l, we can first obtain the super-resolved image I_s. Then a discriminator D is used to distinguish the I_s and high-resolution image I_h, which enforces the SR network to generate the real-like images (*i.e.*, $P(I_s) \approx P(I_h)$). However, the vanilla discriminator only takes into account the coarse-grained distribution of images, while ignoring the semantics of images. This will cause the SR network to produce fake and even worse textures. A promising texture generation should satisfy its semantic information. Therefore, we aim to achieve the semantic-aware discriminator, which leverages the semantics of high-resolution image I_h as a condition. Here, we represent the semantic extractor in the large vision model as ϕ, and we aim to achieve more fine-grained semantic-aware texture generation, which targets for $P(I_s|\phi(I_h)) = P(I_h|\phi(I_h))$.

Therefore, as shown in the Fig. 2, the high-resolution image I_h will be fed into a fixed pretrained semantic extractor from large vision model ϕ to obtain the semantics

$\phi(I_h)$. Then a semantic-aware fusion block (SeFB) is used to further warp the super-resolved image feature f^s and high-resolution image feature f^h to the discriminator as f_s^s, f_s^h by regarding the semantics as the query. Based on the semantic-aware features, the discriminator can achieve the semantic-aware distribution measurement. Notably, this process does not increase the parameters or computational cost for the SR network in the inference stage, as discriminator is no longer needed. In the next section, we will clarify our semantic-ware discriminator in detail.

3.3. Semantic-aware Discriminator

Our semantic-aware discriminator consists of two essential components, *i.e.*, the semantic feature extractor and the semantic-aware fusion block (SeFB). By incorporating these components into the popular discriminators, we can obtain our Semantic-aware discriminators.

3.3.1 Semantic Excavation

A successful semantic excavation plays an irreplaceable role in the Semantic-aware Discriminator. The intuition is that more fine-grained semantics will promote the discriminator to measure finer-grained distributions. Therefore, the recently popular pretrained vision models (PVMs), which possess more powerful representation capabilities, are proper for semantic excavation. Among them, the CLIP model has emerged as a prominent backbone and has been applied in various tasks [8, 40, 47, 49]. Consequently, we adopt the pretrained CLIP "RN50" model as the semantic extractor, since we require the pixel-wise semantic information for SR. Specifically, the "RN50" is composed of four layers, with the resolution of features being down-sampled as the layers increase and the semantics becoming more abstract. To investigate which layer is more suitable for our semantic excavation, we systematically conduct experiments for these four layers, and experimentally find that the semantic features from the third layer are optimal. Here, we represent the semantic extractor as ϕ, and the semantic features f_s are extracted by feeding the high-resolution image I_h into ϕ.

3.3.2 Semantic-aware Fusion Block

After obtaining the semantics S_h, we exploit our proposed semantic-aware fusion block (SeFB) to implement the semantic guidance of Discriminator. The architecture of SeFB is shown in Fig. 2(c). Given the features of the super-resolved/high-quality images f^s/f^h, we aim to warp the semantic-aware textures from images to the discriminator, which enforce the discriminator to focus on the distribution of semantic-aware textures. Therefore, in Fig. 2(c), the semantics S_h is passed to the self-attention module and then

is fed to the cross-attention module as the query:

$$Q = \text{LN}(\text{SA}(\text{LN}(\text{GN}(S_h)))), \quad (4)$$

where LN, GN, SA denotes the layer normalization, group normalization, and self-attention module. Then, the features of super-resolved/high-quality images f^s/f^h are passed into the convolution layer and are regarded as the keys K^s/K^h and values Q^s/Q^h. Finally, the warped semantic-aware image features with cross-attention are concatenated with original enhanced features to form the final features f_s^s/f_s^h as:

$$f_s^{(*)'} = Softmax(Q^{(*)}K^{(*)T}/\sqrt{d_k})V^{(*)} \quad (5)$$
$$f_s^{(*)} = \text{Conv}(\text{Concat}(\text{GELU}(\text{LN}(f_s^{(*)'})), \text{Conv}(f^{(*)}))),$$

where $f_s^{(*)}$ denote the warped semantic-aware image features. $*$, d_k, and GELU are s/h, scale factor, and the Gaussian Error Linear Units, respectively. Conv and Concat represent the convolution layer and the concatenate operation.

3.4. Extension to Various Discriminators

In this paper, we incorporate our proposed SeD into two popular discriminators, including a patch-wise discriminator and a pixel-wise discriminator. As illustrated in Fig. 2(d), the patch-wise semantic-aware discriminator consists of three SeFBs and two convolution layers. For the pixel-wise discriminator, we follow the approach in [60] and exploit the U-Net architecture as the backbone. We substitute original convolution layers with our proposed SeFBs in the shallow feature extraction stage.

4. Experiments

4.1. Experiment setups

Network structures. For the generator, we use the competitive RRDB [59] as the backbone. To demonstrate the generalization ability of our proposed SeD, we also employ the popular transformer-based SR network SwinIR [34] as our generator. For the discriminator, we incorporate our SeD into two types of discriminators, *i.e.*, the patch-wise discriminator [22] and the pixel-wise discriminator [52], namely, P+SeD and U+SeD. We also validate our SeD on an image-wise discriminator based on VGG [59], dubbed V+SeD, which is provided in the **Supplementary**.

Implementation details for classical image SR. Following previous works [18, 34], we train our network using 3450 images from DF2K [1, 56]. We test our methods on five benchmarks: Set5 [2], Set14 [68], Urban100 [21], Manga109 [43] and DIV2K validation set [1]. We synthesize all the ×4 training samples using MATLAB bicubic downsampling kernel. We apply the same data augmentation to all the training samples following previous

arts [30, 34, 64]. The patch size of HR images is 256×256. We implement the experiments on 4 NVIDIA V100 GPUs with PyTorch and a batch size of 8 on each device. The parameters of the generator are initialized with pretrained PSNR-oriented model. We employ Adam optimizer with a fixed learning rate of 1e-4 to optimize our network for 300k iterations. We provide the implementation details for real-world image SR in the **Supplementary**.

4.2. Results on classical image SR

We conduct a quantitative comparison between our SeD and the state-of-the-art GAN-based SR methods. SFTGAN [58] incorporates a semantic map extracted from input LR images to better restore textures. ESRGAN [59] and USRGAN [69] both use RRDB [59] as the backbone of the generator. LDL [35] employs an artifact map into discriminator with RRDB to better reconstruct finer textures. Therefore, we exploit the RRDB as the generator for a fair comparison with them. To validate the generalization capability, we also validate our SeD with another generator architecture, *i.e.*, transformer-based SR backbone, SwinIR [34].

As demonstrated in Table 1, our SeD outperforms existing GAN-based methods in the perception metric (*i.e.*, LPIPS), including SFTGAN [58], ESRGAN [59], USRGAN [69], LDL [35] and DualFormer [41], with the comparable and even more high objective qualities. It is noteworthy that, for the last several rows of the Table 1, we validate the effectiveness of our SeD by incorporating our SeD into different generators (*i.e.*, RRDB [59] and SwinIR [34]) and discriminators (patch-wise discriminator, *i.e.*, "+P" and U-shape pixel-wise discriminator, *i.e.*, "+U"). For "RRDB+P", we can observe that our SeD (*i.e.*, "RRDB+P+SeD") outperforms the vanilla discriminator in LPIPS under all circumstances, even with up to 1.0dB PSNR gains in Manga109 dataset. This indicates the superiority of incorporating semantic guidance in our SeD, as it can discriminate more fine-grained textures during training, resulting in better perceptual quality. Notice that, our SeD also performs well on the transformer-based SR method SwinIR (*i.e.*, "SwinIR+P+SeD", and "SwinIR+U+SeD"). This demonstrates the generalization capacity of SeD on different generator architectures.

Fig. 3 presents the visual comparison between state-of-the-art methods and our SeD on classical image SR. We can notice that other methods restore images with virtual textures, *e.g.*, twisted rain drops in the first row; unnatural textures on the surface of the bridge in the second row; mistaking the dotted texture in front of the car for stripes in the last row, *etc*. Compared with other methods, our SeD generates visual-pleasant results without introducing artifacts. This proves that incorporating semantic guidance through SeD can empower the discriminator to distinguish finer details. We provide more visual results in the **Supplementary**.
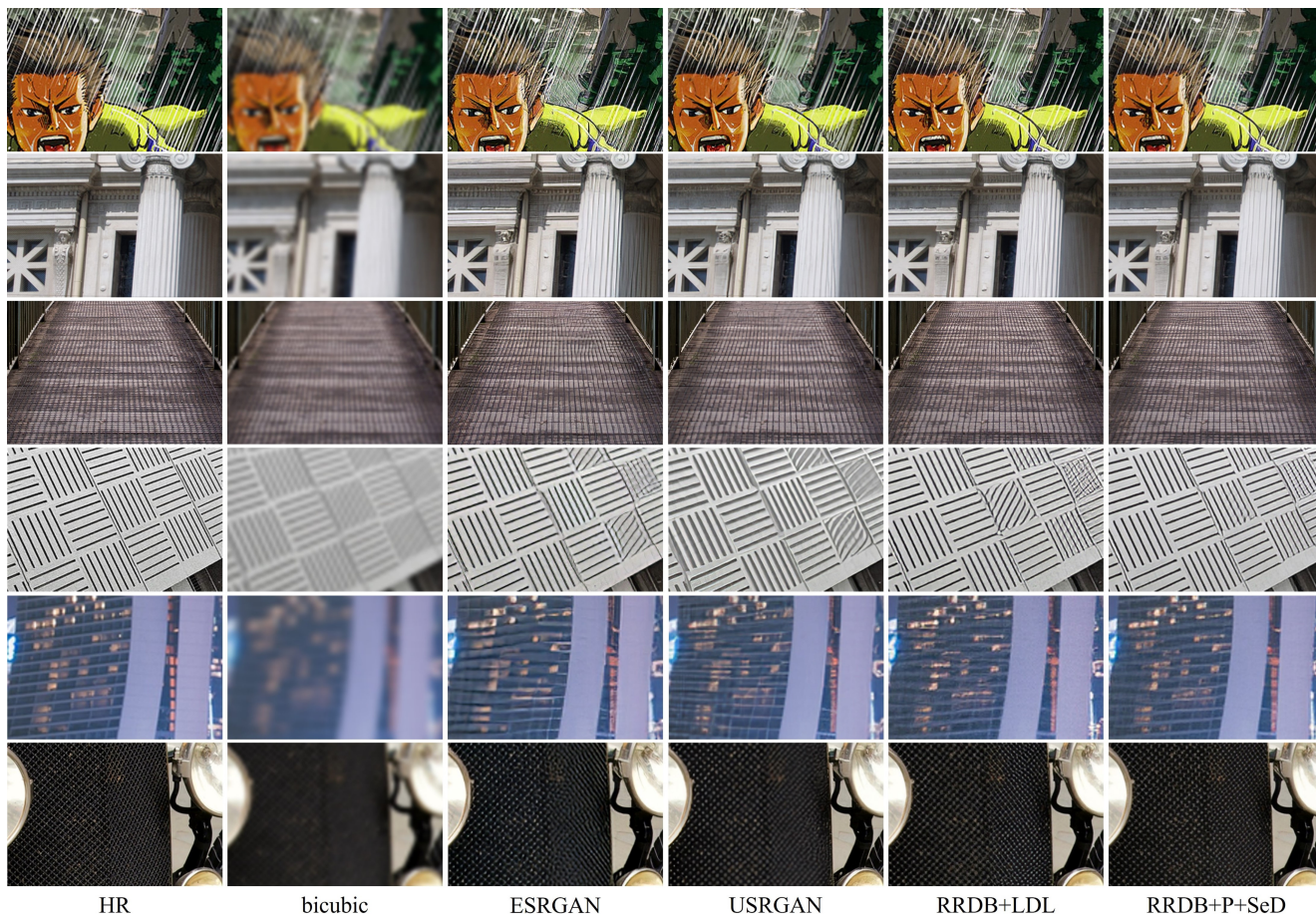
Figure 3. Visual comparison (zoom-in for better view) to state-of-the-art GAN-based SR methods. We demonstrate patch-wise SeD here because it shows better subjective quality. With SeD, the SR network is capable of restoring photo-realistic textures.

| HR | bicubic | ESRGAN | USRGAN | RRDB+LDL | RRDB+P+SeD |

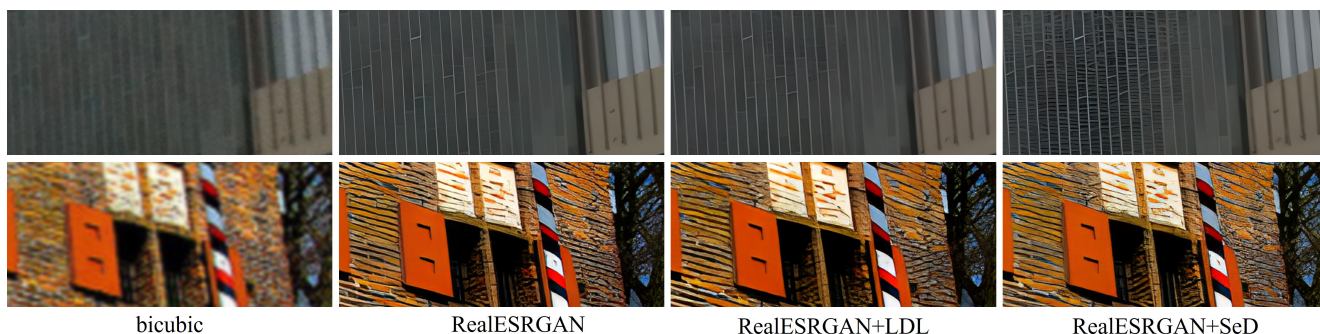| bicubic | RealESRGAN | RealESRGAN+LDL | RealESRGAN+SeD |

Figure 4. Visual comparison (zoom-in for better view) to state-of-the-art GAN-based real-world SR methods. We demonstrate pixel-wise SeD here to align with previous works [35, 60].

4.3. Results on real-world image SR

We evaluate the performance of SeD on the real-world image SR problem by incorporating it into state-of-the-art methods, including Real-ESRGAN [60] and LDL [35]. Typically, Real-ESRGAN and LDL both utilize RRDB as a generator, but LDL replaces the vanilla U-Net discriminator with there artifact mapped one. We also conduct experiments with a more powerful transformer-based generator, SwinIR [34], the qualitative results of SwinIR are provided in the **Supplementary**. All of the above-mentioned meth-

Method	Set5	Set14	DIV2K	Urban100	Manga109
SFTGAN	0.080/30.06/0.848	-/-/-	0.133/28.08/0.771	0.134/24.34/0.723	0.072/28.17/0.856
ESRGAN	0.076/30.44/0.852	0.133/26.28/0.699	0.115/28.20/0.777	0.123/24.37/0.734	0.065/28.41/0.859
USRGAN	0.079/30.91/0.866	0.143/27.15/0.736	0.132/28.77/0.793	0.133/24.89/0.750	0.063/28.75/0.872
RRDB+LDL	0.069/31.03/0.861	0.121/26.94/0.721	0.101/28.95/0.795	0.110/25.50/0.767	0.055/29.41/0.875
RRDB+DualFormer	0.068/31.40/0.872	0.121/27.53/0.741	0.103/29.30/0.802	0.115/25.73/0.774	0.053/29.90/0.886
RRDB+P	0.070/30.67/0.860	0.130/26.92/0.724	0.111/28.71/0.792	0.120/24.86/0.752	0.058/28.60/0.872
RRDB+P+SeD	**0.064**/31.22/0.867	**0.117**/27.37/0.736	**0.094**/29.27/0.802	**0.106**/25.93/0.779	**0.048**/29.99/0.888
RRDB+U	0.072/31.13/0.869	0.127/27.52/0.739	0.110/29.28/0.802	0.125/25.61/0.768	0.056/29.49/0.882
RRDB+U+SeD	**0.069**/31.73/0.880	**0.123**/27.94/0.757	**0.102**/29.85/0.818	**0.112**/26.20/0.788	**0.047**/30.46/0.897
SwinIR+LDL	0.065/31.03/0.861	0.118/27.22/0.732	0.094/29.12/0.801	0.102/26.23/0.792	0.047/30.14/0.888
SwinIR+P	0.070/31.49/0.876	0.127/27.66/0.747	0.103/29.66/0.815	0.107/26.22/0.790	0.048/30.18/0.895
SwinIR+P+SeD	**0.061**/31.44/0.870	**0.115**/27.53/0.742	**0.090**/29.53/0.810	**0.097**/26.45/0.794	**0.044**/30.48/0.896
SwinIR+U	**0.064**/31.38/0.869	0.120/27.64/0.744	**0.095**/29.56/0.810	0.103/26.09/0.786	0.049/29.99/0.889
SwinIR+U+SeD	0.067/31.64/0.874	**0.117**/27.84/0.750	0.096/29.79/0.816	**0.102**/26.46/0.796	**0.045**/30.58/0.898

Table 1. Quantitative comparison between GAN-based SR methods and the proposed SeD. Here, we use "+P" to denote a PatchGAN discriminator; "+U" to denote a U-Net discriminator; "+SeD" to denote our implementation based on two different discriminator architectures. The best perceptual results of each group are highlighted in bold. Each result is in term of LPIPS↓/PSNR↑/SSIM↑, ↑ and ↓ mean that the larger or smaller score is better, respectively.

ods utilize the pixel-wise U-Net discriminator, therefore, we conduct our experiments with the pixel-wise discriminator "U+SeD".

We demonstrate the quantitative comparison in Table 2. Our SeD surpasses the vanilla discriminator on most of the real-world datasets in terms of different backbones and training strategies. The results suggest that our SeD is capable of discriminating heavily distorted real-world images, and learning the fine-grained texture generation ability. As shown in Fig. 4, for the first line, images restored by SeD contain more realistic textures; for the second line, SeD eliminates the smooth problem and generates vivid textures for bricks. We provide more visualizations in the **Supplementary**.

4.4. Ablation studies for SeD

We conduct ablation studies to investigate the most suitable way of introducing semantic guidance into discriminator. We focus on three aspects: i) What is the best way to fuse semantic features in SeFB? ii) Which layer of CLIP contains the most useful semantic information for texture discrimination? iii) Which pretrained model is more suitable as a semantic extractor?

Effects of different fusion methods. There are a series of fusion methods to introduce the priors, including feature concatenation, spatial attention, or channel attention. We compare our proposed SeFB (*i.e.*, SeD-Our) with different fusion strategies in Fig. 3. The channel attention (*i.e.*, SeD-B) destroys the spatial information of semantics, which achieves the most worse performance. The performance of concatenation operation (*i.e.*, SeD-A) and spa-

Method	Type	DPED	OST300	RealSRSet
RealESRGAN	V	5.27	2.82	5.80
	L	5.36	2.83	6.07
	S	**4.49**	**2.73**	**5.34**
SwinIR	V	**4.95**	2.93	5.49
	L	5.65	3.10	5.59
	S	5.16	**2.75**	**5.15**

Table 2. The quantitative comparisons of our proposed SeD with the state-of-the-art methods on Real-world SR in terms of NIQE↓. "V" denotes "vanilla discriminator", "S" denotes "SeD", and "L" denotes "LDL".

Datasets	Fusion methods (LPIPS↓)			
	SeD-A	SeD-B	SeD-C	SeD-our
Set5	0.069	0.085	0.074	**0.064**
Urban100	0.124	0.131	0.125	**0.106**
Manga109	0.058	0.065	0.055	**0.048**

Table 3. Ablation studies for different fusion methods in SeFB. We perform all the experiments on classical image SR, and evaluate with subjective metric LPIPS. "A", "B", "C" denote for concatenate, channel attention, spatial attention, respectively.

tial attention (*i.e.*, SeD-C) are comparable, which is obviously lower than our proposed semantic-aware fusion block (SeFB). More details can be found in the **Supplementary** for the implementations of the above fusion strategies.

Methods	Datasets (LPIPS↓)		
	Set5	Urban100	Manga109
SeD-Layer1	0.066	0.120	0.057
SeD-Layer2	0.068	0.117	0.056
SeD-Layer3	**0.064**	**0.106**	**0.048**
SeD-Layer4	0.069	0.122	0.055

Table 4. Ablation for the semantics from different layers of CLIP.

Method	Datasets (LPIPS↓)		
	Set5	Urban100	Manga109
ResNet-50	0.066	0.117	0.055
CLIP	**0.064**	**0.106**	**0.048**

Table 5. Ablation studies for different semantic extractors.

The semantics of different layers. It is noteworthy that the semantics from different layers of CLIP [46] is different. With the increase of layers, the semantics will be more representative while the resolution becomes lower. The essence of effective semantic guidance lies in extracting semantic features that are sufficiently deep yet retain essential spatial information to ensure the quality of restoration guidance. There are in total four layers inside CLIP semantic extractor, and the spatial resolution is halved after each layer. To investigate which layer is the best choice for the guidance, we conduct the experiments in Table 4. From the table, In the first and second layers, despite the semantic feature possessing a relatively large spatial resolution, its lack of concentration could result in suboptimal semantic guidance for the discrimination process. For the last layer, the spatial resolution is too small, which may causes discriminator ignoring the pixel-wise consistency. As the result, the semantics from the third layer performs best, which is used in our paper.

The effects of the different semantic extractors. To delve deeper into the question of whether finer-grained semantic understanding enhances guidance quality, we compare two distinct semantic extractors. Specifically, we evaluate a pretrained ResNet-50 [19] using the ImageNet [10] dataset against the "RN50" model from the pretrained CLIP, which utilizes a vision-language dataset, as detailed in Table. 5. The results of our experiments indicate that the "RN50" model pretrained with vision-language data outperforms the former. This superior performance is attributed to the language descriptions in CLIP's training regimen, which provide more comprehended and detailed semantic context compared to traditional digital labels used in classification tasks. This finding underscores the potential of language-enhanced models in achieving a more refined understanding of semantics, thereby offering more effective guidance in discriminative processes.

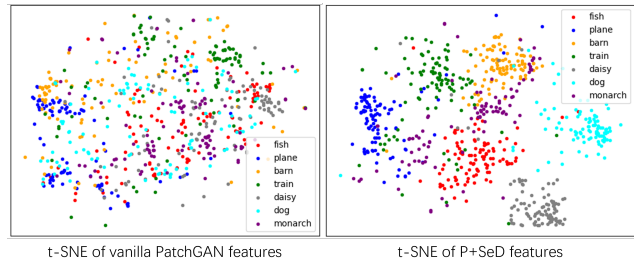### 4.5. T-SNE visualization of discriminator features

Figure 5. The t-SNE visualization of discriminator features. The 7 categories are "fish", "plane", "barn", "train", "daisy", "dog", "monarch" from ImageNet, respectively.

To validate that our semantic guidance promotes the discriminator to perform more fine-grained discrimination, we visualize the intermediate features for both vanilla discriminator and our SeD with t-SNE [57]. We randomly select 7 categories from ImageNet [10], with 100 images each category. Then, we feed all the images into a vanilla PatchGAN discriminator and our P+SeD. We visualize the features after the first SeFB of SeD, and the features after the first BN layer of PatchGAN discriminator, respectively. As shown in Fig. 5, the features of SeD are well-clustered, while the features of vanilla patch discriminator are disorganized. This provides the evidence that under the guidance of semantics, SeD is capable of performing more fine-grained discrimination, as claimed in the paper.

5. Conclusion

In this paper, we propose a simple but effective Semantic-aware Discriminator (SeD) for Image Super-resolution. We find the distribution learning of previous discriminators is overly coarse-grained, which causes the virtual or counter-intuitive textures. To mitigate this, we introduce the semantics of the image from pretrained vision models (PVMs) as the condition of the discriminator with our proposed semantic-aware fusion block (SeFB), which encourages the SR network to learn the fine-grained semantic-aware textures. Moreover, our SeD is easy to be incorporated into most GAN-based SR methods and achieves excellent performance. Extensive experiments on classic SR and Real-world SR have demonstrated the effectiveness of our SeD.

Acknowledgements

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] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 5

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[4] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K Wong. Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing*, 30:1219–1231, 2020. 2

[5] 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. 1

[6] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 3

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 3

[8] Marcos V Conde and Kerem Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3956–3960, 2021. 4

[9] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 1, 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 8

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[12] 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, 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 3

[15] 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 (CVPR) Workshops*, pages 1103–1112, 2022. 1

[16] Shangqi Gao and Xiahai Zhuang. Multi-scale deep neural networks for real image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[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. 1

[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. 5

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 8

[20] Ruifei He, Shuyang Sun, Jihan Yang, Song Bai, and Xiaojuan Qi. Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9161–9171, 2022. 3

[21] 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

[22] 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. 1, 5

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 4

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[25] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional net-

works. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[26] 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. 3, 4

[27] 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. 1, 2, 3, 4

[28] Bingchen Li, Xin Li, Yiting Lu, Sen Liu, Ruoyu Feng, and Zhibo Chen. Hst: Hierarchical swin transformer for compressed image super-resolution. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 651–668. Springer, 2023. 2

[29] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European conference on computer vision (ECCV)*, pages 517–532, 2018. 2

[30] Xin Li, Xin Jin, Jianxin Lin, Sen Liu, Yaojun Wu, Tao Yu, Wei Zhou, and Zhibo Chen. Learning disentangled feature representation for hybrid-distorted image restoration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 313–329. Springer, 2020. 5

[31] Xin Li, Bingchen Li, Xin Jin, Cuiling Lan, and Zhibo Chen. Learning distortion invariant representation for image restoration from a causality perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1714–1724, 2023. 1

[32] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023. 1

[33] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[34] 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. 1, 2, 5, 6

[35] 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. 5, 6

[36] 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

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[38] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 3

[39] 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 (CVPR) Workshops*, pages 457–466, 2022. 1, 2

[40] Zhihe Lu, Jiawang Bai, Xin Li, Zeyu Xiao, and Xinchao Wang. Beyond sole strength: Customized ensembles for generalized vision-language models. *arXiv preprint arXiv:2311.17091*, 2023. 4

[41] 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. 5

[42] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Learning the degradation distribution for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6063–6072, 2022. 2

[43] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 5

[44] Yingxue Pang, Xin Li, Xin Jin, Yaojun Wu, Jianzhao Liu, Sen Liu, and Zhibo Chen. Fan: frequency aggregation network for real image super-resolution. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 468–483. Springer, 2020. 2

[45] 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. 4

[46] 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. 3, 8

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3, 4

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

[49] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038*, 2022. 4

[50] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[51] Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8322–8332, 2022. 3

[52] 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. 1, 2, 5

[53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 3

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 4

[55] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 3

[56] 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 workshops*, pages 114–125, 2017. 5

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

[58] 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. 5

[59] 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. 1, 2, 3, 4, 5

[60] 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. 1, 2, 4, 5, 6

[61] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 2

[62] Xiaoqian Xu, Pengxu Wei, Weikai Chen, Yang Liu, Mingzhi Mao, Liang Lin, and Guanbin Li. Dual adversarial adaptation for cross-device real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5667–5676, 2022. 2

[63] 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 (CVPR)*, 2020. 1, 2

[64] Ren Yang, Radu Timofte, Xin Li, Qi Zhang, Lin Zhang, Fanglong Liu, Dongliang He, Fu Li, He Zheng, Weihang Yuan, et al. Aim 2022 challenge on super-resolution of compressed image and video: Dataset, methods and results. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 174–202. Springer, 2023. 2, 5

[65] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3

[66] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018. 3

[67] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 3

[68] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 5

[69] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3217–3226, 2020. 5

[70] 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. 1, 2, 4

[71] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 1, 3

[72] 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. 1, 2

[73] 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. 1, 2

[74] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Jianming Zhang, Sohrab Amirghodsi, Qing Liu, and Jiebo Luo. Panoptically guided image inpainting with image-level and object-level semantic discriminators, 2023. 2