Pixel-Aware Stable Diffusion for Realistic Image Super-Resolution and Personalized Stylization

Tao Yang[1], Rongyuan Wu[2], Peiran Ren[3], Xuansong Xie[3], and Lei Zhang[2]

[1] Bytedance Inc.
[2] The Hong Kong Polytechnic University
[3] Alibaba Group

Abstract. Diffusion models have demonstrated impressive performance in various image generation, editing, enhancement and translation tasks. In particular, the pre-trained text-to-image stable diffusion models provide a potential solution to the challenging realistic image super-resolution (Real-ISR) and image stylization problems with their strong generative priors. However, the existing methods along this line often fail to keep faithful pixel-wise image structures. If extra skip connections between the encoder and the decoder of a VAE are used to reproduce details, additional training in image space will be required, limiting the application to tasks in latent space such as image stylization. In this work, we propose a pixel-aware stable diffusion (PASD) network to achieve robust Real-ISR and personalized image stylization. Specifically, a pixel-aware cross attention module is introduced to enable diffusion models perceiving image local structures in pixel-wise level, while a degradation removal module is used to extract degradation insensitive features to guide the diffusion process together with image high level information. An adjustable noise schedule is introduced to further improve the image restoration results. By simply replacing the base diffusion model with a stylized one, PASD can generate diverse stylized images without collecting pairwise training data, and by shifting the base model with an aesthetic one, PASD can bring old photos back to life. Extensive experiments in a variety of image enhancement and stylization tasks demonstrate the effectiveness of our proposed PASD approach. Our source codes are available at https://github.com/yangxy/PASD/.

Keywords: Pixel-Aware Stable Diffusion · Realistic Image Super-Resolution · Image Stylization

1 Introduction

Real-world images often suffer from a mixture of complex degradations, such as low resolution, blur, noise, etc., in the acquisition process. While image restoration methods have achieved significant progress, especially in the era of deep learning [8, 32], they still tend to generate over-smoothed details, partially due to the pursue of image fidelity in the methodology design. By relaxing the constraint on image fidelity, realistic image super-resolution (Real-ISR) aims to reproduce perceptually realistic image details from the degraded observation. The

generative adversarial networks (GANs) [11] and the adversarial training strategy have been widely used for Real-ISR [4, 27, 54, 67] and achieved promising results. However, GAN-based Real-ISR methods are still limited in reproducing rich and realistic image details and tend to generate unpleasant visual artifacts. Meanwhile, GAN-based methods have also been widely used in various image stylization tasks such as cartoonization and old-photo restoration. For example, Chen *et al.* [6] proposed CartoonGAN to generate cartoon stylization by using unpaired data for training. However, different models need to be trained for different styles. Wan *et al.* [49] introduced a triplet domain translation network to restore old photos. While achieving promising results, the multi-stage procedure of this method makes it complex to use.

Recently, denoising diffusion probabilistic models (DDPMs) have shown outstanding performance in tasks of image generation [16], and it has become a strong alternative to GAN due to its powerful capability in approximating diverse and complicated distributions. With DDPM, the pre-trained text-to-image (T2I) and text-to-video (T2V) latent diffusion models [15, 41, 43, 46] have been popularly used in numerous downstream tasks [2, 45, 64]. Diffusion models have also been adopted to solve image restoration tasks. A denoising diffusion restoration model (DDRM) is proposed in [21] to solve inverse problem by taking advantage of a pre-trained denoising diffusion generative model. However, DDRM assumes a linear image degradation model, limiting its application to more practical scenarios such as Real-ISR. Considering that the pre-trained T2I models such as Stable Diffusion (SD) [43] can generate high-quality natural images, Zhang and Agrawala [64] proposed ControlNet, which enables conditional inputs like edge maps, segmentation maps, etc., and demonstrated that the generative diffusion priors are also powerful in conditional image synthesis. Unfortunately, ControlNet is not suitable for pixel-wise conditional control (see Fig. 1 for an example). Liu *et al.* [35] and Wang *et al.* [50] demonstrated that pre-trained SD priors can be employed for image colorization and Real-ISR, respectively. However, they resorted to a skipped connection to pass pixel-level details for image restoration, requiring extra training in image space and limiting the model capability to tasks performed in latent space such as image stylization.

In this work, we aim to develop a flexible model to achieve Real-ISR and personalized stylization by using pre-trained T2I models such as SD [43], targeting at reconstructing photo-realistic pixel-level structures and textures. Our idea is to introduce pixel-aware conditional control into the diffusion process so that robust and perceptually realistic outputs can be achieved. To this end, we present a pixel-aware cross attention (PACA) module to perceive pixel-level information without using any skipped connections. A degradation removal module is employed to reduce the impact of unknown image degradations, alleviating the burden of diffusion module to handle real-world low-quality images. We also demonstrate that the high-level classification/detection/captioning information extracted from the input image can further boost the Real-ISR performance. Inspired by recent works [10,33], we present an adjustable noise schedule to further boost the performance of Real-ISR and image stylization tasks. In particular,
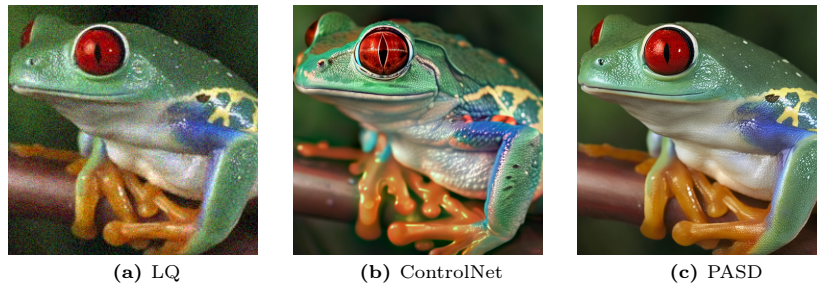
(a) LQ   (b) ControlNet   (c) PASD

Fig. 1: From left to right: an input LQ image, the Real-ISR outputs by ControlNet [64] and our PASD method. One can see that the output by ControlNet has clear content inconsistency with the input, while PASD preserves the structure well in pixel level.

the proposed method, namely pixel-aware stable diffusion (PASD), can perform personalized stylization tasks (*e.g.*, caroonization and old photo restoration) by simply shifting the base model to a personalized one. Extensive experiments demonstrate the effectiveness and flexibility of PASD.

2   Related Work

Realistic Image Super-Resolution. Though deep learning based image super-resolution [8,32] has achieved significant progress, they still suffer from over-smoothed details due to the high illness of the task by minimizing the fidelity objectives (*e.g.*, PSNR, SSIM). Realistic image super-resolution (Real-ISR) aims to reproduce perceptually photo-realistic image details by optimizing not only the fidelity but also the perception objectives. The GAN [11] network and its adversarial training strategies are widely in Real-ISR [27,54]. Basically, a generator network is used to reconstruct the desired high-quality (HQ) image from the low-quality (LQ) input, while a discriminator network is used to judge whether the HQ output is perceptually realistic. In the early study, bicubic downsampling or some simple degradations [8,26] are used to simulate the LQ-HQ training pairs. Cai *et al.* [3] collected a real-world dataset with paired LQ-HQ images by zooming camera lens. Zhang *et al.* [63] and Wang *et al.* [52] later modeled complex degradations by shuffling degradation types and using a high-order process, respectively. Recently, Yang *et al.* [60] took the advantages of handcrafted degradation models and generative diffusion models to synthesize realistic LQ-HQ training pairs.

Though GAN-based models have dominated the previous research in Real-ISR, adversarial training is unstable and the GAN-based Real-ISR methods often bring unnatural visual artifacts. Liang *et al.* [30] proposed a locally discriminative learning approach to suppress the GAN-generated artifacts, yet it is difficult to introduce additional details. Recently, inspired by the success of generative priors in face restoration tasks [51,58], some works have been proposed to leverage the priors learned by diffusion models [16] and pre-trained T2I models [43] to solve the Real-ISR problems and obtain interesting results [4,21,50,57]. In

this work, we aim to develop an SD based Real-ISR model which can achieve pixel-level restoration of image details and textures.

Personalized Stylization. Inspired by the powerful learning capacity of deep neural networks, Gatys *et al.* [9] presented an optimization based method to transfer the style of a given artwork to a content image. This work was extended by many following researches [19, 29, 66]. However, all these methods require an extra image as style input. This problem can be alleviated by resorting to an image-to-image framework [5, 6, 68]. Due to the lack of pairwise training data, some works [37, 59] focus on portrait stylization with the help of StyleGAN [20]. With the rapid development of SD models [43], some works [2, 64] generate stylized images by using proper instruction prompts, achieving impressive results. However, these methods fail to maintain pixel-wise image structures in the stylization process, and they lack the ability to mimic the appearance of subjects in a given reference set. To meet the specific needs of different users, Ruiz *et al.* [45] and Kumari *et al.* [25] proposed personalized stylization approaches for T2I diffusion models. While achieving pleasant stylization results, these methods require extra training procedures for different reference sets.

Old photo restoration, which is a challenging task due to the unknown mixed degradations, can also be viewed as a stylization problem. Most existing works focus on a specific issue such as crack removal [61], super-resolution [4, 52] and facial region enhancement [51, 58]. Wan *et al.* [49] proposed to address this problem using a novel triplet domain translation network. They narrowed the domain gap between real old photos and synthetic ones in the compact latent space and learned to restore old photos via latent space translation. This multi-stage approach yields interesting, yet it is unstable and is not easy to use.

Diffusion Probabilistic Models. The seminal work of DDPM [16] has demonstrated strong capability in generating high quality natural images. Considering that DDPMs require hundreds of sampling steps in the denoising process, Song *et al.* [47] proposed DDIM to accelerate the sampling speed. Following works extend DDPM/DDIM by adapting high-order solvers [36] and distillations [38]. Rombach *et al.* [44] extended DDPM to latent space and demonstrated impressive results with less computational costs. This work sparks the prosperity of large pre-trained T2I and T2V diffusion models such as SD [43], Imagen [15]. It has been demonstrated that T2I diffusion priors are more powerful than GAN priors in handling diverse natural images [41, 43, 46]. Kawar *et al.* [22] applied complex text-guided semantic editing to real images. ControlNet [64] enables conditional inputs, such as edge maps, segmentation maps, keypoints, *etc.*, to T2I models. Liu *et al.* [35] and Wang *et al.* [50] respectively utilized generative diffusion priors to image colorization and super-resolution.

3   Pixel-Aware Stable Diffusion Network

Our method is based on generative diffusion priors. In particular, we utilize the powerful pre-trained SD [43] model, while alternative diffusion models such as DALLE2 [41] and Imagen [46] can also be adopted. The architecture of our
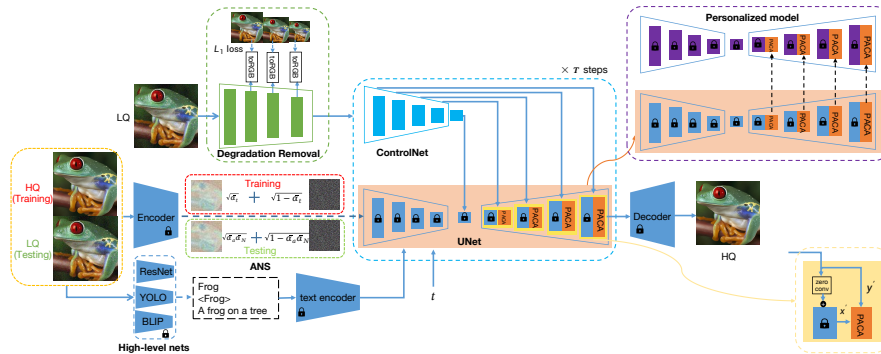
Fig. 2: Architecture of the proposed pixel-aware stable diffusion (PASD) network. PASD consists of several modules, including Degradation Removal, ControlNet, PACA, ANS, and High-level Nets. During training, the encoder maps the HQ image to a latent representation, which is then added by noise to yield a noisy latent. In testing, the LQ image is used to generate the noisy input, while ANS is employed for flexible perception-fidelity balance. The noisy latent is fed to the UNet along with timestep, high-level information, and the output of ControlNet after Degradation Removal conditioned on the LQ image. In particular, the output of ControlNet is added to the UNet via PACA in latent space. PASD can be readily used for personalized stylization by simply switching the base diffusion model to a personalized one.

pixel-aware stable diffusion (PASD) network is depicted in Fig. 2. One can see that in addition to the pre-trained SD model, PASD has four main modules: a degradation removal module to extract degradation insensitive low-level control features, a high-level information extraction module to extract semantic control features, an adjustable noise schedule (ANS) and a pixel-aware cross-attention (PACA) module to perform pixel-level guidance for diffusion. In addition to the Real-ISR task, our PASD can be readily used for personalized stylization by simply switching the base diffusion model to a personalized one.

3.1 Degradation Removal Module

Real-world LQ images usually suffer from complex and unknown degradations. We thus employ a degradation removal module to reduce the impact of degradations and extract "clean" features from the LQ image to control the diffusion process. As shown in Fig. 2, we adopt a pyramid network to extract multi-scale feature maps with 1/2, 1/4 and 1/8 scaled resolutions of the input LQ image. Intuitively, it is anticipated that these features can be used to approximate the HQ image at the corresponding scale as close as possible so that the subsequent diffusion module could focus on recovering realistic image details, alleviating the burden of distinguishing image degradations. Therefore, we introduce an intermediate supervision by employing a convolution layer "toRGB" to turn every single-scale feature maps into the HQ RGB image space. We apply an L_1 loss on each resolution scale to force the reconstruction at that scale to be close to the pyramid decomposition of the HQ image: $\mathcal{L}_{\mathcal{DR}} = \sum_s ||\mathbf{I}_{hq}^s - \mathbf{I}_{sr}^s||_1$, where

$\mathbf{I}_{hq}^s$ and $\mathbf{I}_{sr}^s$ represent the HQ ground-truth and ISR output at scale s. Note that this module is only required in the Real-ISR task.

3.2   Pixel-Aware Cross Attention (PACA)

The main challenge of utilizing pre-trained T2I diffusion priors for image restoration tasks lies in how to enable the diffusion process be aware of image details and textures in pixel-level. The well-known ControlNet can support task-specific conditions (*e.g.*, edges, segmentation masks) well but fail for pixel-level control. Given a feature map $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$ from U-Net, where $\{h, w, c\}$ are feature height, width and channel numbers, and a skipped feature map $\mathbf{y} \in \mathbb{R}^{h \times w \times c}$ from ControlNet, Zhang and Agrawala [64] proposed a unique type of convolution layer $\mathcal{Z}$ called "zero convolution" to connect them:

$$\tilde{\mathbf{x}} = \mathbf{x} + \mathcal{Z}(\mathbf{y}), \tag{1}$$

where $\tilde{\mathbf{x}}$ is the output feature map. The zero convolution is easy-to-implement. However, simply adding the feature maps from the two networks may fail to pass pixel-level precise information, leading to structure inconsistency between the input LQ and output HQ images. Fig. 1 shows an example. One can see that by simply applying ControlNet to the LQ input, there are obvious structure inconsistencies in the output image by ControlNet.

To address this problem, some methods employ a skipped connection outside the U-Net [50] to add image details. However, this introduces additional training in image feature domain, and limits the application of the trained network to tasks performed in latent space (*e.g.*, image stylization). In this work, we introduce a simple pixel-aware cross attention (PACA) to solve this issue. We reshape $\mathbf{x}$ and $\mathbf{y}$ to $\mathbf{x}' \in \mathbb{R}^{h*w \times c}$ and $\mathbf{y}' \in \mathbb{R}^{h*w \times c}$, and consider $\mathbf{y}'$ as the context input. The PACA (see the brown-colored block in Fig. 2) can be computed as follows:

$$PACA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}) \cdot \mathbf{V}, \tag{2}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are calculated according to operations $to\_q(\mathbf{x}'), to\_k(\mathbf{y}')$ and $to\_v(\mathbf{y}')$, respectively.

The conditional feature input $\mathbf{y}'$ is of length $h*w$, which equals to the total number of pixels of latent feature $\mathbf{x}$. Since feature $\mathbf{y}'$ has not been converted into the latent space by the Encoder, it preserves well the original image structures. Therefore, our PASD model can manage to perceive pixel-wise information from the conditional input $\mathbf{y}'$ via PACA. As can be seen in the experimental result section, with the help of PACA, the output of our PASD network can reproduce realistic and faithful image structures and textures in pixel-level.

3.3   Adjustable Noise Schedule (ANS)

As discussed in previous works [10, 33], the noise schedule used in SD [43] suffers train-test discrepancy. In training, the noise schedule leaves some residual

signal even at the terminal diffusion timestep N, leading to non-zero signal-to-noise ratio (SNR). This weakens the model performance at test time when we sample from random Gaussian noise without the signal information. To address this issue, we propose an adjustable noise schedule (ANS) by introducing signal information from the input image at test time.

The residual signals at training stage are from the HQ ground-truth data, which are unavailable at test time. In applications such as Real-ISR, we can embed the input LQ latent into the initial random Gaussian noise at terminal diffusion timestep N as a compensation:

$$\mathbf{z}_N = \sqrt{\bar{\alpha}_N}\mathbf{z}_{LR} + \sqrt{1 - \bar{\alpha}_N}\mathbf{z}, \tag{3}$$

where $\mathbf{z}_N$, $\mathbf{z}_{LR}$, $\bar{\alpha}_N$, $\mathbf{z}$ are respectively the latent input at timestep N, the LQ latent, the cumulative product of α, and the initial random Gaussian noise. This remedy can partially alleviate the discrepancy issue and has been adopted in [50, 57]. However, the train-test discrepancy still exists due to the different origins of residual signals, which can harm the restoration results when the LQ image suffers from severe degradations.

To suppress the side effect of residual signal from LQ data, we introduce an additional Gaussian noise $\mathbf{z}'$ with level $\bar{\alpha}_a \in [0, 1]$ into Eq. (3) as: $\mathbf{z}_N = \sqrt{\bar{\alpha}_a}(\sqrt{\bar{\alpha}_N}\mathbf{z}_{LR} + \sqrt{1 - \bar{\alpha}_N}\mathbf{z}) + \sqrt{1 - \bar{\alpha}_a}\mathbf{z}' = \sqrt{\bar{\alpha}_a\bar{\alpha}_N}\mathbf{z}_{LR} + \sqrt{\bar{\alpha}_a - \bar{\alpha}_a\bar{\alpha}_N}\mathbf{z} + \sqrt{1 - \bar{\alpha}_a}\mathbf{z}'$. Since Gaussian noises $\mathbf{z}$ and $\mathbf{z}'$ are independent, the combination of them is equivalent to another Gaussian noise $\mathbf{z}''$. The above formula can be re-written as follows:

$$\mathbf{z}_N = \sqrt{\bar{\alpha}_a\bar{\alpha}_N}\mathbf{z}_{LR} + \sqrt{1 - \bar{\alpha}_a\bar{\alpha}_N}\mathbf{z}''. \tag{4}$$

In this way, by choosing a proper value of $\bar{\alpha}_a$, we can adjust the strength of the residual signal $\mathbf{z}_{LR}$ to enable flexible perception-fidelity trade-off.

3.4   High-Level Information

Our method is based on the pre-trained SD model where text is used as the input, while in tasks such as Real-ISR, the LQ image is available as the input. Though some SD-based Real-ISR methods [50] adopt the null-text prompt, it has been demonstrated that content-related captions could improve the synthesis results [43]. As shown in Fig. 2, we employ the pre-trained ResNet [13], YOLO [42] and BLIP [28] networks to extract image classification, object detection and image caption information from the LQ input, and employ the CLIP [40] encoder to convert the text information into image-level features, providing additional semantic signal to control the diffusion process.

The classifier-free guidance [17] is adopted in our method: $\tilde{\epsilon}(\mathbf{z}_t, \mathbf{c}) = \epsilon(\mathbf{z}_t, \mathbf{c}) + \omega\epsilon(\mathbf{z}_t, \mathbf{c}_{neg})$, where $\tilde{\epsilon}(\mathbf{z}_t, \mathbf{c})$ and $\epsilon(\mathbf{z}_t, \mathbf{c}_{neg})$ are conditional and unconditional ϵ-predictions [16], $\mathbf{c}$ and $\mathbf{c}_n$ are respectively the positive and negative text prompts, $\mathbf{z}_t$ is the latent feature at step t, and ω adjusts the guidance scale. The unconditional ϵ-prediction $\epsilon(\mathbf{z}_t, \mathbf{c}_{neg})$ can be achieved with negative prompts. In practice, we empirically combine words like "noisy", "blurry", "low resolution" as negative

prompts, which play a key role to trade off mode coverage and sample quality during inference. It is optional but could boost much the Real-ISR performance.

3.5   Application to Personalized Stylization

Personalized stylization. Thanks to the open source of SD [43] and the recently developed techniques such as DreamBooth [45] and LORA [18], the community becomes highly prosperous. Contributors can upload a large amount of personalized models finetuned on SD with self-collected data. Since PASD is based on pretrained SD model and the pretrained weights are frozen during model training, it is easy to replace the base model with personalized ones at test time [12] (as illustrated in the top-right corner of Fig. 2) so that PASD can re-target the output domain and produce stylized results.

Unlike previous methods [5, 6, 68] that achieve stylization ability by learning a pixel-to-pixel mapping function using adversarial training, our PASD approach decouples stylization generation and pixel-to-pixel mapping, opening a new door for image stylization. By fine-tuning personalized SD models with a batch of style images or downloading different personalized models from online communities [4], one can easily generate various stylized results with our PASD method. In this paper, we use cartoonization as a typical stylization task in experiments.

Old photo restoration. Apart from cartoonization, another popular family of personalized models are the aesthetic ones, *i.e.*, those trained on images with a particular aesthetic taste. One typical task of this kind is old photo restoration. By replacing the base model with an aesthetic one, PASD can improve the quality and aesthetics of the input old photo image simultaneously, as will be demonstrated in our experiments.

3.6   Training Strategy

In the model training, we first obtain the latent representation $\mathbf{z}_0$ of an HQ image, and progressively add noise to it to yield a noisy latent $\mathbf{z}_t$, where t is a randomly sampled diffusion step. Given a number of conditions such as diffusion step t, LQ input $\mathbf{I}_{lq}$ and text prompt $\mathbf{c}$, we learn a PASD network ϵ_θ to predict the noise added to the noisy latent $\mathbf{z}_t$. The optimization objective is:

$$\mathcal{L}_{\mathcal{DF}-\epsilon} = \mathbb{E}_{\mathbf{z}_0,t,\mathbf{c},\mathbf{I}_{lq},\epsilon\sim\mathcal{N}(0,1)}\big[||\epsilon - \epsilon_\theta(\mathbf{z}_t,t,\mathbf{c},\mathbf{I}_{lq})||_2^2\big]. \tag{5}$$

During the training of Real-ISR models, we jointly update the degradation removal module. The total loss is $\mathcal{L} = \mathcal{L}_{\mathcal{DF}-\epsilon} + \gamma\mathcal{L}_{\mathcal{DR}}$, where γ is a balancing parameter. We simply set $\gamma = 1$ in the experiments. We freeze all the parameters in pre-trained SD, and only train the newly added modules, including the degradation removal module, ControlNet and PACA. The employed ResNet, YOLO and BLIP and CLIP networks for high-level information extraction are also fixed. During training, we randomly replace 50% of the text prompts with null-text prompts. This encourages our PASD model to perceive semantic contents from input LQ images as a replacement of text prompts.
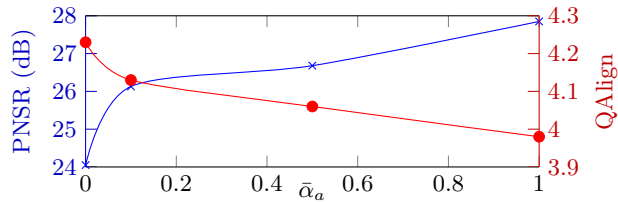
---
[4] https://civitai.com/; https://huggingface.co/

Fig. 3: The curves of PSNR/QAlign versus $\bar{\alpha}_a$.

Input $\bar{\alpha}_a = 1$ $\bar{\alpha}_a = 0.5$ $\bar{\alpha}_a = 0$
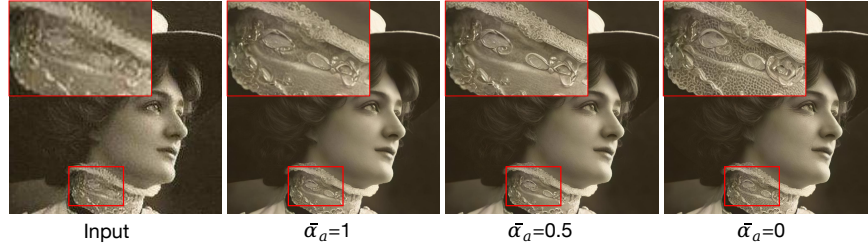
Fig. 4: Visual results with different values of $\bar{\alpha}_a$.

4   Experiments

4.1   Experiment Setup

We adopt the Adam optimizer [24] to train PASD with a batch size of 4. The learning rate is fixed as 5×10^{-5}. The model is updated for $500K$ iterations with 8 NVIDIA Tesla 32G-V100 GPUs.

Training and testing datasets. For the Real-ISR task, we train PASD on DIV2K [48], Flickr2K [1], OST [53], and the first $10,000$ face images from FFHQ [20]. We employ the degradation pipeline of Real-ESRGAN [52] to synthesize LQ-HQ training pairs. We evaluate our approach on both synthetic and real-world datasets. The synthetic dataset is generated from the DIV2K validation set following the Real-ESRGAN degradation pipeline. For real-world test dataset, we use the RealSR [3] and DRealSR [55] for evaluation.

For the task of cartoonization, we simply reuse the PASD model trained for Real-ISR task and shift the base model with stylized ones obtained from online communities. We conduct comparisons on the first 100 face images from FFHQ as well as the first 100 images from Flicker2K.

For the task of old photo restoration, we also adopt the pre-trained PASD model in the task of Real-ISR. Unlike cartoonization, we replace the base model with aesthetic ones. We collect 100 old photos from Internet for testing.

Evaluation metrics. For quantitative evaluation of Real-ISR models, we employ the widely used perceptual metrics, including FID [14], LPIPS [65], DISTS [7], NIQE [39], MUSIQ [23] and QAlign [56], to compare the competing Real-ISR models. The PSN and SSIM indices (evaluated on the Y channel in YCbCr space) are also reported for reference only because they are not suitable to evaluate generative models. For the tasks of cartoonization and old photo restoration, we employ FID [14], MUSIQ [23] and QAlign [56] for evaluation since the ground-truth images are unavailable.
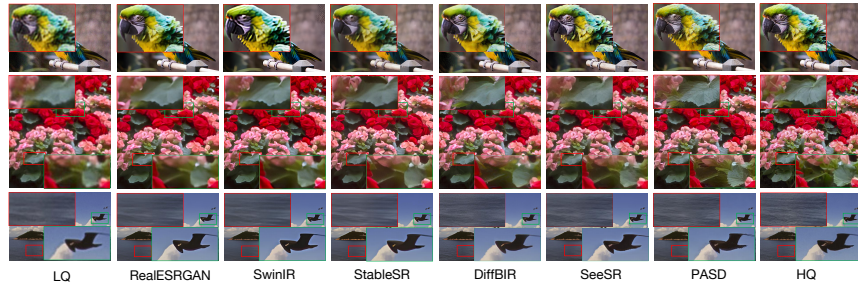
| LQ | RealESRGAN | SwinIR | StableSR | DiffBIR | SeeSR | PASD | HQ |

Fig. 5: Realistic image super-resolution results by different methods.

Table 1: The PSNR, SSIM, LPIPS, FID, DISTS, MUSIQ, NIQE and QAlign indices of different Real-ISR models on synthesized (DIV2K) and real-world (RealSR and DRealSR) test datasets. The best and second best scores are in bold and underscore.

Datasets	Metrics	RealESRGAN	FeMaSR	SwinIR	ResShift	StableSR	DiffBIR	SeeSR	PASD
DIV2K valid	PSNR↑	<u>21.9697</u>	21.0272	21.6136	21.8165	**22.3377**	21.2447	21.7714	21.6494
	SSIM↑	**0.6164**	0.5543	0.6028	0.5589	<u>0.6162</u>	0.5415	0.5943	0.5543
	LPIPS↓	0.3717	0.3925	0.3931	0.4631	0.4144	0.4189	**0.3577**	<u>0.3640</u>
	FID↓	68.6481	66.2590	71.1709	55.1131	50.9362	57.7150	**47.3327**	<u>50.7819</u>
	DISTS↓	0.2092	0.1858	0.2097	0.2363	0.2191	<u>0.1785</u>	0.1959	**0.1778**
	MUSIQ↑	61.6423	60.0450	58.7220	56.3572	49.3028	**66.9539**	<u>66.3071</u>	66.1278
	NIQE↓	3.6754	3.5621	3.6044	6.6093	4.4371	**2.8572**	3.9528	<u>3.3992</u>
	QAlign↑	4.2460	3.6302	3.9947	4.1360	3.6496	4.3099	<u>4.3146</u>	**4.3175**
RealSR	PSNR↑	25.8450	25.4330	26.0279	<u>26.2353</u>	26.2057	24.9872	**26.5952**	25.9301
	SSIM↑	0.7734	0.7540	**0.7802**	0.7047	<u>0.7767</u>	0.6812	0.7689	0.7105
	LPIPS↓	0.2729	0.2927	<u>0.2594</u>	0.4594	**0.2565**	0.3633	0.2796	0.2806
	FID↓	67.0156	63.4422	64.1658	61.6524	109.1054	<u>55.1668</u>	58.3206	**47.3440**
	DISTS↓	0.1685	0.1941	0.1609	0.2563	**0.1584**	0.1870	0.1859	<u>0.1604</u>
	MUSIQ↑	59.6881	58.7741	59.6442	53.3410	60.7065	<u>65.5173</u>	64.2653	**65.5979**
	NIQE↓	4.6781	4.7577	4.6453	7.4331	4.9309	**3.6873**	5.2673	<u>4.1886</u>
	QAlign↑	3.9186	3.6895	3.8204	3.8199	3.8686	<u>4.0999</u>	3.8884	**4.1250**
DRealSR	PSNR↑	27.9116	26.5869	27.8427	28.2573	**29.3013**	27.2030	<u>29.1033</u>	29.0948
	SSIM↑	0.8249	0.7688	0.8209	0.7295	**0.8462**	0.7073	<u>0.8278</u>	0.7937
	LPIPS↓	0.2818	0.3374	0.2838	0.5443	**0.2724**	0.4639	0.2796	<u>0.2803</u>
	FID↓	23.1844	19.5815	24.6355	17.7943	17.6825	16.8249	<u>16.2228</u>	**14.2049**
	DISTS↓	0.1464	0.1766	0.1461	0.2715	<u>0.1435</u>	0.2107	0.1782	**0.1429**
	MUSIQ↑	**35.2563**	31.7808	34.6197	26.4684	33.1994	33.3549	30.5885	<u>34.6331</u>
	NIQE↓	4.7146	4.2176	4.5669	7.1426	5.4360	**2.9683**	5.3846	<u>4.1390</u>
	QAlign↑	<u>4.3029</u>	4.2651	4.2824	4.2366	4.2427	4.2492	4.2178	**4.3822**

In addition, for all tasks we invite 15 volunteers to conduct a user study on 40 real-world images. Each volunteer is asked to choose the most preferred one among the outputs of all competing methods, which are presented to the volunteers in random order.

4.2   Effectiveness of the Adjustable Noise Schedule

We first use the task of Real-ISR to discuss the setting and advantages of our proposed ANS. In order to find out how $\bar{\alpha}_a$ affects the performance, we set $\bar{\alpha}_a \in \{0, 0.1, 0.5, 1\}$) and perform experiments on the RealSR test dataset [3]. The curves of PSNR/QAlign versus $\bar{\alpha}_a$ are plotted in Fig. 3. One can see that the PSNR performance increases while the QAlign score decreases as $\bar{\alpha}_a$
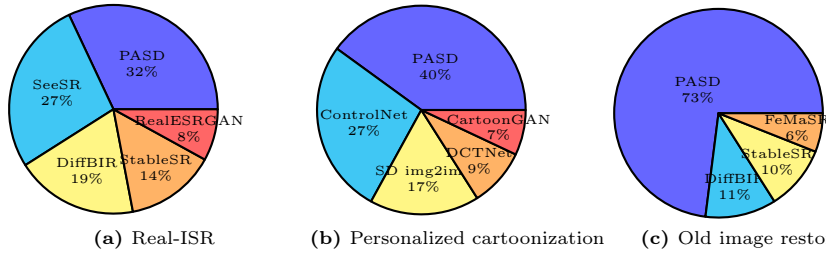
(a) Real-ISR   **(b)** Personalized cartoonization   **(c)** Old image restoration

Fig. 6: User study results of (a) Real-ISR, (b) personalized cartoonization, and (c) old photo restoration tasks.

grows, demonstrating that the proposed $\bar{\alpha}_a$ can be employed to enable flexible perception-fidelity trade-off. Fig. 4 visualizes Real-ISR results with different values of $\bar{\alpha}_a$. We can see that with the increase of $\bar{\alpha}_a$, PASD tends to improve the fidelity while generate less realistic details. In practice, we choose $\bar{\alpha}_a$ from the values of $\bar{\alpha}_n$, where $n \in \{1, 2...N\}$, for convenience. In all of our following experiments, we empirically set $n = 900$, *i.e.*, $\bar{\alpha}_{900} = 0.1189$, to achieve a good balance between fidelity and perception quality.

4.3   Experimental Results

Realistic image super-resolution. We compare the proposed PASD method with two categories of Real-ISR algorithms. The first category is GAN-based methods, including Real-ESRGAN [52], FeMaSR [4], and SwinIR [31]. The second category is diffusion-based models, including ResShift [62], StableSR [50], DiffBIR [34], and SeeSR [57]. The quantitative evaluation results on the test data are presented in Tab. 1, from which we can have the following observations.

First, in term of fidelity measures PSNR/SSIM, the diffusion-based methods are not advantageous over GAN-based methods. This is because diffusion models have higher generative capability and hence may synthesize more perceptually realistic but less faithful details, resulting in lower PSNR/SSIM indices. Second, the diffusion-based methods, especially the proposed PASD, perform better than GAN-based methods in most perception metrics. This conforms to our observation on the visual quality (see Fig. 5) of their Real-ISR output. Third, PASD achieves the best QAlign scores, which is a no-reference image quality assessment index based on large vision-language models, on all the three test datasets.

Fig. 5 visualizes the Real-ISR results of competing methods. It can be seen that our PASD method can generate more realistic details with better visual quality (see the synthesized textures in fur, flowers, leaves, feathers, sea, etc.). Fig. 6a presents the results of subjective user study. PASD receives the most rank-1 votes, confirming its superiority in generating realistic image details. More visual results can be found in the **supplementary material**.

Personalized cartoonization. Similar to the Real-ISR task, we compare the proposed PASD with two categories of stylization algorithms. The first category is GAN-based methods, including CartoonGAN [6], AnimeGAN [5] and DCTNet [37]. We re-train these models with a batch of stylized images gener-

Table 2: The FID, MUSIQ and QAlign indices of different cartoonization models on test data. The best and second best scores are in bold and underscore.

Datasets	Metrics	CartoonGAN	AnimeGAN	DCTNet	InstructPix2Pix	SD Img2img	ControlNet	PASD
	FID↓	53.7454	58.4010	50.6869	39.3259	63.0608	_37.9589_	**37.6698**
FFHQ	MUSIQ↑	71.9848	70.3372	62.6950	72.9510	_74.8963_	74.7521	**75.0221**
	QAlign↑	3.6064	3.4988	3.6812	3.9030	**4.0076**	3.9233	_3.9978_
	FID↓	72.5560	78.4741	81.0789	_71.2098_	75.5553	72.2742	**70.3800**
Flicker2K	MUSIQ↑	71.5768	72.7070	72.2595	_74.1627_	**75.9558**	72.3908	73.1320
	QAlign↑	3.7277	3.6835	3.8483	3.8842	3.9008	**3.9331**	_3.9290_
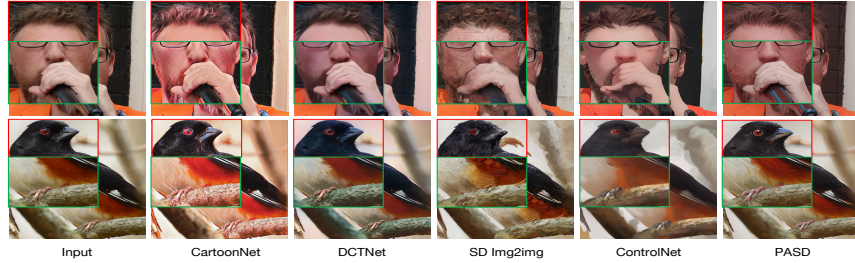
Fig. 7: Stylization (cartoonization) results by different methods on real-world images.
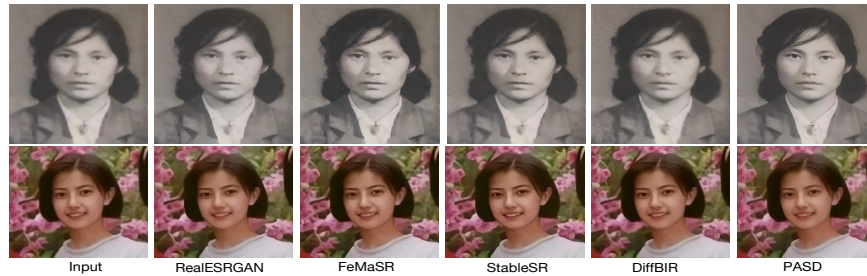
Fig. 8: Old image restoration results by different methods on real-world images.

Table 3: The FID, MUSIQ and QAlign indices of different old image restoration models on self-collected data. The best and second best scores are in bold and underscore.

Metrics	RealESRGAN	FeMaSR	SwinIR	Wan *et al.*	StableSR	DiffBIR	PASD
FID↓	265.5364	263.5393	266.9950	268.3455	264.3178	_262.1837_	**240.2576**
MUSIQ↑	59.4676	_61.2766_	55.3312	32.1470	53.2194	60.3904	**64.4023**
QAlign↑	3.6660	3.6347	3.8232	3.0120	3.8636	_3.9191_	**3.9797**

ated by a personalized diffusion model, *i.e.*, ToonYou [5]. The second category is diffusion-based algorithms, including InstructPix2Pix [2], SD img2img [43] and ControlNet [64]. We replace their base models with the personalized model for fair comparison. Tab. 2 shows the quantitative evaluation results. It can be seen that PASD achieves the best or second best results in most indices.

Fig. 7 shows some cartoonization results. One can see that compared with GAN-based methods, the results of PASD is much cleaner. Compared with the diffusion-based models, PASD can better preserve image details such as human hair. Due to the limited space, we only present results with the style of ToonYou here. Please note that PASD can generate various stylization results by simply
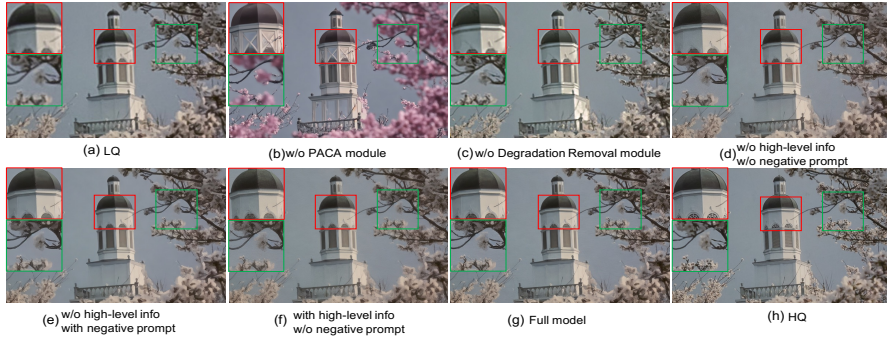
---

[5] https://civitai.com/models/30240/toonyou

Fig. 9: Real-ISR results by different variants of PASD.

Table 4: Quantitative results of different variants of PASD on RealSR test dataset.

Exp.	Degradation Removal	High-level info	Negative prompt	PSNR↑	FID↓	LPIPS↓	Runtime(s)↓
(a)		✓	✓	26.1108	56.7890	0.3822	14.3221
(b)	✓			**27.8731**	53.8988	0.3080	**8.0368**
(c)	✓	✓		27.0901	52.3380	0.2851	_8.7366_
(d)	✓		✓	_27.3781_	_50.2502_	_0.2809_	13.3163
(e)	✓	✓	✓	25.9301	**47.3440**	**0.2806**	14.5889

switching the base diffusion model to a personalized one without any additional training procedure. More stylization results, including the results on image colorization, can be found in the **supplementary materials**.

As in the task of Real-ISR, we also conducted a user study for subjective assessment on the image stylization performance. Fig. 6b shows the results. Clearly, PASD is preferred by most subjects.

Old photo restoration. We compare PASD with Wan *et al.* [49] and several real-world SR methods, including RealESRGAN [52], FeMaSR [4], SwinIR [31], StableSR [50], and DiffBIR [34]. We re-use the PASD model trained for Real-ISR task but replace its base model with an aesthetic one ,*i.e.* majicMIX realistic [6].

Tab. 3 shows the quantitative evaluation results. It can be seen that PASD achieves the best results in all three indices. Fig. 8 visualizes some old photo restoration results. Compared with the competing methods, PASD can better recover vivid image details such as human hair. Fig. 6c presents the results of subjective user study. Clearly, PASD is preferred by the majority of subjects. More visual results can be found in the **supplementary material**.

4.4 Ablation Studies

We perform a series of ablation studies of the proposed PASD network, including the importance of PACA, the role of degradation removal module, and the role of high-level information. We visualize the Real-ISR results of different variants of PASD in Fig. 9, and report the quantitative results and runtime in Tab. 4.

Importance of PACA. We evaluate a variant of PASD by excluding the PACA module from it, *i.e.*, the features $\mathbf{y}$ extracted from ControlNet are simply added to features $\mathbf{x}$. As shown in Fig. 9(b), the output becomes inconsistent

---
[6] https://civitai.com/models/43331/majicmix-realistic

with the LQ input in colors and structures, etc. This verifies the importance of PACA in perceiving pixel-wise local structures.

Role of degradation removal module. To evaluate the effect of degradation removal module, we remove the "toRGB" modules as well as the pyramid $\mathcal{L}_{DR}$ loss during model training. As can be seen in Fig. 9(c) and Tab. 4, removing the degradation removal module leads to dirty outputs and worse PSNR, FID and LPIPS indices.

Role of high-level information. The high-level information and negative prompt are optional but very useful for PASD. We simply replace them with null-text prompt to evaluate their effects. As shown in Fig. 9(d), replacing both high-level information and negative prompt with null-text prompt results in dirty outputs with less realistic details, which is also verified by the worse FID and LPIPS indices in Tab. 4. Abandoning high-level information leads to over-smoothed results, as illustrated in Fig. 9(e). The output can become dirty without negative prompt (see Fig. 9(f)). Our full model takes advantages of both high-level information and negative prompt, and achieves a good balance between clean-smooth and detailed-dirty outputs (see Fig. 9(g) and the best FID score in Tab. 4).

Runtime analysis on different modules. The runtime is reported as the average over 10 runs to process a 256×256 image on a NVIDIA Tesla 32G-V100 GPU. We use the DDIM [47] sampler for 20 steps. By comparing Exps. (a) and (e) in Tab. 4, one can see that the degradation removal module has little effect on the runtime. Without the negative prompt module, the runtime nearly cuts in half because the classifier-free guidance can be removed (see Exps.(c) and (e)). Finally, the high-level information module only increase a little the runtime (see Exps.(d) and (e)).

5 Conclusion and Limitation

We proposed a pixel-aware diffusion network, namely PASD, for realistic image restoration and personalized stylization. By introducing a pixel-aware cross attention module, PASD succeeded in perceiving image local structures in pixel-level and achieved robust and perceptually realistic Real-ISR results. An adjustable noise schedule was also proposed, which helped PASD to achieve flexible perception-fidelity trade-off during the inference stage. By replacing the base model to a personalized one, PASD could produce diverse stylization results with highly consistent semantic contents with the input. The proposed PASD was simple to implement, and our extensive experiments demonstrated its effectiveness and flexibility across different tasks, showing its great potentials for handling complex image restoration and stylization tasks.

Though PASD can achieve pixel-level enhancement, it still suffers from the balance between fidelity and perception. In addition, it may fail to reproduce faithful details when the input image is heavily degraded or the semantic information is inaccurate. A more robust degradation estimation module can be designed, and more precise semantic information can be extracted to further improve the performance of PASD, which will be considered in our future work.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPRW (2017) 9
2. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023) 2, 4, 12
3. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV (2019) 3, 9, 10
4. Chen, C., Shi, X., Qin, Y., Li, X., Han, X., Yang, T., Guo, S.: Real-world blind super-resolution via feature matching with implicit high-resolution priors. In: ACM MM (2022) 2, 3, 4, 11, 13
5. Chen, J., Liu, G., Chen, X.: Animegan: A novel lightweight gan for photo animation. In: CVPR (2020) 4, 8, 11
6. Chen, Y., Lai, Y.K., Liu, Y.J.: Cartoongan: Generative adversarial networks for photo cartoonization. In: CVPR (2018) 2, 4, 8, 11
7. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE TPAMI **44** (2022) 9
8. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. In: ECCV (2014) 1, 3
9. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. In: Arxiv (2015) 4
10. Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S.S., Shah, A., Yin, X., Parikh, D., Misra, I.: Emu video: Factorizing text-to-video generation by explicit image conditioning. In: ArXiv (2023) 2, 6
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: NeurIPS. pp. 2672–2680 (2014) 2, 3
12. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: ArXiv (2023) 8
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770—-778 (2016) 7
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017) 9
15. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition video generation with diffusion models. In: Arxiv (2022) 2, 4
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851 (2020) 2, 3, 4, 7
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: Arxiv (2021) 7
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022) 8
19. Johnson, J., Alahi, A., Li, F.F.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016) 4
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019) 4, 9
21. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. In: NIPS (2022) 2, 3

22. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR (2023) 4

23. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV (2021) 9

24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Arxiv (2015) 9

25. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023) 4

26. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: CVPR (2017) 3

27. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017) 2, 3

28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023) 7

29. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: NeurIPS (2017) 4

30. Liang, J., Zeng, H., Zhang, L.: Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In: CVPR (2022) 3

31. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: Swinir: Image restoration using swin transformer. ArXiv (2021) 11, 13

32. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPRW (2017) 1, 3

33. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. In: ArXiv (2023) 2, 6

34. Lin, X., He, J., Chen, Z., Lyu, Z., Fei, B., Dai, B., Ouyang, W., Qiao, Y., Dong, C.: Diffbir: Towards blind image restoration with generative diffusion prior. In: ArXiv (2023) 11, 13

35. Liu, H., Xing, J., Xie, M., Li, C., Wong, T.T.: Improved diffusion-based image colorization via piggybacked models. In: ArXiv (2023) 2, 4

36. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In: NeurIPS (2022) 4

37. Men, Y., Yao, Y., Cui, M., Lian, Z., Xie, X.: Dct-net: Domain-calibrated translation for portrait stylization. In: ACM TOG (2022) 4, 11

38. Meng, C., Rombach, R., Gao, R., Kingma, D.P., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: CVPR (2023) 4

39. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013) 9

40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 7

41. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. In: Arxiv (2022) 2, 4

42. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016) 7

43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2021) 2, 3, 4, 6, 7, 8, 12

44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 4
45. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023) 2, 4, 8
46. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Tim Salimans, J.H., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Arxiv (2022) 2, 4
47. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021) 4, 14
48. Timofte, R., Agustsson, E., Gool, L.V., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPRW. pp. 114–125 (2017) 9
49. Wan, Z., Zhang, B., Chen, D., Zhang, P., Chen, D., Liao, J., Wen, F.: Bringing old photos back to life. In: CVPR (2021) 2, 4, 13
50. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. In: Arxiv (2023) 2, 3, 4, 6, 7, 11, 13
51. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: CVPR (2021) 3, 4
52. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCVW (2021) 3, 4, 9, 11, 13
53. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR (2018) 9
54. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: ECCVW (2018) 2, 3
55. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: ECCV (2020) 9
56. Wu, H., Zhang, Z., Zhang, W., Chen, C., Li, C., Liao, L., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In: ArXiv (2023) 9
57. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: CVPR (2023) 3, 7, 11
58. Yang, T., Ren, P., Xie, X., , Zhang, L.: Gan prior embedded network for blind face restoration in the wild. In: CVPR (2021) 3, 4
59. Yang, T., Ren, P., Xie, X., Hua, X., Zhang, L.: Beyond a video frame interpolator: A space decoupled learning approach to continuous image transition. In: ECCVW (2022) 4
60. Yang, T., Ren, P., Xie, X., Zhang, L.: Synthesizing realistic image restoration training pairs: A diffusion approach. In: Arxiv (2023) 3
61. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: CVPR (2018) 4
62. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. In: NeurIPS (2023) 11
63. Zhang, K., Liang, J., Gool, L.V., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV. pp. 4791–4800 (2021) 3
64. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: NeurIPS (2023) 2, 3, 4, 6, 12
65. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 9

66. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: CVPR (2023) 4
67. Zhou, S., Chan, K.C., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. In: NeurIPS (2022) 2
68. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017) 4, 8