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

Tao Yang[1*], **Peiran Ren**[1], **Xuansong Xie**[1], **Lei Zhang**[2]

[1]DAMO Academy, Alibaba Group
[1]Department of Computing, The Hong Kong Polytechnic University

Abstract

Realistic image super-resolution (Real-ISR) aims to reproduce perceptually realistic image details from a low-quality input. The commonly used adversarial training based Real-ISR methods often introduce unnatural visual artifacts and fail to generate realistic textures for natural scene images. The recently developed generative stable diffusion models provide a potential solution to Real-ISR with pre-learned strong image priors. However, the existing methods along this line either fail to keep faithful pixel-wise image structures or resort to extra skipped connections to reproduce details, which requires additional training in image space and limits their extension to other related 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 as well as personalized stylization. In specific, 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. By simply replacing the base diffusion model with a personalized one, our method can generate diverse stylized images without the need to collect pairwise training data. PASD can be easily integrated into existing diffusion models such as Stable Diffusion. Experiments on Real-ISR and personalized stylization demonstrate the effectiveness of our proposed approach. The source code and models can be found at https://github.com/yangxy/PASD.

Introduction

Images often suffer from a mixture of complex degradations, such as low resolution, blur, noise, etc., in the acquisition process. While image restoration methods (Yang et al. 2008) have achieved significant progress, especially in the era of deep learning (Dong et al. 2014; Lim et al. 2017), 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) (Goodfellow et al. 2014) and the adversarial training strategy have been widely used for Real-ISR (Ledig et al. 2017; Wang et al. 2018b) and achieved promising results. However, GAN-based Real-ISR methods tend to generate unpleasant visual artifacts. Some methods (Zhou et al. 2022) learn discrete codebooks to reduce the uncertainty in GAN-based face restoration, and some methods (Chen et al. 2022) restore images by matching distorted features to their distortion-free counterparts. Nonetheless, the above methods are limited in reproducing rich and realistic image details.

Recently, denoising diffusion probabilistic models (DDPMs) have shown outstanding performance in tasks of image generation (Ho, Jain, and Abbeel 2020), image-to-image translation (Saharia et al. 2022a), etc. DDPM is a strong alternative to GAN in many downstream tasks (Dhariwal and Nichol 2021; Rombach et al. 2022) due to its powerful capability in approximating diverse and complicated distributions. Denoising diffusion implicit model (DDIM) (Song, Meng, and Ermon 2021) was proposed to accelerate the sampling speed of DDPM. The DDPM/DDIM based pre-trained text-to-image (T2I) and text-to-video (T2V) models (Rombach et al. 2021; Ramesh et al. 2022; Saharia et al. 2022b; Ho et al. 2022) have been popularly used in numerous downstream tasks, including personalized image generation (Ruiz et al. 2023; Kumari et al. 2023), image editing (Hertz et al. 2022; Brooks, Holynski, and Efros 2023; Kawar et al. 2023), image inpainting (Yang et al. 2022a) and conditional image synthesis (Zhang and Agrawala 2023). They have also been adopted to solve image restoration tasks. A denoising diffusion restoration model (DDRM) is proposed in (Kawar et al. 2022) to solve inverse problem by taking advantage of a pre-trained denoising diffusion generative model. However, a liner image degradation model is assumed to be known in DDRM, limiting its application to more practical scenarios such as Real-ISR.

Considering that the pre-trained T2I models such as Stable Diffusion (SD) (Rombach et al. 2021) can generate high-quality natural images, Zhang and Agrawala (Zhang and Agrawala 2023) 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). Qin *et al*. (Qin et al. 2023) extended ControlNet by introducing UniControl to

*yangtao9009@gmail.com

Figure 1: An input LQ image (left) and the Real-ISR output (right) by ControlNet. One can see clearly the content inconsistency between them.

enable more diverse visual conditions. Liu *et al*. (Liu et al. 2023) and Wang *et al*. (Wang et al. 2023) 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 investigate the problem of Real-ISR with pre-trained T2I models such as SD, targeting at reconstructing photo-realistic structures and textures. Our idea is to introduce pixel-aware conditional control into the diffusion process so that robust and perceptually realistic Real-ISR results 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. Last but not least, the proposed pixel-aware stable diffusion (PASD) network can perform personalized stylization tasks by simply shifting the base model to a personalized one.

Related Work

Realistic Image Super-Resolution. Though deep learning based image super-resolution (Dong et al. 2014; Lim et al. 2017) 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 both fidelity and perception objectives. The GAN (Goodfellow et al. 2014) network and its adversarial training strategies are widely in Real-ISR (Ledig et al. 2017; Wang et al. 2018b). 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. However, adversarial training is unstable, and the GAN-based Real-ISR methods often bring unnatural visual artifacts. Liang *et al*. (Liang, Zeng, and Zhang 2022) proposed a locally discriminative learning approach to suppress the GAN-generated artifacts, yet it is difficult to introduce more details. Recently, inspired by the success of generative priors in face restoration tasks (Yang et al. 2021; Wang et al. 2021a), some works exploit the priors extracted from VQGAN (Esser, Rombach, and Ommer 2021), diffusion model (Ho, Jain, and Abbeel 2020) and pre-trained T2I model (Rombach et al. 2021) to solve the Real-ISR problems and led to interesting results (Chen et al. 2022; Kawar et al. 2022; Wang et al. 2023).

In the early study, bicubic downsampling or some simple degradations (Dong et al. 2014; Lai et al. 2017; Gu et al. 2019) are used to simulate the LQ-HQ training pairs. Zhang *et al*. (Zhang et al. 2021) and Wang *et al*. (Wang et al. 2021b) later modeled complex degradations by shuffling degradation types and using a high-order process, respectively. Cai *et al*. (Cai et al. 2019) collected a real-world dataset with paired LQ-HQ images by zooming camera lens. Lugmayr *et al*. (Lugmayr, Danelljan, and Timofte 2019) learned a distribution mapping network with unpaired data.

Personalized Stylization. Inspired by the powerful learning capacity of deep neural networks, Gatys *et al*. (Gatys, Ecker, and Bethge 2015) presented an optimization based method to transfer the style of a given artwork to a content image. This work was extended and developed by many following researchers (Johnson, Alahi, and Li 2016; Li et al. 2017; Zhang et al. 2023). However, these methods require an extra image as style input. This problem can be alleviated by resorting to an image-to-image framework (Zhu et al. 2017; Chen, Lai, and Liu 2018; Chen, Liu, and Chen 2020). Due to the lack of pairwise training data, some works (Men et al. 2022; Yang et al. 2022b) focus on portrait stylization with the help of StyleGAN (Karras, Laine, and Aila 2019). With the rapid development of SD models, some works (Brooks, Holynski, and Efros 2023; Zhang and Agrawala 2023) 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. In addition, these methods 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*. (Ruiz et al. 2023) and Kumari *et al*. (Kumari et al. 2023) proposed personalized stylization approaches for T2I diffusion models.

Diffusion Probabilistic Models. The seminal work of DDPM (Ho, Jain, and Abbeel 2020) demonstrates promising capability in generating high quality natural images. Considering that DDPMs require hundreds of sampling steps in the denoising process, Song *et al*. (Song, Meng, and Ermon 2021) proposed DDIM to accelerate the sampling speed. Following works extend DDPM/DDIM by adapting high-order solvers (Lu et al. 2022) and distillations (Meng et al. 2023). Rombach *et al*. (Rombach et al. 2022) extended DDPM to latent space and demonstrated impressive results with less computational resources. This work sparks the prosperity of large pre-trained T2I and T2V diffusion models such as SD (Rombach et al. 2021), Imagen (Ho et al. 2022). It has been demonstrated that T2I diffusion priors are

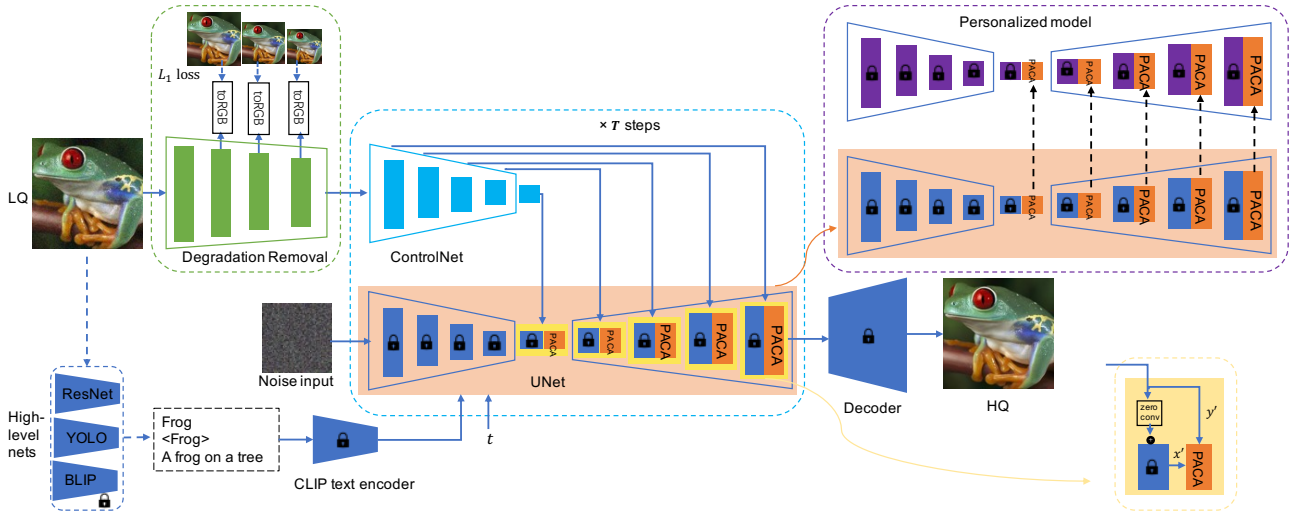

Figure 2: Architecture of the proposed pixel-aware stable diffusion (PASD) network.

more powerful than GAN priors in handling diverse natural images (Rombach et al. 2021; Ramesh et al. 2022; Saharia et al. 2022b). Kawar *et al*. (Kawar et al. 2023) applied complex text-guided semantic edits to real images. ControlNet (Zhang and Agrawala 2023) enables conditional inputs, such as edge maps, segmentation maps, keypoints, etc., to T2I models. Liu *et al*. (Liu et al. 2023) and Wang *et al*. (Wang et al. 2023) respectively utilized generative diffusion priors to image colorization and super-resolution.

Pixel-Aware Stable Diffusion Network

Our method is based on generative diffusion priors. In particular, we utilize the powerful pre-trained SD (Rombach et al. 2021) model, while alternative diffusion models such as DALLE2 (Ramesh et al. 2022) and Imagen (Saharia et al. 2022b) can also be adopted. The architecture of our pixel-aware SD (PASD) network is depicted in Fig. 2. One can see that in addition to the pre-trained SD model, PASD has three 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, and a pixel-aware cross-attention (PACA) module to perform pixel-level guidance for diffusion. While our PASD is mainly designed for the Real-ISR task, it can be readily used for personalized stylization by simply switching the base diffusion model to a personalized one.

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: $\mathcal{L}_{\mathcal{DR}} = \sum_s ||\mathbf{I}^s_{hq} - \mathbf{I}^s_{sr}||_1$, where $\mathbf{I}^s_{hq}$ and $\mathbf{I}^s_{sr}$ represent the HQ ground-truth and ISR output at scale s. Note that L_{DR} is only required in the Real-ISR task.

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 (Zhang and Agrawala 2023) 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 areas such as leaves, flowers, moustaches and glasses.

To address this problem, some methods employ a skipped connection outside the U-Net (Wang et al. 2023) 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

Table 1: The PSNR, SSIM, FID, CLIP-FID, LPIPS, DISTS and MUSIQ indices of different Real-ISR models on both synthesized (DIV2K) and real-world (RealSR and DRealSR) test datasets.

Datasets	Metrics	BSRGAN	Real-ESRGAN	LDL	FeMaSR	SwinIR-GAN	LDM	SD Upscaler	StableSR	PASD
DIV2K valid	PSNR↑	**23.4105**	23.1465	22.7449	21.8552	22.6459	21.4763	21.2083	20.8778	21.8530
	SSIM↑	0.6078	**0.6212**	0.6172	0.5426	0.6051	0.5572	0.5466	0.5256	0.5215
	FID↓	92.0801	85.8453	91.0641	83.7743	85.2116	93.0184	92.1536	69.5598	**68.7851**
	CLIP-FID↓	13.3123	13.1279	14.1164	10.6802	13.0361	14.0566	12.9387	**9.9444**	10.3318
	LPIPS↓	0.4263	**0.4030**	0.4160	0.4100	0.4055	0.4497	0.4302	0.4376	0.4304
	DISTS↓	0.1787	0.1691	0.1770	0.1711	0.1692	0.1844	0.1747	0.1761	**0.1680**
	MUSIQ↑	55.9025	57.2905	56.3724	59.0301	56.8953	40.2183	63.4425	48.1331	**65.8658**
RealSR	PSNR↑	**26.7457**	26.0231	25.6422	25.4993	26.2911	24.6439	25.871	26.0326	24.8065
	SSIM↑	0.7767	0.7742	0.7696	0.7518	**0.7816**	0.6468	0.6430	0.7714	0.6947
	FID↓	62.7142	66.5168	72.3305	62.8607	63.8406	61.8616	67.2451	70.2201	**57.7270**
	CLIP-FID↓	7.449	7.8258	9.1593	**6.3895**	8.2282	11.9536	9.2767	8.3158	7.2620
	LPIPS↓	0.2674	0.2709	0.2761	0.2961	**0.2591**	0.3967	0.3079	0.2617	0.2926
	DISTS↓	0.1354	0.1431	0.1504	0.1328	0.1345	0.1656	0.1603	0.1382	**0.1263**
	MUSIQ↑	63.9907	62.6299	63.4630	61.4377	65.0789	51.4555	62.7663	67.2719	**67.9338**
DRealSR	PSNR↑	**28.3408**	27.9152	27.7215	26.5862	27.8456	24.3063	26.4961	27.5304	25.0244
	SSIM↑	0.8205	0.8247	**0.8333**	0.7683	0.8206	0.6556	0.7698	0.8303	0.6943
	FID↓	19.7078	23.1816	25.6747	19.5665	24.6497	27.0092	24.2128	**18.2130**	19.9174
	CLIP-FID↓	3.6615	3.6877	3.7812	**3.3268**	3.6527	7.2814	5.4362	3.4574	3.4227
	LPIPS↓	0.2929	0.2818	0.2785	0.3374	0.2838	0.4348	0.4055	0.2750	**0.2739**
	DISTS↓	0.0870	0.0901	0.0962	0.0994	0.0925	0.1259	0.1024	**0.0853**	0.0958
	MUSIQ↑	55.0083	54.8643	54.9784	56.3682	55.3148	54.4335	50.0446	51.9831	**63.2654**

$\mathbf{x}$ 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:

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

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 Fig. 8, with the help of PACA, the output of our PASD network can reproduce realistic and faithful image structures and textures in pixel-level.

High-Level Information

Our method is based on the pre-trained SD model where text is used as the input, while Real-ISR typically takes the LQ image as the input. Though some SD-based Real-ISR methods (Wang et al. 2023) employ the null-text prompt, it has been demonstrated that content-related captions could improve the synthesis results (Rombach et al. 2021). As shown in Fig. 2, we employ the pre-trained ResNet (He et al. 2016), YOLO (Redmon et al. 2016) and BLIP (Li et al. 2023) networks to extract image classification, object detection and image caption information from the LQ input, and employ the CLIP (Radford et al. 2021) encoder to convert the text information into image-level features, providing additional semantic signal to control the diffusion process.

The classifier-free guidance (Ho and Salimans 2021) technique is adopted here:

$$\tilde{\epsilon}(\mathbf{z}_t,\mathbf{c}) = \epsilon(\mathbf{z}_t,\mathbf{c}) + \omega\epsilon(\mathbf{z}_t,\mathbf{c}_{neg}), \quad (3)$$

where $\tilde{\epsilon}(\mathbf{z}_t,\mathbf{c})$ and $\epsilon(\mathbf{z}_t,\mathbf{c}_{neg})$ are conditional and unconditional ϵ-predictions (Ho, Jain, and Abbeel 2020). $\mathbf{c}$ 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", *etc*., as negative prompts. The negative prompts 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.

Personalized Stylization

Our method is primarily designed for Real-ISR, which can be considered as an image-to-image translation problem (*i.e*., from LQ images to HQ images). Thanks to the proposed PACA, the translation can be done in pixel-level. Inspired by the recent work of AnimateDiff (Guo et al. 2023), we can replace the base model of our PASD network with a personalized model during inference (as illustrated in the top-right corner of Fig. 2) so that it can produce stylization results. Unlike previous methods (Zhu et al. 2017; Chen, Lai, and Liu 2018; Chen, Liu, and Chen 2020) 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 [1], one can easily generate various stylized results with our PASD method.

[1] https://civitai.com/, https://huggingface.co/

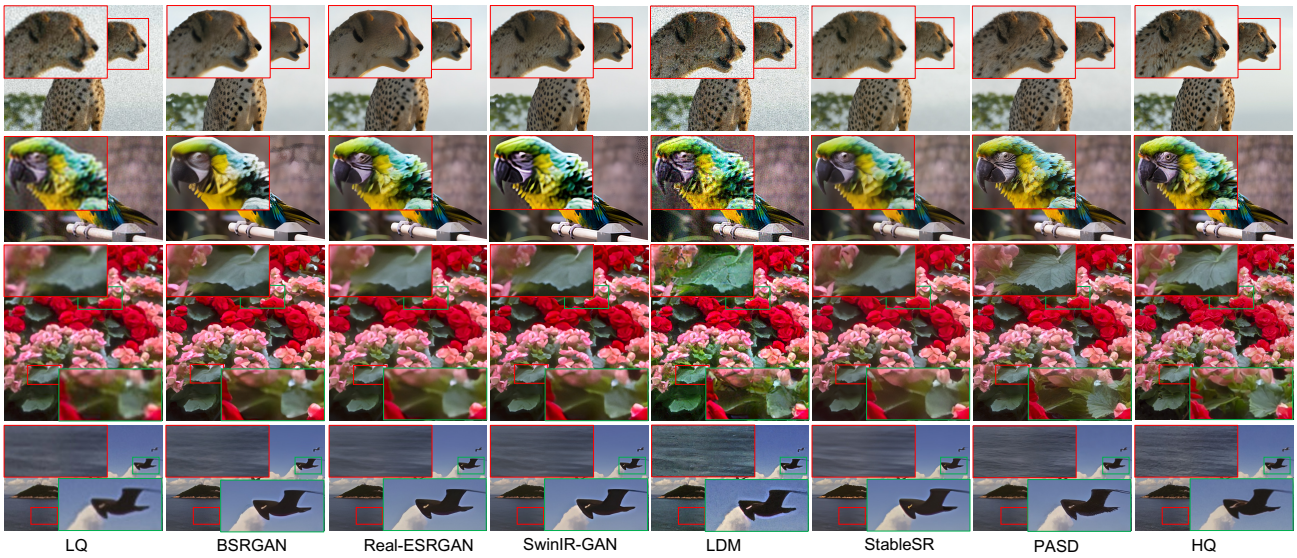

Figure 3: Realistic image super-resolution results by different methods on test images.

Table 2: The FID, CLIP-FID and MUSIQ indices of different stylization models on test data.

Datasets	Metrics	CartoonGAN	AnimeGAN	DCT-Net	Instruct-Pix2Pix	SD image2image	ControlNet	PASD
FFHQ	FID↓	110.6474	121.8184	122.6714	108.0571	130.0608	**98.9197**	106.4948
	CLIP-FID↓	53.7454	58.4010	50.6869	39.3259	63.0608	37.9589	**37.6698**
	MUSIQ↑	71.9848	70.3372	62.6950	72.9510	74.8963	74.7521	**75.0221**
Flicker2K	FID↓	168.2210	177.2038	177.8979	171.2087	163.3093	**153.4681**	161.5938
	CLIP-FID↓	72.5560	78.4741	81.0789	71.2098	75.5553	72.2742	**70.3800**
	MUSIQ↑	71.5768	72.7070	72.2595	74.1627	**75.9558**	72.3908	73.1320

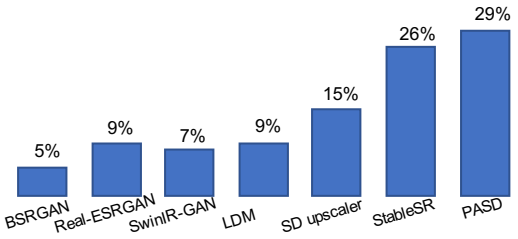

Figure 4: User study results of different Real-ISR methods.

Training Strategy

Given an HQ image, we first obtain its latent representation $\mathbf{z}_0$. The diffusion algorithm progressively adds noise to the latent image and yields a noisy latent $\mathbf{z}_t$, where t is the diffusion step and is randomly sampled. Given a number of conditions including 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 of the diffusion model is:

$$\mathcal{L}_{\mathcal{DF}} = \mathbb{E}_{\mathbf{z}_0,t,\mathbf{c},\mathbf{I}_{lq},\epsilon\sim\mathcal{N}(0,1)}\left[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}, \mathbf{I}_{lq})||_2^2\right]. \quad (4)$$

During the training of Real-ISR models, we jointly update the degradation removal module. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\mathcal{DF}} + \alpha\mathcal{L}_{\mathcal{DR}} \quad (5)$$

where α is a balancing parameter. We empirically set $\alpha = 1$.

Fine-tuning all the parameters in the pre-trained SD model would be very expensive. As in previous works (Zhang and Agrawala 2023; Wang et al. 2023), we freeze all the parameters in SD, and only train the newly added modules, including 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.

Experiments

Experiment Setup

We adopt the Adam optimizer (Kingma and Ba 2015) 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 dataset. We train PASD on DIV2K (Timofte et al. 2017), Flickr2K (Agustsson and Timofte 2017), OST (Wang et al. 2018a), and the first 10000 face images from FFHQ (Karras, Laine, and Aila 2019). We employ the degradation pipeline of Real-ESRGAN (Wang et al. 2021b) to synthesize LQ-HQ training pairs.

In the task of Real-ISR, we evaluate our approach on both synthetic and real-world datasets. The synthetic dataset is generated from the DIV2K validation set following the

Figure 5: Stylization (cartoonization) results by different methods on real-world images.

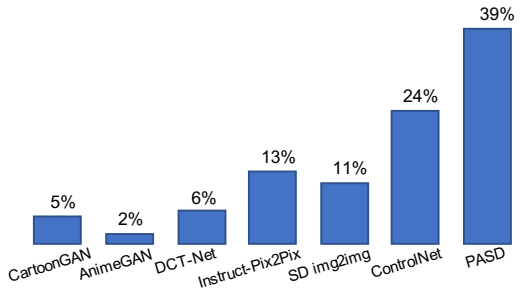

Figure 6: User study results of different stylization methods.

degradation pipeline of Real-ESRGAN (Wang et al. 2021b). For real-world test dataset, we use two benchmarks, *i.e.*, RealSR (Cai et al. 2019) and DRealSR (Wei et al. 2020), for evaluation. As for the task of personalized stylization, we conduct comparisons on the first 100 face images from FFHQ as well as the first 100 images from Flicker2K.

Evaluation metrics. For quantitative evaluation of Real-ISR models, we employ the widely used perceptual metrics, including FID (Heusel et al. 2017), LPIPS (Zhang et al. 2018), DISTS (Ding et al. 2022) and MUSIQ (Ke et al. 2021), to compare the competing Real-ISR models. FID (Heusel et al. 2017) is widely used to evaluate the image perceptual quality in image generation tasks. It extracts image features using Inception V3 (Szegedy et al. 2015) trained on ImageNet. We also adopt a variant of FID, *i.e.*, CLIP-FID (Kynkäänniemi et al. 2023), which uses the CLIP (Radford et al. 2021) features, in evaluation. 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. Since ground-truth images are unavailable in personalized stylization tasks, we employ FID, CLIP-FID and MUSIQ in evaluation.

For both Real-ISR and stylization tasks, we invite 20 volunteers to conduct a user study on 40 real-world images. Each volunteer is asked to choose the most preferred one among all the outputs of competing methods, which are presented to the volunteers in random order.

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 BSRGAN (Zhang et al. 2021), Real-ESRGAN (Wang et al. 2021b), SwinIR-GAN (Liang et al. 2021), LDL (Liang, Zeng, and Zhang 2022) and FeMaSR (Chen et al. 2022). The second category is diffusion-based models, including LDM (Rombach et al. 2021), SD upscaler (Rombach et al. 2021) and StableSR (Wang et al. 2023). 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. 8) of their Real-ISR output. Third, our proposed PASD achieves the best MUSIQ scores, which is a no-reference image quality assessment index, on all the three test datasets.

Fig. 8 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, etc.). Fig. 4 presents the results of subjective user study. The proposed 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 Stylization. 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 (Chen, Lai, and Liu 2018), AnimeGAN (Chen, Liu, and Chen 2020) and DCT-Net (Men et al. 2022). We re-train these models with a batch of stylized images generated by a personalized diffusion model,

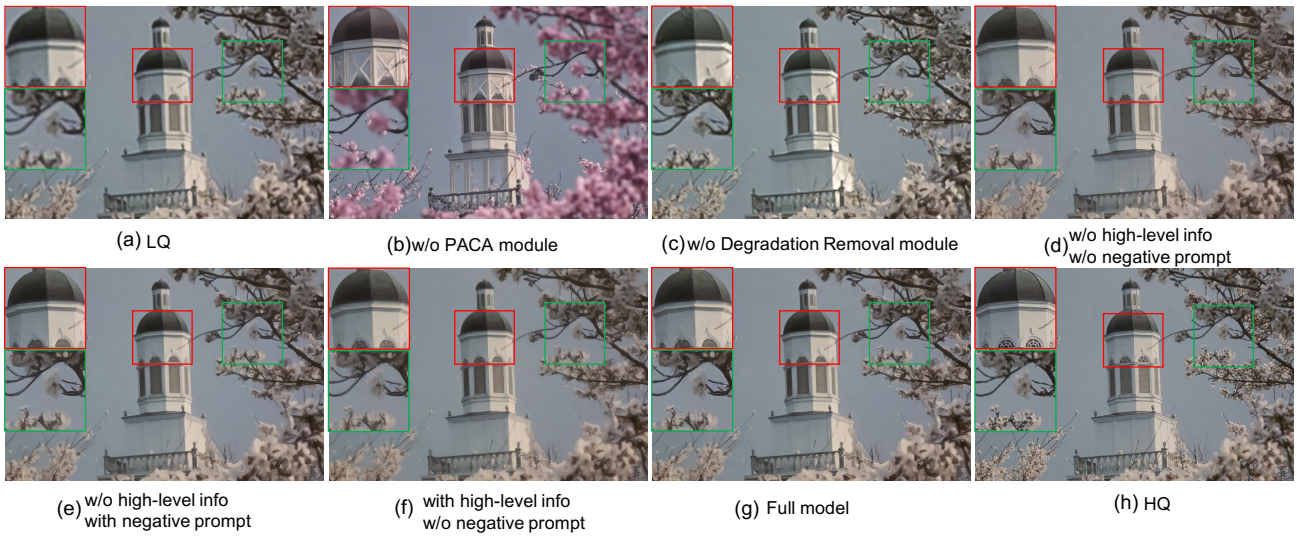

Figure 7: Real-ISR results by different variants of PASD.

Table 3: Quantitative results of different variants of PASD on RealSR test dataset.

Exp.	Degradation Removal	High-level info	Negative prompt	PSNR↑	FID↓	LPIPS↓
(a)		✓	✓	23.0590	24.9677	0.4722
(b)	✓			**26.0218**	7.8604	0.3958
(c)	✓	✓		25.3468	9.1088	0.3770
(d)	✓		✓	24.7121	8.5100	0.3590
PASD	✓	✓	✓	24.8065	**7.2620**	**0.2926**

i.e., ToonYou [2]. The second category is diffusion-based algorithms, including InstructPix2Pix (Brooks, Holynski, and Efros 2023), SD img2img (Rombach et al. 2021) and ControlNet (Zhang and Agrawala 2023). We replace their base models with the personalized model for fair comparison.

Tab. 2 shows the quantitative evaluation results. It can be seen that our PASD method achieves the best or second best results in most indices. Fig. 9 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 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 conduct a user study for subjective assessment for stylization. Fig. 6 shows the results. Clearly, PASD is preferred by most subjects.

Ablation Studies

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. 7(b), the output becomes inconsistent 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. 7(c) and Tab. 3, 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. 7(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. 3. Abandoning high-level information leads to over-smoothed results, as illustrated in Fig. 7(e). The output can become dirty without negative prompt (see Fig. 7(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. 7(g)).

Conclusion

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. By replacing the base model to a personalized one, PASD could also 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.

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

Appendix

More Real-ISR Results

In Fig. 8, we show more visual comparisons between our method with state-of-the-art Real-ISR methods, including Real-ESRGAN (Wang et al. 2021b), SwinIR-GAN (Liang et al. 2021), LDM (Rombach et al. 2021) and StableSR (Wang et al. 2023). Similar conclusions to the main paper can be made. With the help of PACA module, our PASD can provide pixel-level guidance on the image generation, reproducing more realistic fine details and less visual artifacts.

Various Stylization Results

As mentioned in the main paper, by simply switching the base diffusion model to a personalized one, our proposed PASD can do various stylization tasks without any additional training procedure. In the main paper, we have provided the results by using the ToonYou style. In Fig 9 of this supplementary file, we show more types of stylization results by using the personalized base models of Disney 3D, Oil painting and Shinkai. One can see that our PASD method can keep very well the pixel-wise image details while performing style transfer.

Image Colorization

Our PASD can serve as a generic solution for various pixel-wise image-to-image tasks. In addition to Real-ISR and personalized stylization, we also apply it to image colorization and show the results in this supplementary file.

Figure 10 shows the qualitative comparisons between PASD and the state-of-the-art image colorization methods, including DeOldify (Anti 2019), BigColor (Kim et al. 2022), CT2 (Weng et al. 2022) and DDColor (Kang et al. 2023). One can see that our PASD generates more photo-realistic and vivid colorization results. In particular, it significantly alleviates the color bleeding effect, which often happens in the compared methods.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*.

Anti, J. 2019. jantic/deoldify: A deep learning based project for colorizing and restoring old image (and videos!). Https://github.com/jantic/DeOldify.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*.

Chen, C.; Shi, X.; Qin, Y.; Li, X.; Han, X.; Yang, T.; and Guo, S. 2022. Real-World Blind Super-Resolution via Feature Matching with Implicit High-Resolution Priors. In *ACM MM*.

Chen, J.; Liu, G.; and Chen, X. 2020. AnimeGAN: A Novel Lightweight GAN for Photo Animation. In *CVPR*.

Chen, Y.; Lai, Y.-K.; and Liu, Y.-J. 2018. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. In *CVPR*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Arxiv*.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2022. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE TPAMI*, 44.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Image Super-Resolution Using Deep Convolutional Networks. In *ECCV*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A Neural Algorithm of Artistic Style. In *Arxiv*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks. In *NeurIPS*, 2672–2680.

Gu, J.; Lu, H.; Zuo, W.; and Dong, C. 2019. Blind super-resolution with iterative kernel correction. In *CVPR*.

Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; and Dai, B. 2023. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *ArXiv*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770—-778.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. In *ArXiv*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; and Salimans, T. 2022. IMAGEN VIDEO: HIGH DEFINITION VIDEO GENERATION WITH DIFFUSION MODELS. In *Arxiv*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *Arxiv*.

Johnson, J.; Alahi, A.; and Li, F.-F. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

Kang, X.; Yang, T.; Ouyang, W.; Ren, P.; Li, L.; and Xie, X. 2023. DDColor: Towards Photo-Realistic and Semantic-Aware Image Colorization via Dual Decoders. In *ICCV*.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising Diffusion Restoration Models. In *NIPS*.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *CVPR*.

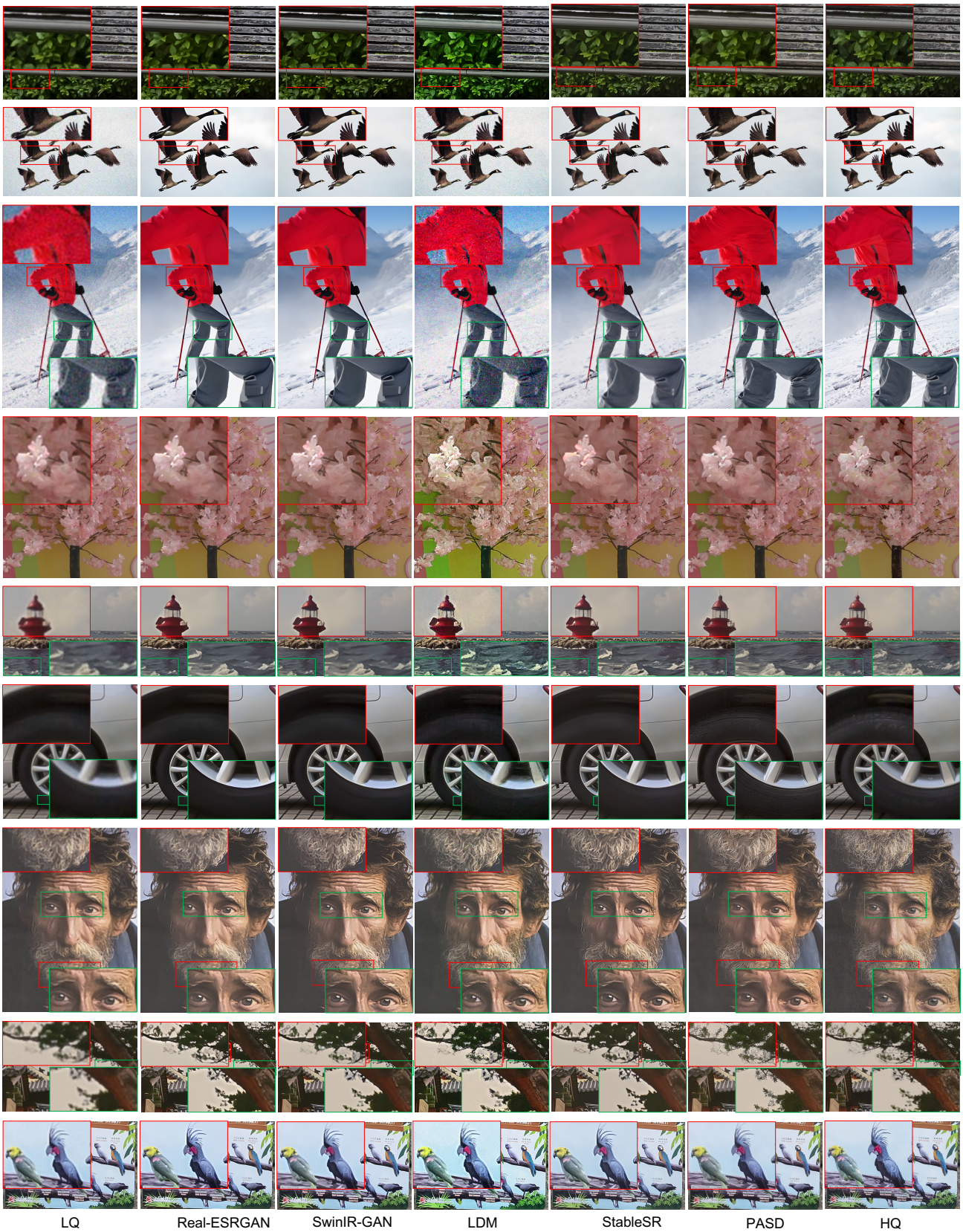

Figure 8: Realistic image super-resolution results by different methods. Please zoom-in for better comparison.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *ICCV*.

Kim, G.; Kang, K.; Kim, S.; Lee, H.; Kim, S.; Kim, J.; Baek, S.-H.; and Cho, S. 2022. BigColor: Colorization using a Generative Color Prior for Natural Images. In *ECCV*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Arxiv*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *CVPR*.

Kynkäänniemi, T.; Karras, T.; Aittala, M.; Aila, T.; and Lehtinen, J. 2023. The Role of ImageNet Classes in Fréchet Inception Distance. In *ICLR*.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *CVPR*.

Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *CVPR*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal Style Transfer via Feature Transforms. In *NeurIPS*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Gool, L. V.; and Timofte, R. 2021. SwinIR: Image Restoration Using Swin Transformer. *ArXiv*.

Liang, J.; Zeng, H.; and Zhang, L. 2022. Details or Artifacts: A Locally Discriminative Learning Approach to Realistic Image Super-Resolution. In *CVPR*.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Lee, K. M. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *CVPRW*.

Liu, H.; Xing, J.; Xie, M.; Li, C.; and Wong, T.-T. 2023. Improved Diffusion-based Image Colorization via Piggybacked Models. In *ArXiv*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In *NeurIPS*.

Lugmayr, A.; Danelljan, M.; and Timofte, R. 2019. Unsupervised Learning for Real-World Super-Resolution. In *ICCVW*.

Men, Y.; Yao, Y.; Cui, M.; Lian, Z.; and Xie, X. 2022. DCT-Net: Domain-Calibrated Translation for Portrait Stylization. In *ACM TOG*.

Meng, C.; Rombach, R.; Gao, R.; Kingma, D. P.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On Distillation of Guided Diffusion Models. In *CVPR*.

Qin, C.; Zhang, S.; Yu, N.; Feng, Y.; Yang, X.; Zhou, Y.; Wang, H.; Niebles, J. C.; Xiong, C.; Savarese, S.; Ermon, S.; Fu, Y.; and Xu, R. 2023. UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild. In *ArXiv*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. In *Arxiv*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. In *CVPR*.

Saharia, C.; Chan, W.; Chang, H.; Lee, C. A.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-Image Diffusion Models. In *Arxiv*.

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.; and Norouzi, M. 2022b. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Arxiv*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising diffusion implicit models. In *ICLR*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going Deeper with Convolutions. In *CVPR*.

Timofte, R.; Agustsson, E.; Gool, L. V.; Yang, M.-H.; and Zhang, L. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 114–125.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023. Exploiting Diffusion Prior for Real-World Image Super-Resolution. In *Arxiv*.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021a. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *CVPR*.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021b. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *ICCVW*.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018a. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Loy, C. C. 2018b. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*.

Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; and Lin, L. 2020. Component Divide-and-Conquer for Real-World Image Super-Resolution. In *ECCV*.

Weng, S.; Sun, J.; Li, Y.; and Li, S. 2022. CT2: Colorization transformer via color tokens. In *ECCV*.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2022a. Paint by Example: Exemplar-based Image Editing with Diffusion Models. In *ArXiv*.

Yang, J.; Wright, J.; Huang, T. S.; and Ma, Y. 2008. Image super-resolution as sparse representation of raw image patches. In *CVPR*.

Yang, T.; Ren, P.; Xie, X.; ; and Zhang, L. 2021. GAN Prior Embedded Network for Blind Face Restoration in the Wild. In *CVPR*.

Yang, T.; Ren, P.; Xie, X.; Hua, X.; and Zhang, L. 2022b. Beyond a Video Frame Interpolator: A Space Decoupled Learning Approach to Continuous Image Transition. In *ECCVW*.

Zhang, K.; Liang, J.; Gool, L. V.; and Timofte, R. 2021. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. In *ICCV*, 4791–4800.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *NeurIPS*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023. Inversion-Based Style Transfer with Diffusion Models. In *CVPR*.

Zhou, S.; Chan, K. C.; Li, C.; and Loy, C. C. 2022. Towards Robust Blind Face Restoration with Codebook Lookup Transformer. In *NeurIPS*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*.

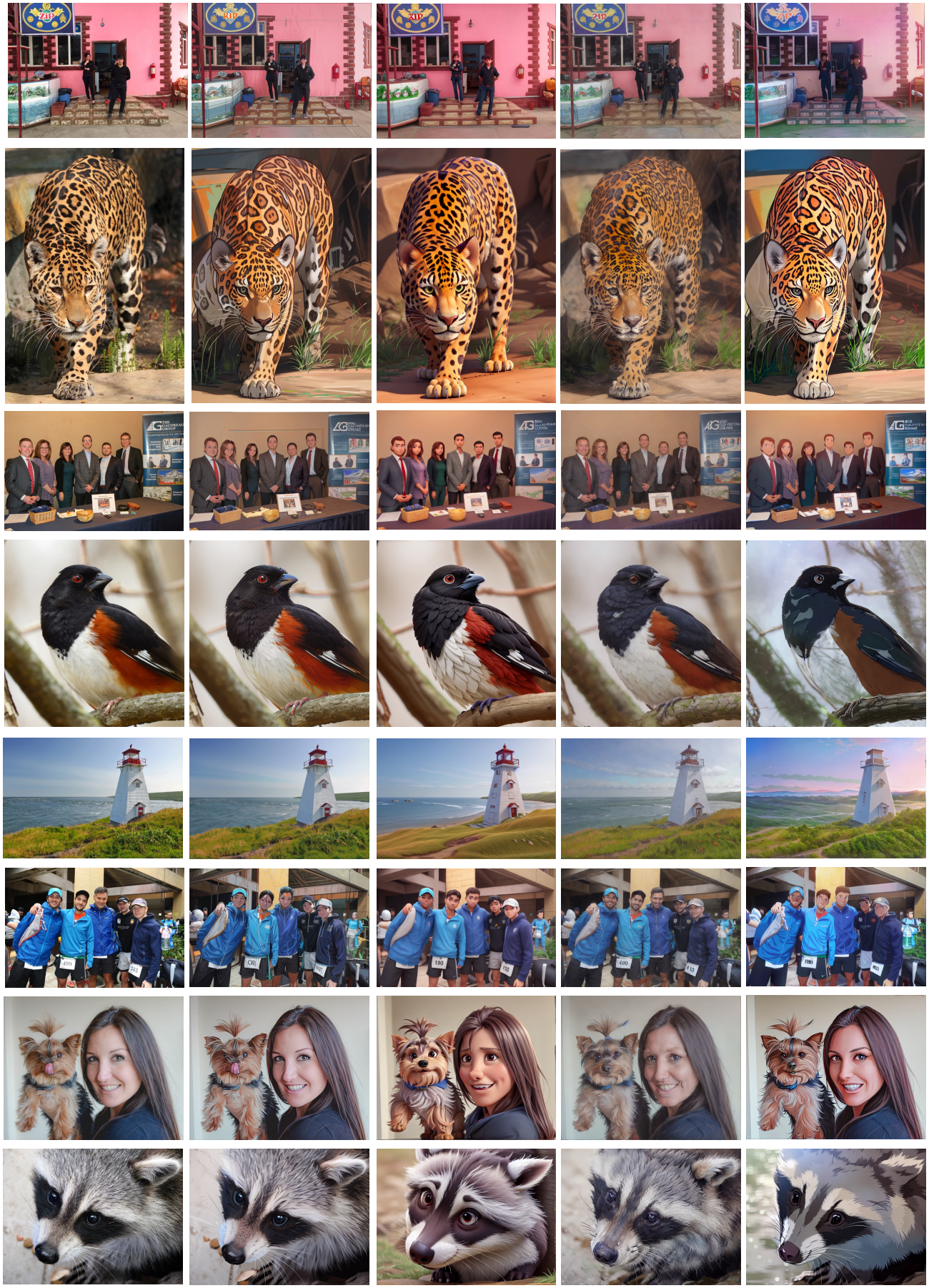

Figure 9: Stylization results by PASD with different base models (ToonYou, Disney 3D, Oil painting, Shinkai) on real-world images.

Figure 10: Qualitative comparison of different colorization methods on real-world images.