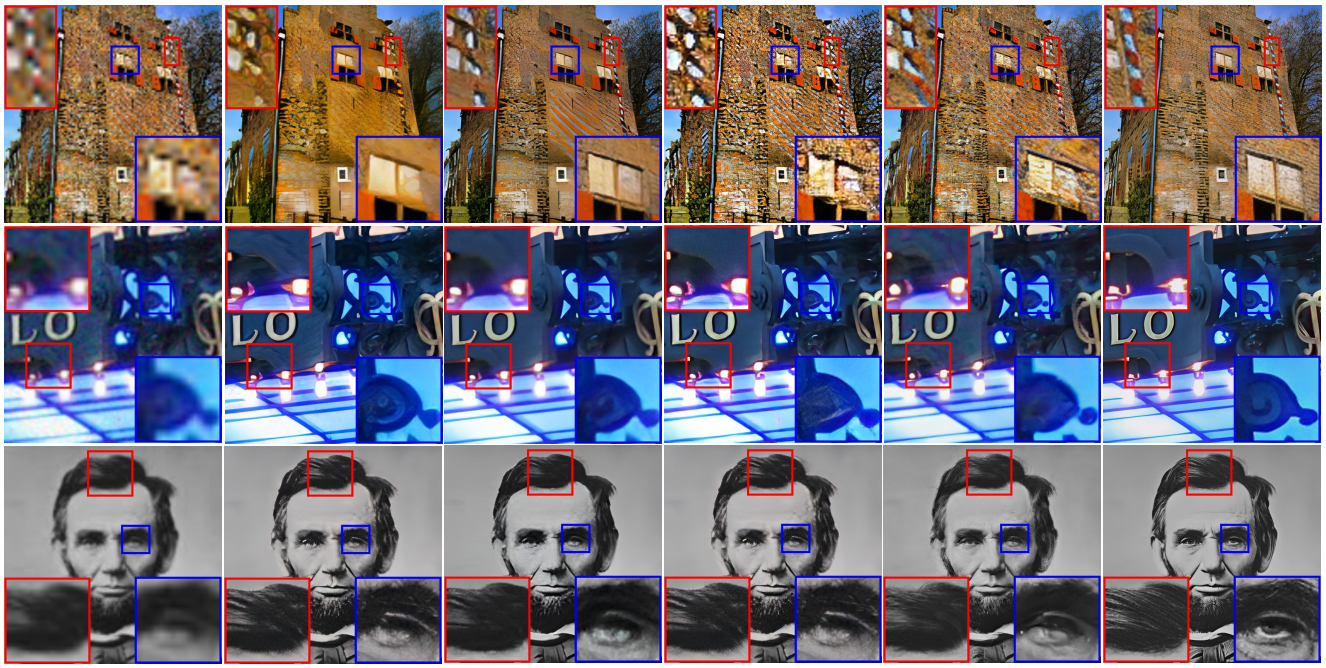
Exploiting Diffusion Prior for Real-World Image Super-Resolution

Jianyi Wang Zongsheng Yue Shangchen Zhou Kelvin C.K. Chan Chen Change Loy
S-Lab, Nanyang Technological University
{jianyi001, zongsheng.yue, s200094, chan0899, ccloy}@ntu.edu.sg



(a) Zoomed LR (b) BSRGAN (c) Real-ESRGAN+ (d) FeMaSR (e) LDM (f) Ours

Figure 1: Qualitative comparisons of BSRGAN [78], Real-ESRGAN+ [64], FeMaSR [8], LDM [51], and our StableSR on real-world examples. (**Zoom in for details**)

Abstract

We present a novel approach to leverage prior knowledge encapsulated in pre-trained text-to-image diffusion models for blind super-resolution (SR). Specifically, by employing our time-aware encoder, we can achieve promising restoration results without altering the pre-trained synthesis model, thereby preserving the generative prior and minimizing training cost. To remedy the loss of fidelity caused by the inherent stochasticity of diffusion models, we introduce a controllable feature wrapping module that allows users to balance quality and fidelity by simply adjusting a scalar value during the inference process. Moreover, we develop a progressive aggregation sampling strategy to overcome the fixed-size constraints of pre-trained diffusion models, enabling adaptation to resolutions of any size. A comprehensive evaluation of our method using both synthetic and real-world benchmarks demonstrates its superiority over current state-of-the-art approaches.

1. Introduction

We have seen significant advancements in diffusion models [27, 45, 57, 74] for image synthesis tasks. Existing research demonstrates that the diffusion prior, embedded in synthesis models like Stable Diffusion [51], can be applied to various downstream content creation tasks, including image [2, 10, 20, 22, 25, 44, 79] and video [43, 48, 70] editing. In this study, we extend our exploration beyond the realm of content creation and examine the potential benefits of using the diffusion prior for super-resolution (SR). This low-level vision task presents an additional challenge, as it requires high image fidelity, which stands in contrast to the stochas-

tic nature of diffusion models.

A common solution to the challenge above involves training a SR model from scratch [36, 51, 53, 55]. To preserve fidelity, these methods use the low-resolution (LR) image as an additional input to constrain the output space. While they have achieved notable success, these approaches often demand significant computational resources to train the diffusion model. Moreover, training a network from scratch can potentially jeopardize the generative priors captured in synthesis models, leading to suboptimal performance in the final network. These limitations have inspired an alternative approach [10, 12, 41, 58, 67], which involves incorporating constraints into the reverse diffusion process of a pre-trained synthesis model. This paradigm avoids the need for model training while leveraging the diffusion prior. However, designing these constraints assumes knowing the image degradations a priori, which is typically unknown and complex. Consequently, such methods exhibit limited generalizability.

In this study, we present **StableSR**, an approach that ***preserves pre-trained diffusion priors without making explicit assumptions about the degradations.*** Specifically, unlike previous works [36, 51, 53, 55] that concatenate the LR image to intermediate outputs, which requires one to train a diffusion model from scratch, our method only needs to fine-tune a lightweight *time-aware encoder* and a few feature modulation layers for the SR task. Our encoder incorporates a time embedding layer to generate time-aware features, allowing the features in the diffusion model to be adaptively modulated at different iterations. Besides gaining improved training efficiency, keeping the original diffusion model frozen help preserve the generative prior. The time-aware encoder also helps maintain fidelity by providing adaptive guidance for each diffusion step during the restoration process, *i.e.*, stronger guidance at earlier iterations and weaker guidance later. Our experiments confirm that this time-aware property is crucial for achieving performance improvements.

To suppress randomness inherited from the diffusion model as well as the information loss due to the encoding process of VQGAN [17], inspired by Codeformer [85], we apply a *controllable feature wrapping module* with an adjustable coefficient to refine the outputs of the diffusion model during the decoding process of VQGAN. Specifically, multi-scale intermediate features from VQGAN encoder are adopted to tune the decoder features in a residual manner. With the adjustable coefficient to control the residual strength, we can further realize a continuous fidelity-realism trade-off to handle both light and heavy degradations.

Applying diffusion models to arbitrary resolutions has remained a persistent challenge. A simple solution would be to split the image into patches and process each independently. However, this method often leads to boundary discontinuity in the output. To address this issue, we introduce a *progressive aggregation sampling strategy*. Our approach involves dividing the image into overlapping patches and fusing these patches using a Gaussian kernel at each diffusion iteration. This process smooths out the boundaries, resulting in a more coherent output.

Adapting generative priors for real-world image super-resolution presents an intriguing yet challenging problem, and in this work, we offer a novel approach as a solution. We introduce a fine-tuning method that leverages pre-trained diffusion models without making explicit assumptions about degradations. We address key challenges, such as fidelity and arbitrary resolution, by proposing simple yet effective modules. With our time-aware encoder, controllable feature wrapping module, and progressive aggregation sampling strategy, our *StableSR* serves as a strong baseline that inspires future research in adopting diffusion priors for restoration tasks.

2. Related Work

Image Super-Resolution. Image Super-Resolution (SR) aims to restore an HR image from its degraded LR observation. Early SR approaches [9, 13–16, 24, 35, 37, 56, 66, 71, 72, 81] assume a pre-defined degradation process, e.g., bicubic downsampling and blurring with known parameters. While these methods can achieve appealing performance on the synthetic data with the same degradation, their performance deteriorates significantly in real-world scenarios due to the limited generalizability.

Recent works have moved their focus from synthetic settings to blind SR, where the degradation is unknown and similar to real-world scenarios. Due to the lack of real-world paired data for training, some methods [19, 40, 60, 62, 69, 77] propose to implicitly learn a degradation model from LR images in an unsupervised manner such as CycleGAN [87] and contrastive learning [46]. In addition to unsupervised learning, other approaches aim to explicitly synthesize LR-HR image pairs that resemble real-world data. Specifically, BSRGAN [78] and Real-ESRGAN [64] present effective degradation pipelines for blind SR in real world. Building upon such degradation pipelines, recent works based on diffusion models [53, 55] further show competitive performance on real-world image SR. In this work, we consider an orthogonal direction of fine-tuning a diffusion model for SR. In this way, the computational cost of network training could be reduced. Moreover, our method allows the exploitation of generative prior encapsulated in the synthesis model, leading to better performance.

Prior for Image Super-Resolution. To further enhance performance in complex real-world SR scenarios, numerous prior-based approaches have been proposed. These techniques deploy additional image priors to bolster the gen-

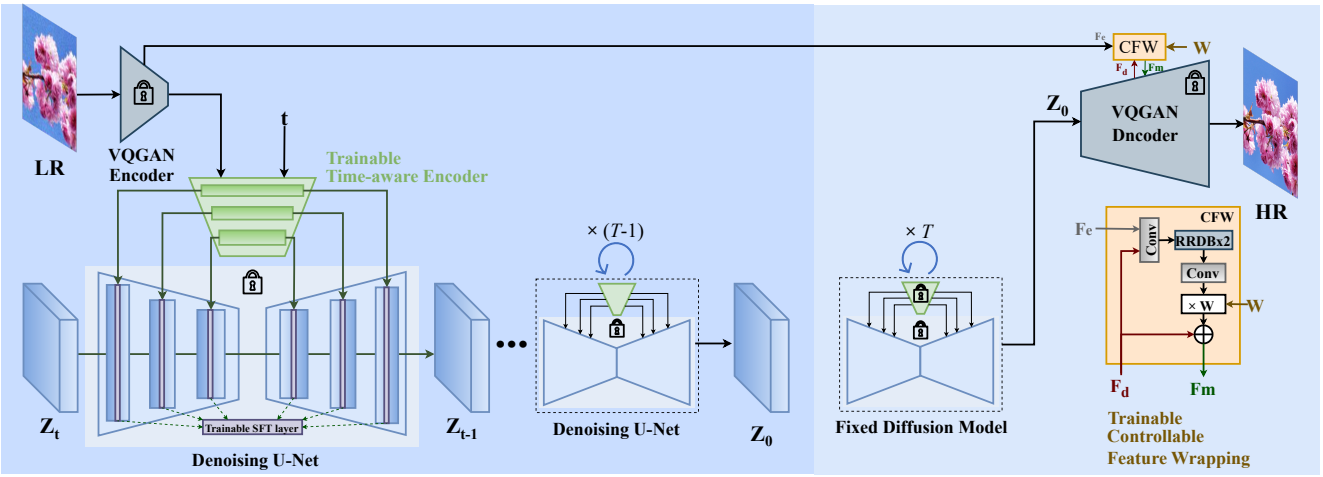


Figure 2: Framework of StableSR. We first finetune the time-aware encoder that is attached to a fixed pre-trained Stable Diffusion model. Features are combined with trainable spatial feature transform (SFT) layers. Such a simple yet effective design is capable of leveraging rich diffusion prior for image SR. Then, the diffusion model is fixed. Inspired by CodeFormer [85], we introduce a controllable feature wrapping (CFW) module to obtain a tuned feature $\boldsymbol{F}_m$ in a residual manner, given the additional information $\boldsymbol{F}_e$ from LR features and features $\boldsymbol{F}_d$ from the fixed VQGAN decoder. With an adjustable coefficient w, CFW can trade between quality and fidelity.

eration of faithful textures. A straightforward method is reference-based SR [31, 73, 82, 84, 86]. This involves using one or several reference high-resolution (HR) images, which share similar textures with the input low-resolution (LR) image, as an explicit prior to aid in generating the corresponding HR output. However, aligning features of the reference with the LR input can be challenging in real-world cases, and such explicit priors are not always readily available. Recent works have moved away from relying on explicit priors, finding more promising performance with implicit priors instead. Wang *et al*. [65] were the first to propose the use of semantic segmentation probability maps for guiding SR in the feature space. Subsequent works [6, 7, 21, 42, 47, 63, 75] employed pre-trained GANs by exploring the corresponding high-resolution latent space of the low-resolution input. While effective, the implicit priors used in these approaches are often tailored for specific scenarios, such as limited categories [6,21,47,65] and faces [42, 63, 75], and therefore lack generalizability for complex real-world SR tasks. Other implicit priors for general image SR include mixtures of degradation experts [38, 76] and VQGAN [8, 83]. However, these methods fall short, either due to insufficient prior expressiveness [38, 76, 83] or inaccurate feature matching [8], resulting in output quality that remains less than satisfactory.

In contrast to existing strategies, we set our sights on exploring the robust and extensive generative prior found in pre-trained diffusion models [45, 49–51, 54]. While recent studies [2, 10, 28, 44, 79] have highlighted the remarkable generative abilities of pre-trained diffusion models, the high-fidelity requirement inherent in super-resolution (SR) makes it unfeasible to directly adopt these methods for this task. Our proposed StableSR, unlike LDM [51], does not necessitate training from scratch. Instead, it fine-tunes directly on a frozen pre-trained diffusion model with only a small number of trainable parameters. This approach is significantly more efficient and demonstrates superior performance in practice.

3. Methodology

Our method employs diffusion prior for SR. Inspired by the generative capabilities of Stable Diffusion [51], we use it as the diffusion prior in our work, hence the name *StableSR* for our method. The main component of StableSR is a time-aware encoder, which is trained along with a frozen Stable Diffusion model to allow for conditioning based on the input image. To further facilitate a trade-off between realism and fidelity, depending on user preference, we follow CodeFormer [85] to introduce an optional controllable feature wrapping module. The overall framework of StableSR is depicted in Fig. 2.

3.1. Guided Finetuning with Time Awareness

To exploit the prior knowledge of Stable Diffusion for SR, we establish the following constraints when designing our model: 1) The resulting model must have the ability to generate a plausible HR image, conditioned on the observed LR input. This is vital because the LR image is the only source of structural information, which is crucial for maintaining high fidelity. 2) The model should introduce only minimal alterations to the original Stable Diffusion model

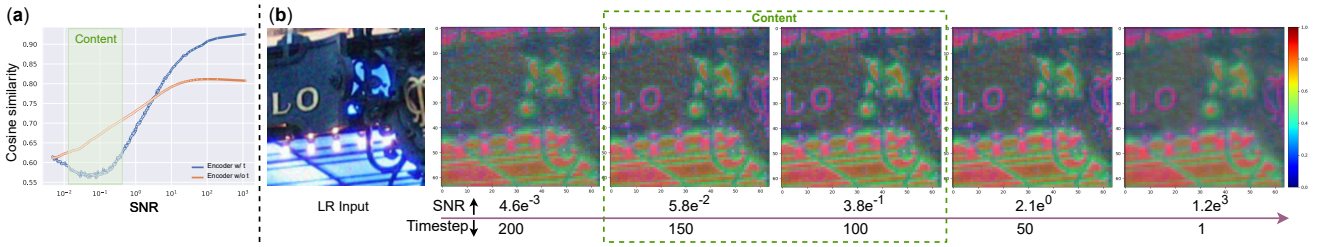


Figure 3: In contrast to a conditional encoder without time embedding, the one equipped with time embedding can adaptively supply guidance to the pre-trained diffusion models. (a), we gauge the cosine similarity between the diffusion model's features pre- and post-SFT at various timesteps, which echoes the strength of the condition originating from the encoder. (b), we further visualize the features of the conditional encoder extracted from the LR image. As shown, the encoder is inclined to provide sharp features when the SNR hovers around 5e^{-2}. This is precisely when the diffusion model requires substantial guidance to generate the desired high-resolution image content. Interestingly, this observation aligns with the findings reported in [11].

to prevent disrupting the prior encapsulated within it.

Feature Modulation. While several existing approaches [3, 18, 25, 45, 51] have successfully controlled the generated semantic structure of a diffusion model via cross-attention, such a strategy can hardly provide detailed and high-frequency guidance due to insufficient inductive bias [39]. To more accurately guide the generation process, we adopt an additional encoder to extract multi-scale features $\{\mathbf{F}^n\}_{n=1}^N$ from the degraded LR image features, and use them to modulate the intermediate feature maps $\{\mathbf{F}^n_{\text{dif}}\}_{n=1}^N$ of the residual blocks in Stable Diffusion via spatial feature transformations (SFT) [65]:

$$\hat{\mathbf{F}}^n_{\text{dif}} = (1+\boldsymbol{\alpha}^n) \odot \mathbf{F}^n_{\text{dif}} + \boldsymbol{\beta}^n;\ \boldsymbol{\alpha}^n, \boldsymbol{\beta}^n = \mathcal{M}^n_\theta(\mathbf{F}^n), \quad (1)$$

where $\boldsymbol{\alpha}^n$ and $\boldsymbol{\beta}^n$ denote the affine parameters in SFT and $\mathcal{M}^n_\theta$ denotes a small network consisting of several convolutional layers. Here n indices the spatial scale of the UNet [52] architecture in Stable Diffusion.

During finetuning, we freeze the weights of Stable Diffusion and train only the encoder and SFT layers. This strategy allows us to insert structural information extracted from the LR image without destroying the generative prior captured by Stable Diffusion.

Time-aware Guidance. We find that incorporating temporal information through a time-embedding layer in our encoder considerably enhances both the quality of generation and the fidelity to the ground truth, since it can adaptively adjust the condition strength derived from the LR features. Here, we provide an analysis of this phenomenon from a signal-to-noise ratio (SNR) standpoint, and later quantitatively and qualitatively confirm it in the ablation study found in Section 4.4.

During the generation process, the SNR of the produced image progressively increases as noise is incrementally removed. A recent study [11] indicates that image content is rapidly populated when the SNR approaches 5e^{-2}. In line with this observation, our proposed encoder is designed to offer comparatively strong conditions to the diffusion model within the range where the SNR hits 5e^{-2}. This is essential because the content generated at this stage significantly influences the super-resolution performance of our method. To further substantiate this, we employ the cosine similarity between the features of Stable Diffusion before and after the SFT to measure the condition strength provided by the encoder. The cosine similarity values at different timesteps are plotted in Fig. 3-(a). As can be observed, the cosine similarity reaches its minimum value around an SNR of 5e^{-2}, indicative of the strongest conditions imposed by the encoder. In addition, we also depict the feature maps extracted from our specially designed encoder in Fig. 3-(b). It is noticeable that the features around the SNR point of 5e^{-2} are sharper and contain more detailed image structures. We hypothesize that these adaptive feature conditions can furnish more comprehensive guidance for SR.

Color Correction. Diffusion models can occasionally exhibit color shifts, as noted in [11]. To address this issue, we perform color normalization on the generated image to align its mean and variance with those of the LR input. In particular, if we let $\boldsymbol{x}$ denote the LR input and $\hat{\boldsymbol{y}}$ represent the generated HR image, the color-corrected output, $\boldsymbol{y}$, is calculated as follows:

$$\boldsymbol{y}^c = \frac{\hat{\boldsymbol{y}}^c - \boldsymbol{\mu}^c_{\hat{\boldsymbol{y}}}}{\boldsymbol{\sigma}^c_{\hat{\boldsymbol{y}}}} \cdot \boldsymbol{\sigma}^c_x + \boldsymbol{\mu}^c_x, \quad (2)$$

where $c \in \{r, g, b\}$ denotes the color channel, $\boldsymbol{\mu}^c_{\hat{\boldsymbol{y}}}$ and $\boldsymbol{\sigma}^c_{\hat{\boldsymbol{y}}}$ (or $\boldsymbol{\mu}^c_x$ and $\boldsymbol{\sigma}^c_x$) are the mean and standard variance estimated from the c-th channel of $\hat{\boldsymbol{y}}$ (or $\boldsymbol{x}$), respectively. We find that this simple correction suffices to remedy the color difference.

3.2. Fidelity-Realism Trade-off

Although the output of the proposed approach is visually compelling, it often deviates from the ground truth due to the inherent stochasticity of the diffusion model. Drawing

Zoomed LR StableSR w/o Agg. StableSR

Figure 4: When dealing with images beyond 512×512, StableSR (w/o aggregation sampling) suffers from obvious block inconsistency by chopping the image into several tiles, processing them separately, and stitching them together. With our proposed aggregation sampling, StableSR can achieve consistent results on large images. The resolution of the shown figure is 1024×1024.

inspiration from CodeFormer [85], we introduce a Controllable Feature Wrapping (CFW) module to flexibly manage the balance between realism and fidelity.

Since Stable Diffusion is implemented in the latent space of VQGAN [17], it is natural to leverage the VQGAN encoder features to modulate the VQGAN decoder features to further improve fidelity. Let $\boldsymbol{F}_e$ and $\boldsymbol{F}_d$ be the VQGAN encoder and decoder features, respectively. We introduce an adjustable coefficient $w \in [0, 1]$ to control the extent of modulation:

$$\boldsymbol{F}_m = \boldsymbol{F}_d + \mathcal{C}(\boldsymbol{F}_e, \boldsymbol{F}_d; \boldsymbol{\theta}) \times w, \tag{3}$$

where $\mathcal{C}(\cdot; \boldsymbol{\theta})$ represents convolution layers with trainable parameter $\boldsymbol{\theta}$. The overall framework is shown in Fig. 2.

In this design, a small w exploits the generation capability of Stable Diffusion, leading to outputs with high realism. In contrast, a large w allows stronger structural guidance from the LR image, enhancing fidelity. We observe that $w = 0.5$ achieves a good balance between quality and fidelity. Note that we only train CFW in this particular stage.

3.3. Aggregation Sampling

Despite its fully convolutional nature, Stable Diffusion tends to produce inferior outputs for resolutions differing from its training settings, specifically $512, \times, 512$. This, in effect, constrains the practicality of StableSR.

A common workaround involves splitting the larger image into several overlapping smaller patches and processing each one individually. While this strategy often yields good results for conventional CNN-based SR methods, it is not directly applicable in the diffusion paradigm. This is because discrepancies between patches are compounded and magnified over the course of diffusion iterations. A typical failure case is illustrated in Fig. 4.

Inspired by AB. Jiménez [4], we apply a progressive patch aggregation sampling algorithm to handle images of arbitrary resolutions. Specifically, we begin by encoding the low-resolution image into a latent feature map, which is then subdivided into multiple overlapping small patches, each with a resolution of 64×64 - matching the training resolution[1]. During each timestep in the reverse sampling, each patch is individually processed through StableSR, with the processed patches subsequently aggregated. To integrate overlapping patches, a weight map of size 64×64 is generated for each patch using a centered Gaussian kernel. Overlapping pixels are then weighted in accordance with their respective Gaussian weight maps. This procedure is reiterated until the final iteration is reached. Our experiments suggest that this progressive aggregation method substantially mitigates discrepancies in the overlapped regions, as depicted in Fig. 4. The algorithm is detailed in the supplementary material.

4. Experiments

4.1. Implementation Details

StableSR is built based on Stable Diffusion 2.1-base[2]. Our time-aware encoder is similar to the contracting path of the denoising U-Net in Stable Diffusion but is much more lightweight (∼105M, including SFT layers). SFT layers are inserted in each residual block of Stable Diffusion for effective control. We finetune the diffusion model of StableSR for 117 epochs with a batch size of 192, and the prompt is fixed as null. We follow Stable Diffusion to use Adam [34] optimizer and the learning rate is set to 5×10^{-5}. The training process is conducted on 512×512 resolution with 8 NVIDIA Tesla 32G-V100 GPUs. For inference, we adopt DDPM sampling [27] with 200 timesteps. To handle images with arbitrary sizes, we adopt the proposed aggregation sampling strategy for images beyond 512×512. As for images under 512×512, we first enlarge the LR images to 512×512 using zero padding and remove the padding parts after generation.

To train CFW, we first generate 100k synthetic LR-HR pairs with 512×512 resolution following the degradation pipeline in Real-ESRGAN [64]. Then we adopt the finetuned diffusion model to generate the corresponding latent codes Z_0 given the above LR images as conditions. The training losses are almost the same as VQGAN, except that we use a fixed adversarial loss weight of 0.025 rather than a self-adjustable one.

4.2. Experimental Settings

Training Datasets. We adopt the degradation pipeline of Real-ESRGAN [64] to synthesize LR/HR pairs on DIV2K [1], DIV8K [23], Flickr2K [59] and OutdoorSceneTraining

[1]The downsampling scale factor of the VQGAN encoder in Stable Diffusion is $8\times$.

[2]https://huggingface.co/stabilityai/stable-diffusion-2-1-base

Table 1: Quantitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks. **Red** and blue colors represent the best and second best performance, respectively.

Datasets	Metrics	RealSR [30]	BSRGAN [77]	Real-ESRGAN+ [64]	DASR [38]	FeMaSR [8]	LDM [51]	StableSR ($w = 0.0$)	**StableSR** ($w = 0.5$)	StableSR ($w = 1.0$)
DIV2K Valid [1]	PSNR ↑	**24.62**	24.58	24.47	24.29	23.06	23.32	22.68	23.26	23.14
	SSIM ↑	0.5970	0.6269	0.6304	**0.6372**	0.5887	0.5762	0.5546	0.5726	0.5681
	LPIPS ↓	0.5276	0.3351	0.3543	0.3112	0.3126	0.3199	0.3393	0.3114	**0.3077**
	FID ↓	49.49	44.22	49.16	37.64	35.87	26.47	25.83	**24.44**	26.14
	CLIP-IQA ↑	0.3534	0.5246	0.5036	0.5276	0.5998	0.6245	0.6529	**0.6771**	0.6197
	MUSIQ ↑	28.57	61.19	55.19	61.05	60.83	62.27	65.72	**65.92**	64.31
RealSR [5]	PSNR ↑	**27.30**	26.38	27.02	25.69	25.06	25.46	24.07	24.65	24.70
	SSIM ↑	0.7579	0.7651	**0.7707**	0.7614	0.7356	0.7145	0.6829	0.7080	0.7157
	LPIPS ↓	0.3570	**0.2656**	0.3134	0.2709	0.2937	0.3159	0.3190	0.3002	0.2892
	CLIP-IQA ↑	0.3687	0.5114	0.3198	0.4495	0.5406	0.5688	0.6127	**0.6234**	0.5847
	MUSIQ ↑	38.26	63.28	41.21	60.36	59.06	58.90	65.81	**65.88**	64.05
DRealSR [68]	PSNR ↑	**30.19**	28.70	29.75	28.62	26.87	27.88	27.43	28.03	27.97
	SSIM ↑	0.8148	0.8028	**0.8262**	0.8052	0.7569	0.7448	0.7341	0.7536	0.7540
	LPIPS ↓	0.3938	0.2858	0.3099	**0.2818**	0.3157	0.3379	0.3595	0.3284	0.3080
	CLIP-IQA ↑	0.3744	0.5091	0.3813	0.4515	0.5634	0.5756	0.6340	**0.6357**	0.5893
	MUSIQ ↑	26.93	57.16	42.41	54.26	53.71	53.72	**58.98**	58.51	56.77
DPED-iphone [29]	CLIP-IQA ↑	0.4496	0.4021	0.2826	0.3389	**0.5306**	0.4482	0.5015	0.4799	0.4250
	MUSIQ ↑	45.60	45.89	32.68	42.42	49.95	44.23	**51.90**	50.48	47.96

[65] datasets. We additionally add 5000 face images from the FFHQ dataset [32] for general cases.

Testing Datasets. We evaluate our approach on both synthetic and real-world datasets. For synthetic data, we follow the degradation pipeline of Real-ESRGAN [64] and generate 3k LR-HR pairs from DIV2K validation set [1]. The resolution of LR is 128×128 and that of the corresponding HR is 512×512. Note that for StableSR, the inputs are first upsampled to the same size as the outputs before inference. For real-world datasets, we follow common settings to conduct comparisons on RealSR [5], DRealSR [68] and DPED-iphone [29]. We further collect 40 images from the Internet for comparison.

Compared Methods. To verify the effectiveness of our approach, we compare our StableSR with several state-of-the-art methods[3], *i.e.*, RealSR[4] [30], BSRGAN [77], Real-ESRGAN+ [64], DASR [38], FeMaSR [8], and latent diffusion model (LDM) [51]. Since LDM is officially trained on images with 256×256 resolution, we finetune it following the same training settings of StableSR for a fair comparison. For other methods, we directly use the official code and models for testing. Note that the results in this section are obtained on the same resolution with training, *i.e.*, 128×128. Specifically, for images from [5,29,68], we crop them at the center to obtain patches with 128×128 resolution. For other real-world images, we first resize them such that the shorter sides are 128 and then apply center cropping. As for other resolutions, one example of StableSR on real-world images under 1024×1024 resolution is shown in Fig. 4. More results are provided in the supplementary material.

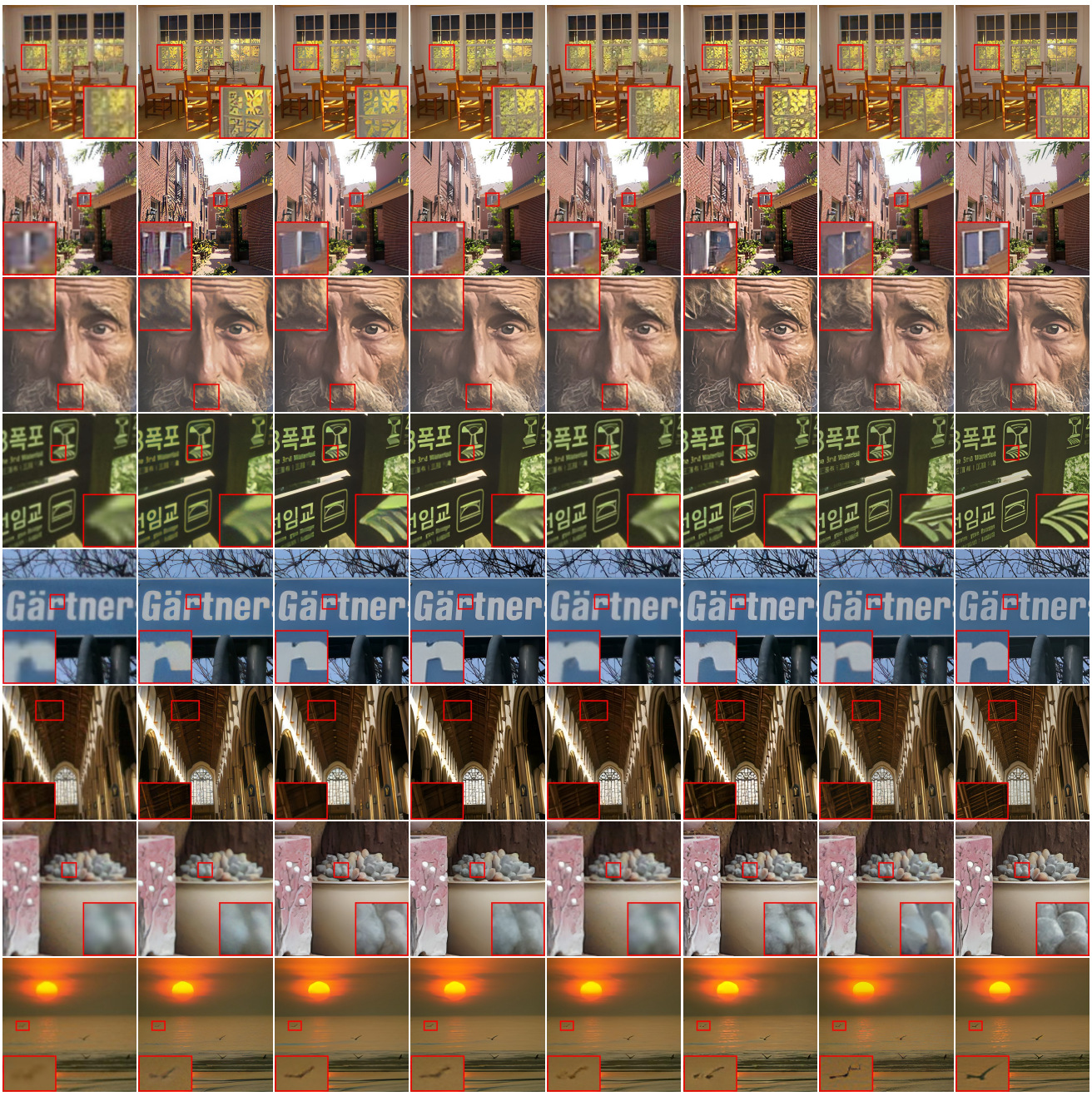
Evaluation Metrics. For benchmarks with paired data, *i.e.*, DIV2K Valid, RealSR and DRealSR, we employ various perceptual metrics including LPIPS [80], FID [26], CLIP-IQA [61] and MUSIQ [33] to evaluate the perceptual quality of generated images. PSNR and SSIM scores (evaluated on the luminance channel in YCbCr color space) are also reported for reference. Since ground-truth images are unavailable in DPED-iphone [29], we follow existing methods [8, 64] to report results on no-reference metrics *i.e.*, CLIP-IQA and MUSIQ for perceptual quality evaluation. Besides, we further conduct a user study on 16 real-world images to verify the effectiveness of our approach against existing methods.

4.3. Comparison with Existing Methods

Quantitative Comparisons. We first show the quantitative comparison on the synthetic DIV2K validation set and three real-world benchmarks. As shown in Table 1, our approach outperforms state-of-the-art SR methods in terms of multiple perceptual metrics including FID, CLIP-IQA and MUSIQ. Specifically, on synthetic benchmark DIV2K Valid, our StableSR ($w = 0.5$) achieves a 24.44 FID score, which is 7.7% lower than LDM and at least 32.9% lower than other GAN-based methods. Besides, our StableSR ($w = 0.5$) achieves the highest CLIP-IQA scores on the two commonly used real-world benchmarks [5, 68], which clearly demonstrates the superiority of StableSR. Note that although DASR and BSRGAN achieve good LPIPS scores, they show inferior performance on other perceptual metrics, i.e., FID, CLIP-IQA and MUSIQ. Moreover, they fail to restore faithful textures and generate blurry results as shown in Fig. 5. In contrast, our StableSR is capable of generating sharp images with realistic details.

Qualitative Comparisons. To demonstrate the effectiveness of our method, we present visual results on real-world images from both real-world benchmarks [5,68] and the in-

[3]SR3 [55] is not included since its official code is unavailable.
[4]We use the latest official model DF2K-JPEG.



(a) Zoomed LR (b) RealSR (c) BSRGAN (d) Real-ESRGAN+ (e) DASR (f) FeMaSR (g) LDM (h) Ours

Figure 5: Qualitative comparisons on several representative real-world samples (128 $\rightarrow$ 512). Our StableSR is capable of removing annoying artifacts and generating realistic details. (**Zoom in for details**)

ternet in Fig. 5. It is observed that StableSR outperforms previous methods in both artifact removal and detail generation. Specifically, StableSR is able to generate faithful details as shown in the first row of Fig. 5, while other methods either show blurry results (DASR, BSRGAN, Real-ESRGAN+, LDM) or unnatural details (RealSR, FeMaSR). Furthermore, for the fourth row of Fig. 5, StableSR generates sharp edges without obvious degradations, whereas other state-of-the-art methods all generate blurry results.

User Study. To further confirm the superiority of StableSR, we conduct a user study on 16 real-world LR images collected from the Internet. We compare our approach with three commonly used SR methods with competitive performance, *i.e.*, BSRGAN, Real-ESRGAN+ and LDM. The comparison is conducted in pairs, *i.e.*, given a LR image as reference, the subject is asked to choose the better HR image generated from either StableSR or BSRGAN/Real-ESRGAN+/LDM. Given the 16 LR images with the three

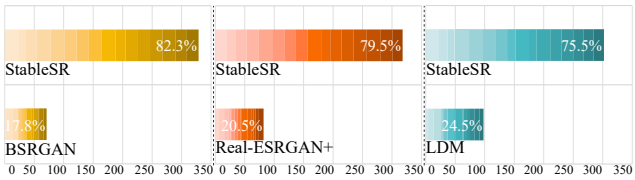


Figure 6: User study on 16 real-world images evaluated by 25 subjects. The comparisons are conducted in pairs, *i.e.*, given one LR image, the subjects are asked to choose the better HR image generated from either StableSR or BSRGAN/Real-ESRGAN+/LDM. Our StableSR outperforms other methods by gaining over 75% of the votes.

compared methods, there are 3×16 pairs evaluated by 25 subjects, resulting in 3×400 votes in total. As depicted in Fig. 6, our StableSR outperforms all three competitive methods by a large margin, gaining over 75% of the votes all the time, indicating the superiority of the proposed StableSR in real-world SR.

4.4. Ablation Study

Importance of Time-aware Guidance and Color Correction. We first investigate the significance of time-aware guidance and color correction. Recall that in Sec. 3.1, we already show that the time-aware guidance allows the encoder to adaptively adjust the condition strength. Here, we further verify its effectiveness on real-world benchmarks [5,68]. As shown in Table 2, removing time-aware guidance (*i.e.*, removing the time-embedding layer) or color correction both lead to worse SSIM and LPIPS. Moreover, the visual comparisons in Fig. 7 also indicate inferior performance without the above two components, suggesting the effectiveness of time-aware guidance and color correction.

Flexibility of Fidelity-realism Trade-off. Our CFW module inspired by CodeFormer [85] allows a flexible realism-fidelity trade-off. Particularly, given a controllable coefficient w with a range of $[0, 1]$, CFW with a small w tends to generate a realistic result while CFW with a larger w improves the fidelity. As shown in Table 1, compared with StableSR ($w = 0.0$), StableSR ($w = 1.0$) achieves higher PSNR and SSIM on all three paired benchmarks, indicating better fidelity. In contrast, StableSR ($w = 0.0$) achieves better perceptual quality with higher CLIP-IQA scores and MUSIQ scores. Similar phenomena can also be observed in Fig. 8. We further observe that a proper w can lead to improvement in both fidelity and perceptual quality. Specifically, StableSR ($w = 0.5$) shows comparable PSNR and SSIM with StableSR ($w = 1.0$) but achieves better perceptual metric scores in Table 1. Hence, we set the coefficient w to 0.5 by default for trading between quality and fidelity.

4.5. Limitation

Despite achieving encouraging performance, StableSR, essentially a diffusion-based approach, requires multi-step sampling for image generation. When using 200 steps, StableSR generates a 512×512 image within 10 seconds on one NVIDIA Tesla 32G-V100 GPU. This is comparable to LDM, but slower than SR methods using only a single forward pass. Fast sampling strategy and model distillation are two promising solutions to improve efficiency, and such exploration is left as our future work.

Table 2: Ablation studies of time-aware guidance and color correction on RealSR and DRealSR benchmarks.

Exp.	Strategies		RealSR / DRealSR		
	Time aware	Color cor.	PSNR ↑	SSIM ↑	LPIPS ↓
(a)		✓	**24.65** / 27.68	0.7040 / 0.7280	0.3157 / 0.3456
(b)	✓		22.24 / 23.86	0.6840 / 0.7179	0.3180 / 0.3544
Ours	✓	✓	**24.65 / 28.03**	**0.7080 / 0.7536**	**0.3002 / 0.3284**

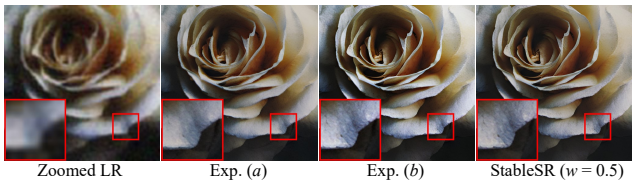


Figure 7: Visual comparisons of time-aware guidance and color correction. Exp. (a) does not apply time-aware guidance, leading to blurry textures. Exp. (b) applies time-aware guidance and can generate sharper details, but obvious color shifts can be observed. With both strategies, StableSR generates sharp textures and avoids color shifts.

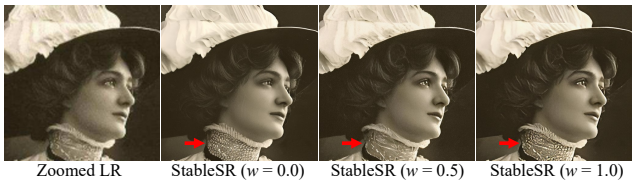


Figure 8: Visual comparisons with different coefficients w for CFW module. It is observed that a small w tends to generate a realistic result while a larger w improved the fidelity.

5. Conclusion

Motivated by the rapid development of diffusion models and their wide applications to downstream tasks, this work discusses an important yet underexplored problem of *how diffusion prior can be adopted for super-resolution.* In this paper, we present StableSR, a new way to exploit diffusion prior for real-world SR while avoiding source-intensive training from scratch. We devote our efforts to tackling the well-known problems, such as high computational cost and fixed resolution, and propose respective solutions, including the time-aware encoder, controllable feature wrapping module, and progressive aggregation sampling scheme. We believe that our exploration would lay a good foundation in this direction, and our proposed StableSR could provide useful insights for future works.

Appendix

A. Details of Time-aware Encoder

As mentioned in Sec. 4.1 of the main paper, the architecture is similar to the contracting path of the denoising U-Net in Stable Diffusion with much fewer parameters (∼105M, including SFT layers) by reducing the number of channels. The detailed settings are listed in Table 3.

Table 3: Settings of the time-aware encoder in StableSR.

Settings	*Value*
in_channels	4
model_channels	256
out_channels	256
num_res_blocks	2
dropout	0
channel_mult	[1, 1, 2, 2]
attention_resolutions	[4, 2, 1]
conv_resample	True
dims	2
use_fp16	False
num_heads	4

B. Details of Aggregation Sampling

Here, we provide further details of our progressive patch aggregation sampling algorithm. As mentioned in Sec. 3.3 of the main paper, for LR inputs with resolutions under 512×512, we first enlarge the LR images to 512×512 using zero padding and remove the padding parts after generation. Here, we mainly focus on the cases beyond 512×512.

Inspired by AB. Jiménez [4], we first transfer the LR image into latent features $\boldsymbol{F}$ using VQGAN encoder. Then $\boldsymbol{F} \in \mathcal{R}^{h\times w}$ is cropped into M over-lapped patches $\{\boldsymbol{F}_{\Omega_n}\}_{n=1}^{M}$ with a resolution[5] of 64×64, where Ω_n is the index set of the ith patch in $\boldsymbol{F}$. For each patch $\boldsymbol{F}_{\Omega_n}$, we generate a weight map $\boldsymbol{w}_{\Omega_n} \in \mathcal{R}^{h\times w}$ whose entries follow up a Gaussian filter in Ω_n and 0 elsewhere. Meanwhile, we also define a padding function $f(\cdot)$ that expands any patch of size 64×64 to the resolution of $h\times w$ by filling zeros outside the region Ω. In each timestep t, we process every patch independently and denote the output as $\epsilon_{\boldsymbol{\theta}}(\mathbf{Z}_{\Omega_n}^{(t)}, \mathbf{F}_{\Omega_n}, t)$, where $\mathbf{Z}_{\Omega_n}^{(t)}$ is the nth patch of the noisy input $\boldsymbol{Z}^{(t)}$, $\boldsymbol{\theta}$ is the parameters of the diffusion model. Next, the results of all the patches are aggregated together as follows:

$$\epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, t) = \sum_{n=1}^{M} \frac{\boldsymbol{w}_{\Omega_n}}{\hat{\boldsymbol{w}}} \odot f\left(\epsilon_{\boldsymbol{\theta}}\left(\boldsymbol{Z}_{\Omega_n}^{(t)}, \mathbf{F}_{\Omega_n}, t\right)\right), \quad (4)$$

where $\hat{\boldsymbol{w}} = \sum_n \boldsymbol{w}_{\Omega_n}$. Based on $\epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, t)$, we can obtain $\boldsymbol{Z}^{(t-1)}$ according to the sampling procedure, denoted as $\text{Sampler}(\mathbf{Z}^{(t)}, \epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, t))$, in the diffusion model. Subsequently, we re-split $\boldsymbol{Z}^{(t-1)}$ into over-lapped patches and repeat the above steps until $t = 1$. The whole process is summed up in Algorithm 1.

[5]The downsampling scale of VQGAN encoder in Stable Diffusion is $8\times$.

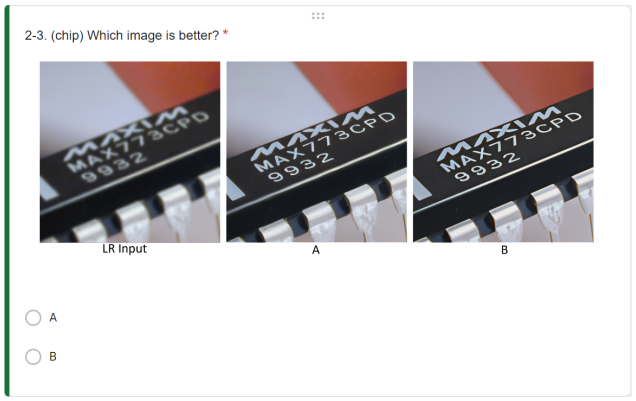


Figure 9: Screenshot of the user interface in the user study.

Algorithm 1 Progressive Patch Aggregation

Require: Cropped Regions $\{\Omega_n\}_{n=1}^{M}$, diffusion steps T, LR latent features $\boldsymbol{F}$.

Initialize $\boldsymbol{w}_{\Omega_n}$ and $\hat{\boldsymbol{w}}$
$\boldsymbol{Z}^{(T)} \sim \mathcal{N}(0, \mathbb{I})$
for $t \in [T, \dots, 0]$ **do**
  for $n \in [1, \dots, M]$ **do**
    Compute $\epsilon_{\boldsymbol{\theta}}\left(\mathbf{Z}_{\Omega_n}^{(t)}, \mathbf{F}_{\Omega_n}, t\right)$
  end for
  Compute $\epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, t)$ following Eq. (4)
  $\boldsymbol{Z}^{(t-1)} = \text{Sampler}(\boldsymbol{Z}^{(t)}, \epsilon_{\boldsymbol{\theta}}(\mathbf{Z}^{(\mathbf{t})}, \mathbf{F}, \mathbf{t}))$
end for
return $\mathbf{Z}_0$

C. User Study Settings

Here, we provide more details about our user study settings. We apply Google Form as the study platform and an example is shown in Fig. 9. The user study is conducted on 16 real-world images evaluated by 25 subjects. Given the LR reference, the subject is asked to choose the better HR image generated from either StableSR or BSRGAN/Real-ESRGAN+/LDM. There are 48 questions in total with random orders for each subject.

D. Additional Visual Results

D.1. Visual Results on Fixed Resolution

In this section, we provide additional qualitative comparisons on real-world images w/o ground truths under the resolution of 512×512. We follow Sec. 4.2 of the main paper to obtain LR images with 128×128 resolution. As

(a) Zoomed LR (b) RealSR (c) BSRGAN (d) Real-ESRGAN+ (e) DASR (f) FeMaSR (g) LDM (h) Ours

Figure 10: More qualitative comparisons on real-world images ($128 \rightarrow 512$). While existing methods typically fail to restore realistic textures under complicated degradations, our StableSR outperforms these methods by a large margin. (**Zoom in for details**)

shown in Fig. 10, StableSR successfully produces outputs with finer details and sharper edges, significantly outperforming state-of-the-art methods.

D.2. Visual Results on Arbitrary Resolution

In this section, we provide additional qualitative comparisons on the original resolution of real-world images w/o ground truths. As shown in Fig. 11, StableSR can generate realistic textures under diverse and complicated real-world scenarios such as buildings and texts, while existing methods either lead to blurry results or introduce unpleasant artifacts.

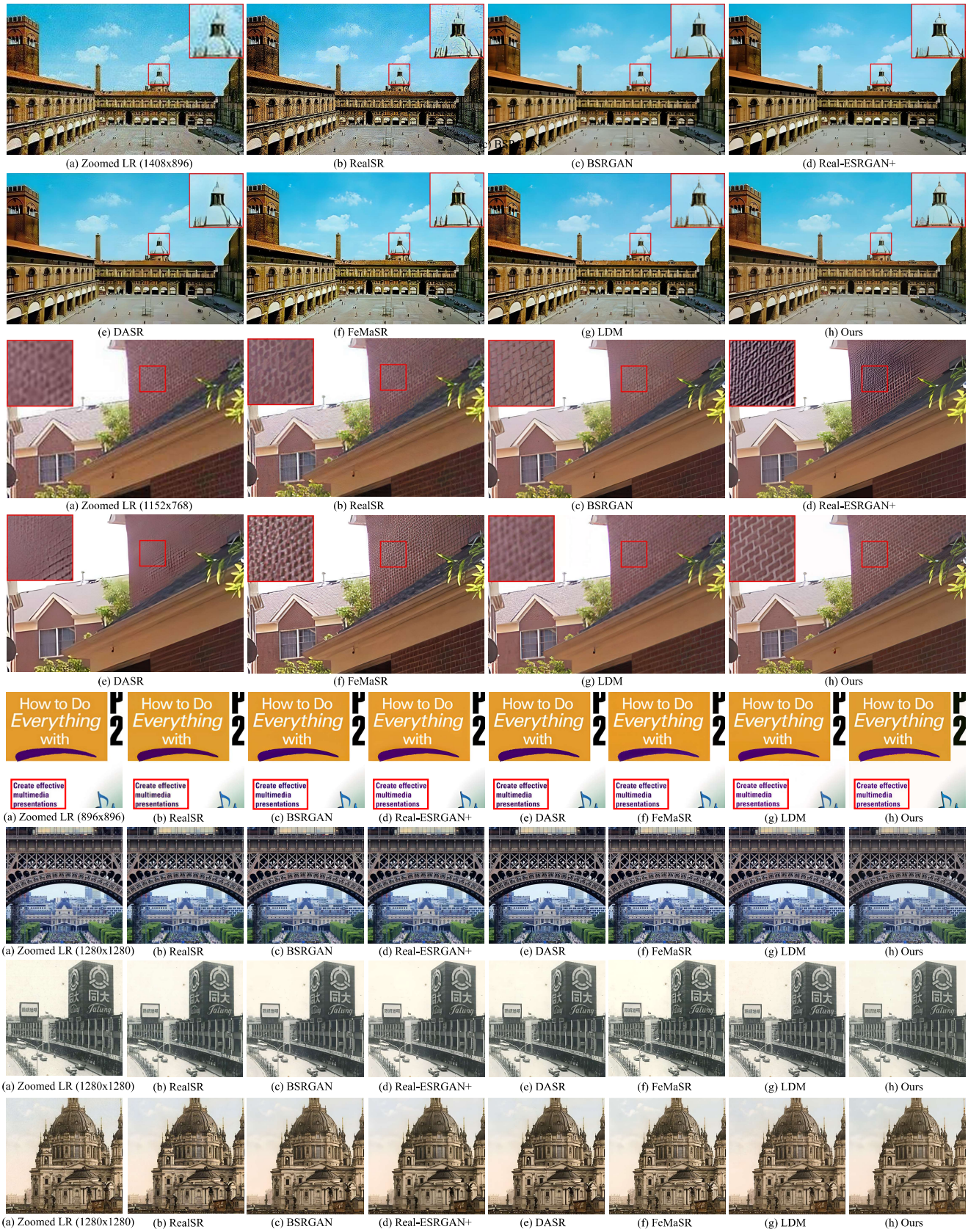


Figure 11: More qualitative comparisons on original real-world images with diverse resolutions. Our StableSR is capable of generating vivid details without annoying artifacts. (**Zoom in for details**)

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (CVPR-W)*, 2017.

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022.

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[4] Álvaro Barbero Jiménez. Mixture of diffusers for scene composition and high resolution image generation. *arXiv e-prints*, pages arXiv–2302, 2023.

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[6] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. GLEAN: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. GLEAN: Generative latent bank for large-factor image super-resolution and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[8] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022.

[9] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[10] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[11] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11472–11481, 2022.

[12] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[13] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[16] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

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

[18] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *Proceedings of International Conference on Learning Representations (ICLR)*, 2023.

[19] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV-W)*, 2019.

[20] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023.

[21] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10696–10706, 2022.

[23] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV-W)*.

[24] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt im-

age editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[28] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022.

[29] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[30] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (CVPR-W)*, 2020.

[31] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[36] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

[37] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV-W)*, 2021.

[38] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[40] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[41] Xiangming Meng and Yoshiyuki Kabashima. Diffusion model based posterior sampling for noisy linear inverse problems. *arXiv preprint arXiv:2211.12343*, 2022.

[42] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[43] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

[44] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[45] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of International Conference on Machine Learning (ICML)*, 2022.

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[47] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[48] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

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

[50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of International Conference on Machine Learning (ICML)*, 2021.

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[53] Hshmat Sahak, Daniel Watson, Chitwan Saharia, and David Fleet. Denoising diffusion probabilistic models for robust image super-resolution in the wild. *arXiv preprint arXiv:2302.07864*, 2023.

[54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

[56] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of International Conference on Machine Learning (ICML)*, 2015.

[58] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023.

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

[60] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[61] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[62] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[63] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[64] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV-W)*, 2021.

[65] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[66] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops (ECCV-W)*, 2018.

[67] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

[68] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[69] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[70] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

[71] Xiangyu Xu, Yongrui Ma, and Wenxiu Sun. Towards real scene super-resolution with raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[72] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[73] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[74] S. Yang, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[75] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[76] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement

learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[77] Jiahui Zhang, Shijian Lu, Fangneng Zhan, and Yingchen Yu. Blind image super-resolution via contrastive representation learning. *arXiv preprint arXiv:2107.00708*, 2021.

[78] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[79] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[80] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[81] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[82] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[83] Yang Zhao, Yu-Chuan Su, Chun-Te Chu, Yandong Li, Marius Renn, Yukun Zhu, Changyou Chen, and Xuhui Jia. Rethinking deep face restoration. In *cvpr*, 2022.

[84] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[85] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[86] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[87] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.