Noname manuscript No.
(will be inserted by the editor)

Exploiting Diffusion Prior for Real-World Image Super-Resolution

Jianyi Wang · Zongsheng Yue · Shangchen Zhou · Kelvin C.K. Chan · Chen Change Loy

arXiv:2305.07015v3 [cs.CV] 5 Nov 2023

Received: date / Accepted: date

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 employ 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. Code and models are available at https://github.com/IceClear/StableSR.

Jianyi Wang
S-Lab, Nanyang Technological University, Singapore
E-mail: jianyi001@ntu.edu.sg

Zongsheng Yue
S-Lab, Nanyang Technological University, Singapore
E-mail: zongsheng.yue@ntu.edu.sg

Shangchen Zhou
S-Lab, Nanyang Technological University, Singapore
E-mail: s200094@ntu.edu.sg

Kelvin C.K. Chan
S-Lab, Nanyang Technological University, Singapore
E-mail: chan0899@ntu.edu.sg

Chen Change Loy (Corresponding author)
S-Lab, Nanyang Technological University, Singapore
E-mail: ccloy@ntu.edu.sg

Keywords Super-resolution · image restoration · diffusion models · generative prior

1 Introduction

We have seen significant advancements in diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Yang et al., 2021a; Nichol et al., 2022) for the task of image synthesis. Existing studies demonstrate that the diffusion prior, embedded in synthesis models like Stable Diffusion (Rombach et al., 2022), can be applied to various downstream content creation tasks, including image (Choi et al., 2021; Avrahami et al., 2022; Hertz et al., 2022; Gu et al., 2022; Mou et al., 2023; Zhang et al., 2023; Gal et al., 2023) and video (Wu et al., 2022; Molad et al., 2023; Qi et al., 2023) editing. In this study, we extend the exploration beyond the realm of content creation and examine the potential benefits of using diffusion prior for super-resolution (SR). This low-level vision task presents an additional non-trivial challenge, as it requires high image fidelity in its generated content, which stands in contrast to the stochastic nature of diffusion models.

A common solution to the challenge above involves training a SR model from scratch (Saharia et al., 2022b; Rombach et al., 2022; Sahak et al., 2023; Li et al., 2022). To preserve fidelity, these methods use the low-resolution (LR) image as an additional input to constrain the output space. While these methods have achieved notable success, they 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

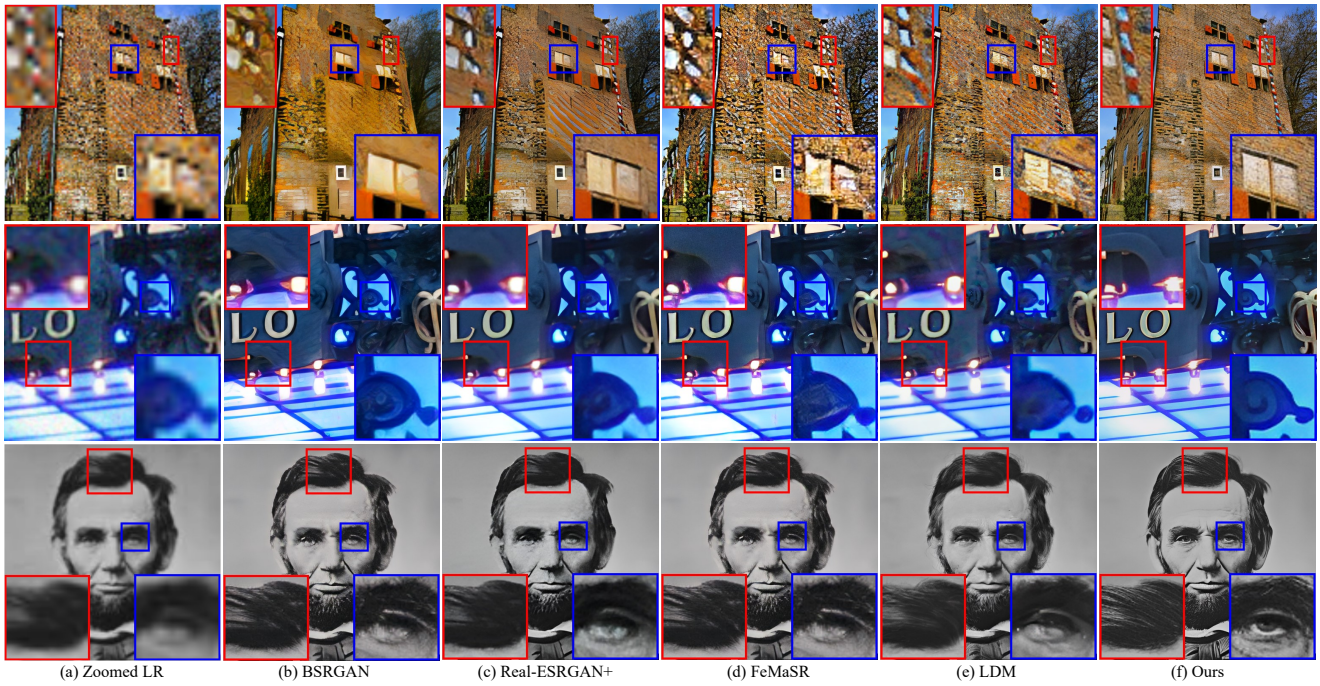

Fig. 1: Qualitative comparisons of BSRGAN (Zhang et al., 2021b), Real-ESRGAN+ (Wang et al., 2021c), FeMaSR (Chen et al., 2022), LDM (Rombach et al., 2022), and our StableSR on real-world examples. (**Zoom in for details**)

alternative approach (Choi et al., 2021; Wang et al., 2022; Chung et al., 2022; Song et al., 2023a; Meng and Kabashima, 2022), 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 as a priori, which are 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 (Saharia et al., 2022b; Rombach et al., 2022; Sahak et al., 2023; Li et al., 2022) 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 helps 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 show 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 the autoencoder (Rombach et al., 2022), inspired by Codeformer (Zhou et al., 2022), we apply a *controllable feature wrapping module* with an adjustable coefficient to refine the outputs of the diffusion model during the decoding process of the autoencoder. Specifically, multi-scale intermediate features from the 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 (Dai et al., 2019; Dong et al., 2014, 2015, 2016; He et al., 2019; Xu et al., 2019; Zhang et al., 2018b; Chen et al., 2021; Liang et al., 2021; Wang et al., 2018b; Ledig et al., 2017; Sajjadi et al., 2017; Xu et al., 2017; Zhou et al., 2020) 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 (Fritsche et al., 2019; Maeda, 2020; Wan et al., 2020; Wang et al., 2021a; Wei et al., 2021; Zhang et al., 2021a) propose to implicitly learn a degradation model from LR images in an unsupervised manner such as CycleGAN (Zhu et al., 2017) and contrastive learning (Oord et al., 2018). In addition to unsupervised learning, other approaches aim to explicitly synthesize LR-HR image pairs that resemble real-world data. Specifically, BSRGAN (Zhang et al., 2021b) and Real-ESRGAN (Wang et al., 2021c) present effective degradation pipelines for blind SR in real world. Building upon such degradation pipelines, recent works based on diffusion models (Saharia et al., 2022b; Sahak et al., 2023) 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 generation of faithful textures. A straightforward method is reference-based SR (Zheng et al., 2018; Zhang et al., 2019; Yang et al., 2020; Jiang et al., 2021; Zhou et al., 2020). 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.* (Wang et al., 2018a) were the first to propose the use of semantic segmentation probability maps for guiding SR in the feature space. Subsequent works (Menon et al., 2020; Gu et al., 2020; Wang et al., 2021b; Pan et al., 2021; Chan et al., 2021, 2022a; Yang et al., 2021b) 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 (Wang et al., 2018a; Gu et al., 2020; Pan et al., 2021; Chan et al., 2021) and faces (Menon et al., 2020; Wang et al., 2021b; Yang et al., 2021b), and therefore lack generalizability for complex real-world SR tasks. Other implicit priors for image SR include mixtures of degradation experts (Yu et al., 2018; Liang et al., 2022) and VQGAN (Zhao et al., 2022; Chen et al., 2022; Zhou et al., 2022). However, these methods fall short, either due to insufficient prior expressiveness (Yu et al., 2018; Zhao et al., 2022; Liang et al., 2022) or inaccurate feature matching (Chen et al., 2022), 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 (Nichol et al., 2022; Rombach et al., 2022; Ramesh et al., 2021; Saharia et al., 2022a; Ramesh et al., 2022). While recent studies (Choi et al., 2021; Avrahami et al., 2022; Hu et al., 2022; Zhang et al., 2023; Mou et al., 2023) 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 (Rombach et al., 2022), 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 pa-

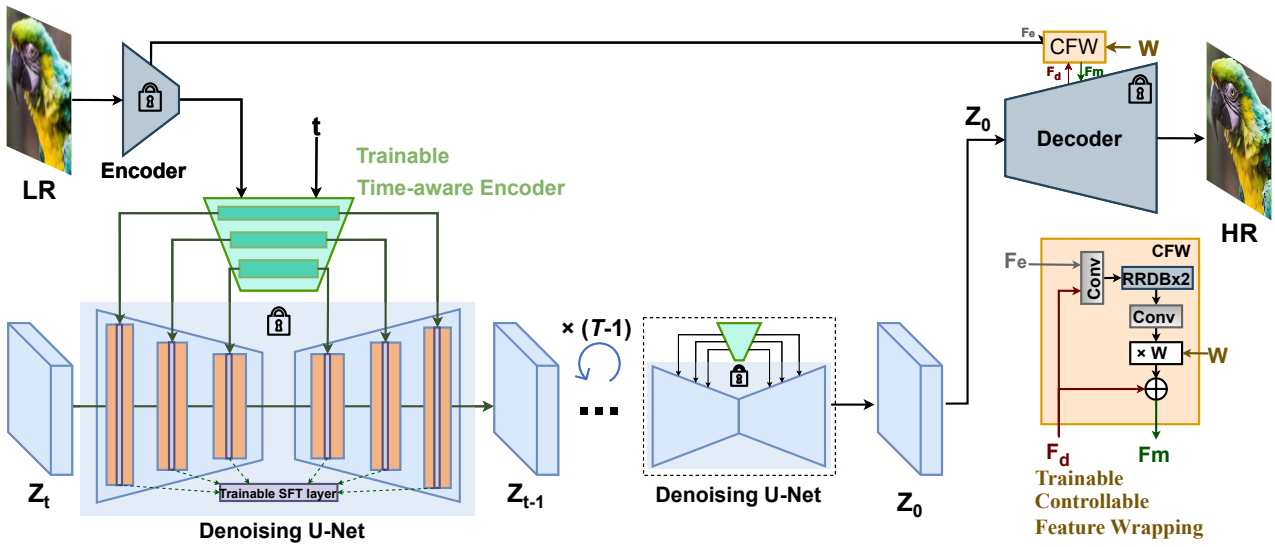

Fig. 2: Framework of StableSR. We first finetune the time-aware encoder that is attached to a fixed pre-trained Stable Diffusion model. Features are combined with trainable spatial feature transform (SFT) layers. Such a simple yet effective design is capable of leveraging rich diffusion prior for image SR. Then, the diffusion model is fixed. Inspired by CodeFormer (Zhou et al., 2022), 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 decoder. With an adjustable coefficient w, CFW can trade between quality and fidelity.

rameters. 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 (Rombach et al., 2022), 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 (Zhou et al., 2022) 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 to prevent disrupting the prior encapsulated within it.

Feature Modulation. While several existing approaches (Nichol et al., 2022; Rombach et al., 2022; Hertz et al., 2022; Feng et al., 2023; Balaji et al., 2022) 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 (Liu et al., 2021). To more accurately guide the generation process, we adopt an additional encoder to extract multi-scale features $\{\boldsymbol{F}^n\}_{n=1}^N$ from the degraded LR image features, and use them to modulate the intermediate feature maps $\{\boldsymbol{F}^n_{\text{dif}}\}_{n=1}^N$ of the residual blocks in Stable Diffusion via spatial feature transformations (SFT) (Wang et al., 2018a):

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

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 (Ronneberger et al., 2015) 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 analyze this phenomenon from a signal-to-noise ratio (SNR) standpoint and later quantitatively and qualitatively validate it in the ablation study.

During the generation process, the SNR of the produced image progressively increases as noise is incrementally removed. A recent study (Choi et al., 2022) 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 (Choi et al., 2022). 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:

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

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.

In addition to adopting color correction in the pixel domain, we further propose wavelet-based color correction for better visual performance. Given any image $\boldsymbol{I}$, we extract its high-frequence component $\boldsymbol{H}^i$ and low-frequence component $\boldsymbol{L}^i$ at the i-th ($1 \le i \le l$) scale via the wavelet decomposition, i.e.,

$$\boldsymbol{L}^i = \mathcal{C}_i(\boldsymbol{L}^{i-1}, \boldsymbol{k}), \;\; \boldsymbol{H}^i = \boldsymbol{L}^{i-1} - \boldsymbol{L}^i, \tag{3}$$

where $\boldsymbol{L}^0 = \boldsymbol{I}$, $\mathcal{C}_i$ denotes the convolutional operator with a dilation of 2^i, and k is the convolutional kernel defined as:

$$\boldsymbol{k} = \begin{bmatrix} 1/16 & 1/8 & 1/16 \\ 1/8 & 1/4 & 1/8 \\ 1/16 & 1/8 & 1/16 \end{bmatrix}. \tag{4}$$

By denoting the l-th low frequence and high frequence components of $\boldsymbol{x}$ (or $\hat{\boldsymbol{y}}$) as $\boldsymbol{L}^l_x$ and $\boldsymbol{H}^l_x$ (or $\boldsymbol{L}^l_y$ and $\boldsymbol{H}^l_y$), the desired HR output $\boldsymbol{y}$ is formulated as follows:

$$\boldsymbol{y} = \boldsymbol{H}^l_y + \boldsymbol{L}^l_x. \tag{5}$$

Intuitively, we replace the low-frequency component $\boldsymbol{L}^l_y$ of $\hat{\boldsymbol{y}}$ with $\boldsymbol{L}^l_x$ to correct the color bias. By default, we adopt color correction in the pixel domain for simplicity.

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 inspiration from CodeFormer (Zhou et al., 2022), 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 an autoencoder, it is natural to leverage the encoder features of the autoencoder to modulate the corresponding decoder features for further fidelity improvement. Let $\boldsymbol{F}_e$ and $\boldsymbol{F}_d$ be the encoder and decoder features, respectively. We introduce an adjustable coefficient $w \in [0, 1]$ to control the extent of modulation:

$$\boldsymbol{F}_m = \boldsymbol{F}_d + \mathcal{C}(\boldsymbol{F}_e, \boldsymbol{F}_d; \boldsymbol{\theta}) \times w, \tag{6}$$

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.

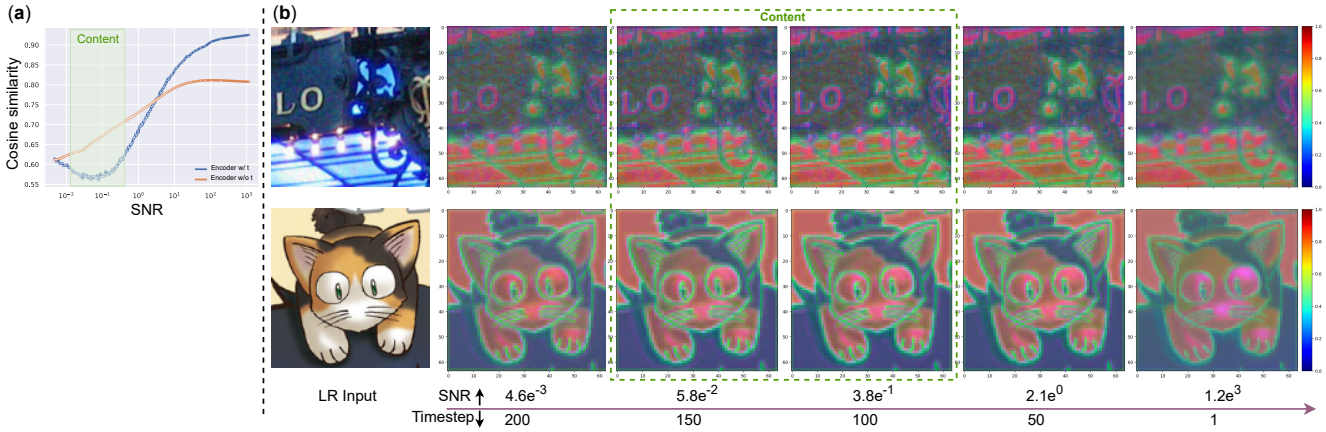

Fig. 3: In contrast to a conditional encoder without time embedding, the one equipped with time embedding can adaptively supply guidance to the pre-trained diffusion models. (a), we gauge the cosine similarity between the diffusion model's features pre- and post-SFT at various timesteps, which echoes 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 in (Choi et al., 2022).

Fig. 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.

3.3 Aggregation Sampling

Due to the heightened sensitivity of the attention layers in Stable Diffusion with respect to the image resolution, it tends to produce inferior outputs for resolutions differing from its training settings, specifically 512×512. This, in effect, constrains the practicality of StableSR.

A common workaround involves splitting the larger image into several overlapping smaller patches and processing each individually. While this strategy often yields good results for conventional CNN-based SR methods, it is not directly applicable to 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 Jiménez (Barbero Jiménez, 2023), 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 $\boldsymbol{F} \in \mathcal{R}^{h \times w}$, which is then subdivided into M overlapping small patches $\{\boldsymbol{F}_{\Omega_n}\}_{n=1}^{M}$, each with a resolution of 64×64 - matching the training resolution[1]. Here, Ω_n is the coordinate set of the nth patch in $\boldsymbol{F}$. 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 $\boldsymbol{w}_{\Omega_n} \in \mathcal{R}^{h \times w}$ whose entries follow up a Gaussian filter in Ω_n and 0 elsewhere is generated for each patch $\boldsymbol{F}_{\Omega_n}$. Overlapping pixels are then weighted in accordance with their respective Gaussian weight maps. In particular, we 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_n. This procedure is reiterated until the final iteration is reached.

Given the output of each patch as $\epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}_{\Omega_n}^{(t)}, \boldsymbol{F}_{\Omega_n}, t)$, where $\boldsymbol{Z}_{\Omega_n}^{(t)}$ is the nth patch of the noisy input $\boldsymbol{Z}^{(t)}$ and $\boldsymbol{\theta}$ is the parameters of the diffusion model, the results of all the patches aggregated together can be formulated

[1] The downsampling scale factor of the autoencoder in Stable Diffusion is 8×.

as follows:

$$\epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, t) = \sum_{n=1}^{M} \frac{\boldsymbol{w}_{\Omega_n}}{\hat{\boldsymbol{w}}} \odot f\left(\epsilon_{\boldsymbol{\theta}}\left(\boldsymbol{Z}_{\Omega_n}^{(t)}, \boldsymbol{F}_{\Omega_n}, t\right)\right), \quad (7)$$

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 Sampler$(\boldsymbol{Z}^{(t)}, \epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, t))$, in the diffusion model. Subsequently, we re-split $\boldsymbol{Z}^{(t-1)}$ into overlapped patches and repeat the above steps until $t = 1$. The whole process is summed up in Algorithm 1. Our experiments suggest that this progressive aggregation method substantially mitigates discrepancies in the overlapped regions, as depicted in Fig. 4.

Algorithm 1 Progressive Patch Aggregation

Require: Cropped Regions $\{\Omega_n\}_{n=1}^{M}$, diffusion steps T, LR latent features $\boldsymbol{F}$.
1: Initialize $\boldsymbol{w}_{\Omega_n}$ and $\hat{\boldsymbol{w}}$
2: $\boldsymbol{Z}^{(T)} \sim \mathcal{N}(0, \mathbb{I})$
3: **for** $t \in [T, \ldots, 0]$ **do**
4: **for** $n \in [1, \ldots, M]$ **do**
5: Compute $\epsilon_{\boldsymbol{\theta}}\left(\boldsymbol{Z}_{\Omega_n}^{(t)}, \boldsymbol{F}_{\Omega_n}, t\right)$
6: **end for**
7: Compute $\epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, t)$ following Eq. (7)
8: $\boldsymbol{Z}^{(t-1)} = \text{Sampler}(\boldsymbol{Z}^{(t)}, \epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, t))$
9: **end for**
10: **return** $\boldsymbol{Z}_0$

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 (Kingma and Ba, 2014) 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 (Ho et al., 2020) 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 such that the shorter side has a length of 512 and rescale the results back to target resolutions after generation.

To train CFW, we first generate 100k synthetic LR-HR pairs with 512×512 resolution following the degradation pipeline in Real-ESRGAN (Wang et al., 2021c). 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 the autoencoder used in LDM (Rombach et al., 2022), 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 (Wang et al., 2021c) to synthesize LR/HR pairs on DIV2K (Agustsson and Timofte, 2017), DIV8K (Gu et al., 2019), Flickr2K (Timofte et al., 2017) and OutdoorSceneTraining (Wang et al., 2018a) datasets. We additionally add 5000 face images from the FFHQ dataset (Karras et al., 2019) 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 (Wang et al., 2021c) and generate 3k LR-HR pairs from DIV2K validation set (Agustsson and Timofte, 2017). 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 (Cai et al., 2019), DRealSR (Wei et al., 2020) and DPED-iPhone (Ignatov et al., 2017). 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] (Ji et al., 2020), BSRGAN (Zhang et al., 2021a), Real-ESRGAN+ (Wang et al., 2021c), DASR (Liang et al., 2022), FeMaSR (Chen et al., 2022), latent diffusion model (LDM) (Rombach et al., 2022), SwinIR-GAN[5] (Liang et al., 2021), and DeepFloyd IF_III (Deep-floyd, 2023). 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

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

[3] SR3 (Saharia et al., 2022b) is not included since its official code is unavailable.

[4] We use the latest official model DF2K-JPEG.

[5] We use the latest official GAN-based SwinIR SR model, *i.e.*, 003_realSR_BSRGAN_DFOWMFC_s64w8_SwinIR-L_x4_GAN.pth.

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	BSRGAN	DASR	Real-ESRGAN+	FeMaSR	LDM	SwinIR-GAN	IF_III	**StableSR**
DIV2K Valid	PSNR ↑	**24.62**	24.58	24.47	24.29	23.06	23.32	23.93	23.36	23.26
	SSIM ↑	0.5970	0.6269	0.6304	**0.6372**	0.5887	0.5762	0.6285	0.5636	0.5726
	LPIPS ↓	0.5276	0.3351	0.3543	**0.3112**	0.3126	0.3199	0.3160	0.4641	0.3114
	FID ↓	49.49	44.22	49.16	37.64	35.87	26.47	36.34	37.54	**24.44**
	CLIP-IQA ↑	0.3534	0.5246	0.5036	0.5276	0.5998	0.6245	0.5338	0.3980	**0.6771**
	MUSIQ ↑	28.57	61.19	55.19	61.05	60.83	62.27	60.22	43.71	**65.92**
RealSR	PSNR ↑	**27.30**	26.38	27.02	25.69	25.06	25.46	26.31	25.47	24.65
	SSIM ↑	0.7579	0.7651	0.7707	0.7614	0.7356	0.7145	**0.7729**	0.7067	0.7080
	LPIPS ↓	0.3570	0.2656	0.3134	0.2709	0.2937	0.3159	**0.2539**	0.3462	0.3002
	CLIP-IQA ↑	0.3687	0.5114	0.3198	0.4495	0.5406	0.5688	0.4360	0.3482	**0.6234**
	MUSIQ ↑	38.26	63.28	41.21	60.36	59.06	58.90	58.70	41.71	**65.88**
DRealSR	PSNR ↑	**30.19**	28.70	29.75	28.62	26.87	27.88	28.50	28.66	28.03
	SSIM ↑	0.8148	0.8028	**0.8262**	0.8052	0.7569	0.7448	0.8043	0.7860	0.7536
	LPIPS ↓	0.3938	0.2858	0.3099	0.2818	0.3157	0.3379	**0.2743**	0.3853	0.3284
	CLIP-IQA ↑	0.3744	0.5091	0.3813	0.4515	0.5634	0.5756	0.4447	0.2925	**0.6357**
	MUSIQ ↑	26.93	57.16	42.41	54.26	53.71	53.72	52.74	30.71	**58.51**
DPED-iphone	CLIP-IQA ↑	0.4496	0.4021	0.2826	0.3389	**0.5306**	0.4482	0.3373	0.2962	0.4799
	MUSIQ ↑	45.60	45.89	32.68	42.42	49.95	44.23	43.30	37.49	**50.48**

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 (Cai et al., 2019; Wei et al., 2020; Ignatov et al., 2017), 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[6](Zhang et al., 2018a), FID (Heusel et al., 2017), CLIP-IQA (Wang et al., 2023) and MUSIQ (Ke et al., 2021) 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 (Ignatov et al., 2017), we follow existing methods (Wang et al., 2021c; Chen et al., 2022) 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 (Cai et al., 2019; Wei et al., 2020), suggesting the superiority of StableSR. Note that although Real-ESRGAN+ and SwinIR-GAN achieve good LPIPS scores, they show inferior performance on other perceptual metrics, i.e., FID, CLIP-IQA and MUSIQ. Moreover, existing methods 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 (Cai et al., 2019; Wei et al., 2020) and the internet in Fig. 5 and Fig. 6. 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). Moreover, as shown in the fourth row of Fig. 5, StableSR generates sharp edges without obvious degradations, whereas other state-of-the-art methods generate blurry results. Figure 6 further demonstrates the superiority of StableSR on images beyond 512×512.

User Study. To further examine the effectiveness of StableSR, we conduct a user study on 16 real-world

[6] We use LPIPS-ALEX by default.

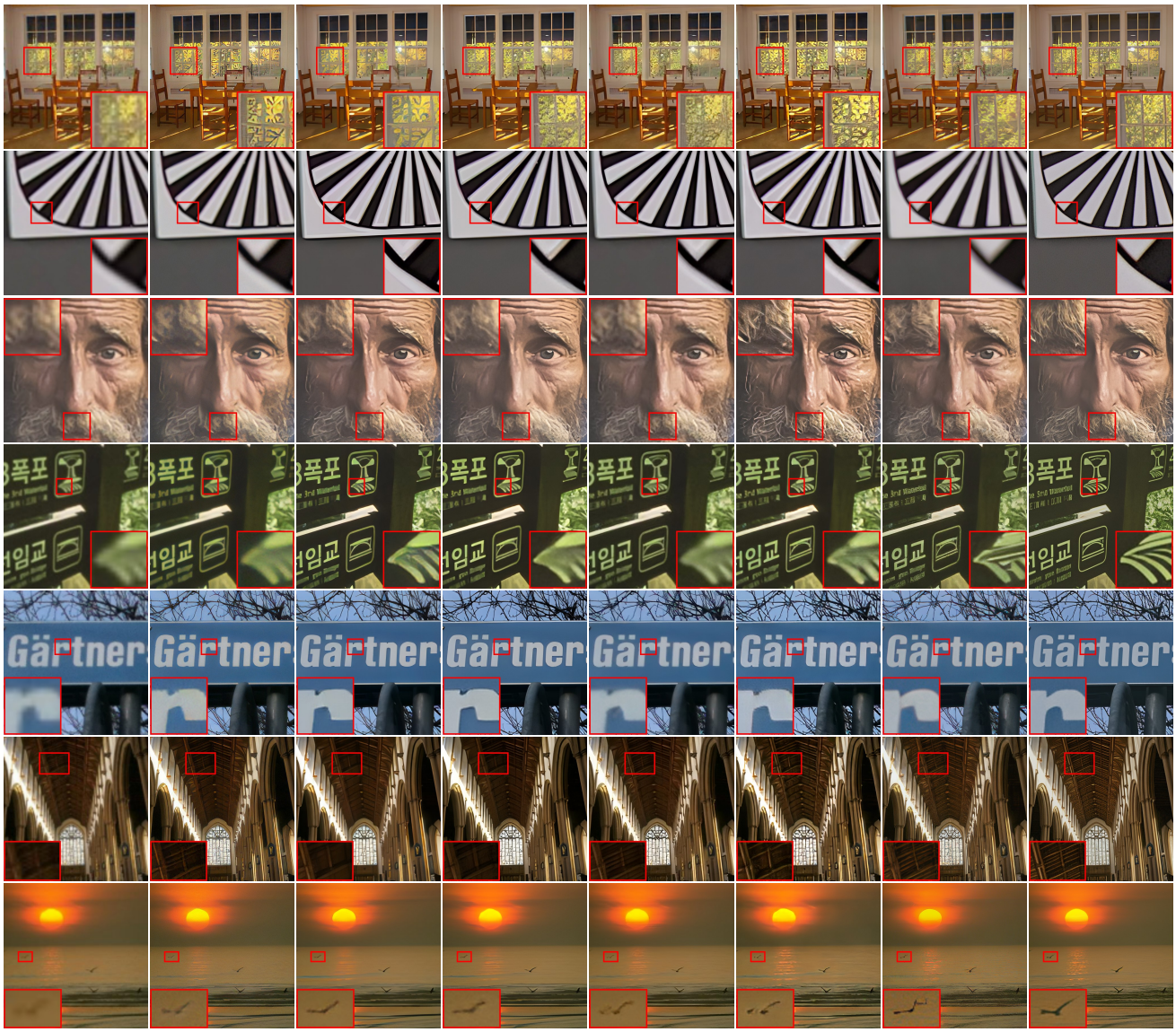

Fig. 5: Qualitative comparisons on several representative real-world samples (128 → 512). Our StableSR is capable of removing artifacts and generating realistic details. (**Zoom in for details**)

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 compared methods, there are 3×16 pairs evaluated by 25 subjects, resulting in 3×400 votes in total. As depicted in Fig. 7, StableSR outperforms all three competitive methods by a large margin, gaining over 75% of the votes all the time.

Comparison with Concurrent Diffusion Applications. We notice that recent concurrent works (Zhang et al., 2023; Deep-floyd, 2023) can also be adopted for image SR. Here, we further conduct comparisons with these methods on real-world images. For fair comparisons, we use DDIM sampling with eta 1.0 and timestep 200 for all the methods, and the seed is fixed to 42. We further set $w = 0.0$ in StableSR to avoid additional improvement due to CFW. For ControlNet-tile (Zhang et al., 2023), we generate additional prompts using stable-diffusion-webui[7] for better performance. For

[7] https://github.com/AUTOMATIC1111/stable-diffusion-webui

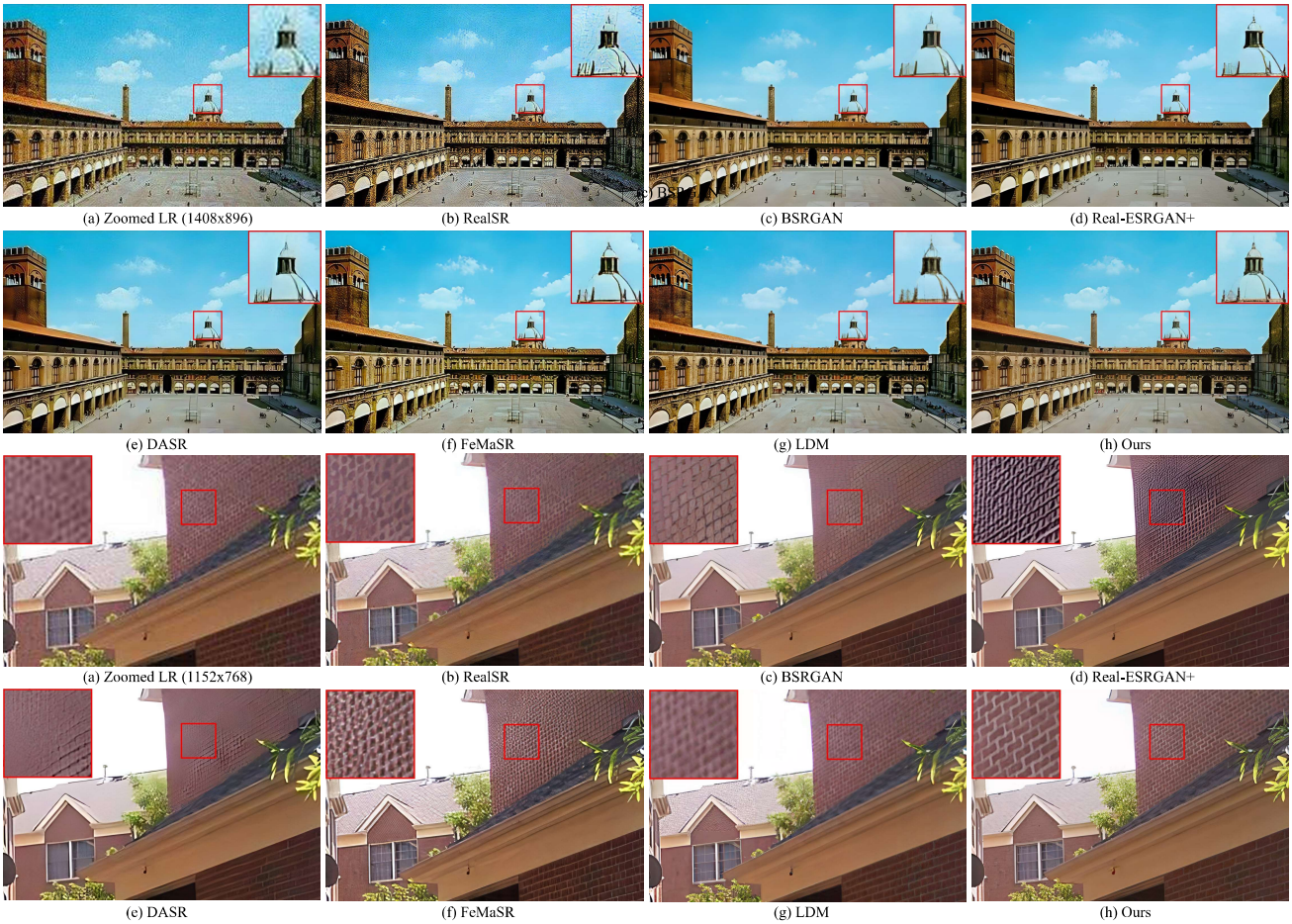

Fig. 6: Qualitative comparisons on real-world images with diverse resolutions beyond 512×512. Our StableSR still outperforms other methods with more vivid details and less annoying artifacts. (**Zoom in for details**)

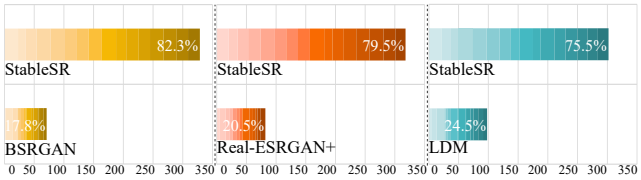

Fig. 7: User study on 16 real-world images evaluated by 25 subjects. Given one LR image, the subjects are asked to choose the better HR image generated from either StableSR or BSRGAN/Real-ESRGAN+/LDM. The proposed StableSR outperforms other methods by gaining over 75% of the votes.

IF_III upscaler (Deep-floyd, 2023), we follow official examples to set noise level to 100 w/o prompts. As shown in Fig. 8, ControlNet-tile shows poor fidelity due to the lack of specific designs for SR. Compared with IF_III upscaler, the proposed StableSR is capable of generating more faithful details with sharper edges, *e.g.*, the text in the first row, the tiger's nose in the third row and the wing of the butterfly in the last row of Fig. 8. Note that IF_III upscaler is trained from scratch, which requires significant computational resources. The visual comparisons suggest the superiority of StableSR.

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 Fig. 3, 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 (Cai et al., 2019; Wei et al., 2020). 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 comparisons in Fig. 9 also indicate inferior performance without the above two components, suggesting the effectiveness of time-aware guidance and color

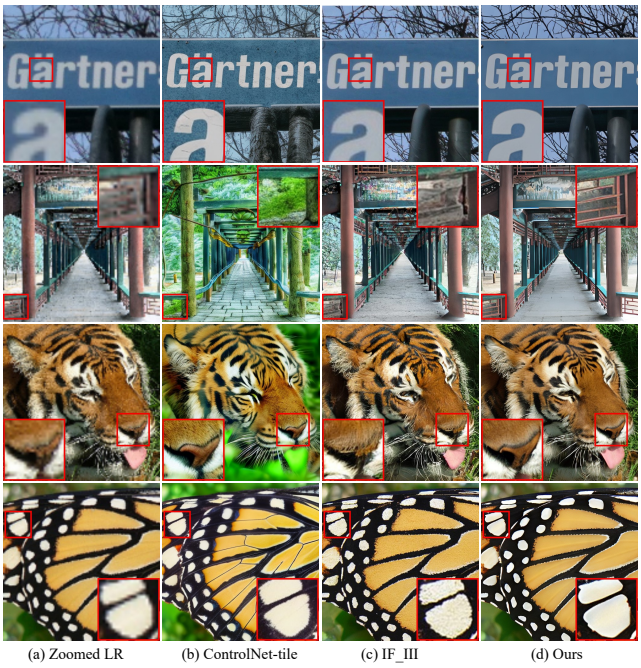

Fig. 8: Qualitative comparisons on real-world images (128 → 512). Our StableSR outperforms ControlNet-tile (Zhang et al., 2023) with higher fidelity and has more realistic and sharper details compared with IF_III upscaler (Deep-floyd, 2023). (**Zoom in for details**)

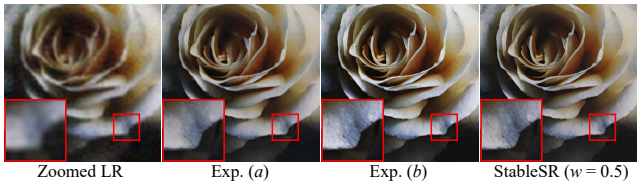

Fig. 9: Visual comparisons of time-aware guidance and color correction. Exp. (a) does not apply time-aware guidance, leading to blurry textures. Exp. (b) applies time-aware guidance and can generate sharper details, but obvious color shifts can be observed. With both strategies, StableSR generates sharp textures and avoids color shifts.

Table 2: Ablation studies of time-aware guidance and color correction on RealSR (Cai et al., 2019) and DRealSR (Wei et al., 2020) benchmarks.

Exp.	Strategies			RealSR / DRealSR		
	Time aware	Pixel Color cor.	Wavelet 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
(c)	✓		✓	23.38 / 26.80	0.6870 / 0.7235	0.3157 / 0.3475
Default	✓	✓		**24.65** / **28.03**	**0.7080** / **0.7536**	**0.3002** / **0.3284**

correction. In addition to directly adopting color correction in the pixel domain, our proposed wavelet color correction can further boost the visual quality, as shown in Fig. 10, which may further facilitate the practical use.

Table 3: Ablation studies of the controllable coefficient w on both synthetic (DIV2K Valid (Agustsson and Timofte, 2017)) and real-world (RealSR (Cai et al., 2019), DRealSR (Wei et al., 2020), and DPED-iPhone (Ignatov et al., 2017)) benchmarks.

Datasets	Metrics	StableSR ($w = 0.0$)	**StableSR** ($w = 0.5$)	StableSR ($w = 0.75$)	StableSR ($w = 1.0$)
DIV2K Valid	PSNR ↑	22.68	23.26	**24.17**	23.14
	SSIM ↑	0.5546	0.5726	**0.6209**	0.5681
	LPIPS ↓	0.3393	0.3114	**0.3003**	0.3077
	FID ↓	25.83	24.44	**24.05**	26.14
	CLIP-IQA ↑	0.6529	**0.6771**	0.5519	0.6197
	MUSIQ ↑	65.72	**65.92**	59.46	64.31
RealSR	PSNR ↑	24.07	24.65	**25.37**	24.70
	SSIM ↑	0.6829	0.7080	**0.7435**	0.7157
	LPIPS ↓	0.3190	0.3002	**0.2672**	0.2892
	CLIP-IQA ↑	0.6127	**0.6234**	0.5341	0.5847
	MUSIQ ↑	65.81	**65.88**	62.36	64.05
DRealSR	PSNR ↑	27.43	28.03	**29.00**	27.97
	SSIM ↑	0.7341	0.7536	**0.7985**	0.7540
	LPIPS ↓	0.3595	0.3284	**0.2721**	0.3080
	CLIP-IQA ↑	0.6340	**0.6357**	0.5070	0.5893
	MUSIQ ↑	**58.98**	58.51	53.12	56.77
DPED-iPhone	CLIP-IQA ↑	**0.5015**	0.4799	0.3405	0.4250
	MUSIQ ↑	**51.90**	50.48	41.81	47.96

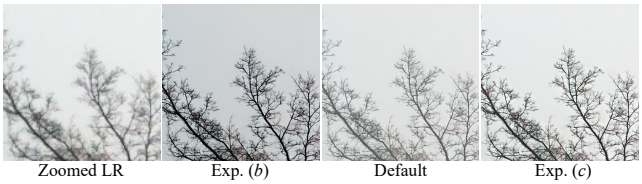

Fig. 10: Visual comparisons of different color correction strategies. With no color correction, obvious color shifts can be observed in Exp. (b). Our color correction via channel matching in Eq. (2) by default can alleviate the color shift problem, while the wavelet color correction of Eq. (5) can further improve the visual quality.

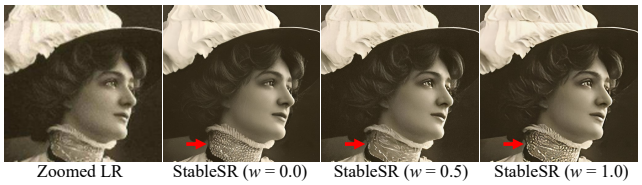

Fig. 11: Visual comparisons with different coefficients w for CFW module. It is observed that a small w tends to generate a realistic result while a larger w improves the fidelity.

Flexibility of Fidelity-realism Trade-off. Our CFW module inspired by CodeFormer (Zhou et al., 2022) allows a flexible realism-fidelity trade-off. In particular, 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 3, compared with StableSR ($w = 0.0$), StableSR with larger values of w (e.g., 0.75) achieves higher PSNR and SSIM on all three paired benchmarks, indicating better fidelity. In contrast, Sta-

Table 4: Complexity comparison of model parameters and running time. All methods are evaluated on 128×128 input images for 4x SR using an NVIDIA Tesla 32G-V100 GPU. For diffusion models, the sampling steps are set to 200. The runtime is averaged by ten runs with a batch size of 1.

	Real-ESRGAN+	FeMaSR	LDM	SwinIR-GAN	IF_III	**StableSR**
Model type	GAN	VQGAN	LDM	GAN	LDM	LDM
Runtime	0.08s	0.12s	5.25s	0.31s	17.78s	15.16s
Trainable Params	16.70M	28.29M	113.62M	28.01M	473.40M	149.91M

bleSR ($w = 0.0$) achieves better perceptual quality with higher CLIP-IQA scores and MUSIQ scores. Similar phenomena can also be observed in Fig. 11. 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 3. Hence, we set the coefficient w to 0.5 by default for trading between quality and fidelity.

4.5 Complexity Comparison

StableSR is a diffusion-based approach and requires multi-step sampling for image generation. As shown in Table 4, when the number of sampling steps is set to 200, StableSR needs 15.16 seconds to generate a 512×512 image on one NVIDIA Tesla 32G-V100 GPU. This is comparable to IF_III upscaler but slower than GAN-based SR methods such as Real-ESRGAN+ and SwinIR-GAN, which require only a single forward pass. Fast sampling strategy (Song et al., 2020; Lu et al., 2022; Karras et al., 2022) and model distillation (Salimans and Ho, 2021; Song et al., 2023b; Luo et al., 2023) are two promising solutions to improve efficiency. Another viable remedy is to shorten the chain of diffusion process (Yue et al., 2023). As for trainable parameters, StableSR has 149.91M trainable parameters, which is only 11.50% of the full model and less than IF_III, *i.e.*, 473.40M. The trainable parameters can be further decreased with more careful design, *e.g.*, adopting lightweight architectures (Chollet, 2017; Howard et al., 2019) or network pruning (Fang et al., 2023). Such exploration is beyond the scope of this paper.

5 Inference Strategies

The proposed StableSR already demonstrates superior performance quantitatively and qualitatively on both synthetic and real-world benchmarks, as shown in Sec. 4. Here, we discuss several effective strategies during the sampling process that can further boost the visual quality of the generated results without additional finetuning.

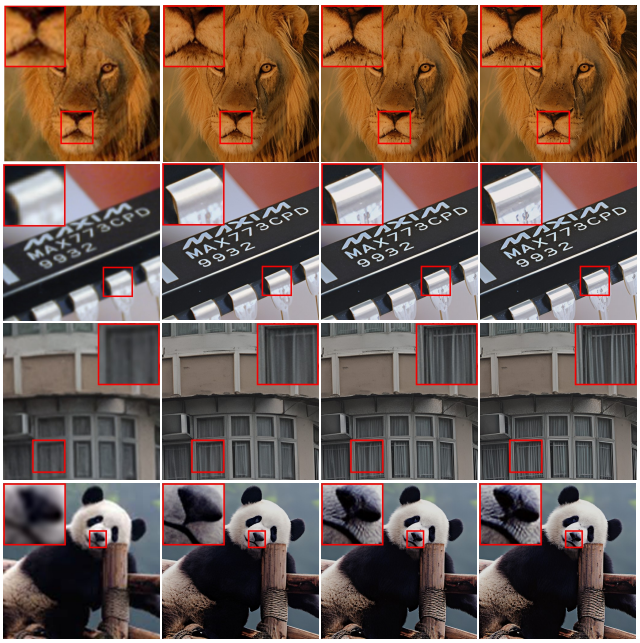

Fig. 12: Qualitative comparisons on classifier-free guidance with negative prompts. Higher guidance scale s leads to sharper edges. (**Zoom in for details**)

5.1 Classifier-free Guidance with Negative Prompts

The default StableSR is trained with null prompts. Interestingly, we observe that StableSR can react to prompts, especially negative prompts. We examine the use of classifier-free guidance (Ho and Salimans, 2021) with negative prompts to further improve the visual quality during sampling. Given two StableSR models conditioned on null prompts $\epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, [], t)$ and negative prompts $\epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, \boldsymbol{c}, t)$, respectively, the new sampling process can be performed using a linear combination of the estimations with a guidance scale s:

$$\tilde{\epsilon}_{\boldsymbol{\theta}} = \epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, \boldsymbol{c}, t) + s\left(\epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, [], t) - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{Z}^{(t)}, \boldsymbol{F}, \boldsymbol{c}, t)\right), \quad (8)$$

where $\boldsymbol{c}$ is the negative prompt for guidance. According to Eq. (8), it is worth noting that $s = 0$ is equivalent to directly using negative prompts without guidance, and $s = 1$ is equivalent to our default settings with the null prompt.

We compare the performance of StableSR with various positive prompts, *i.e.*, (1) "`(masterpiece:2), (best quality:2), (realistic:2), (very clear:2)`", and (2) "`Good photo.`", and negative

Table 5: Comparison of different prompts and guidance strengths. Note that $s = 0$ is equivalent to directly using negative prompts w/o guidance. Positive prompts are (1) "`(masterpiece:2), (best quality:2), (realistic:2), (very clear:2)`", and (2) "`Good photo.`". Negative prompts are (a) "`3d, cartoon, anime, sketches, (worst quality:2), (low quality:2)`", and (b) "`Bad photo.`". The first row is the default settings for StableSR.

Strategies			RealSR / DRealSR				
Pos. Prompts	Neg. Prompts	Guidance Scale	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-IQA ↑	MUSIQ ↑
[]	-	-	24.65 / 28.03	0.7080 / 0.7536	0.3002 / 0.3284	0.6234 / 0.6357	65.88 / 58.51
(1)	-	-	24.68 / 28.03	0.7025 / 0.7461	0.3151 / 0.3378	0.6251 / 0.6370	65.34 / 58.07
(2)	-	-	24.71 / 28.07	0.7049 / 0.7500	0.3118 / 0.3333	0.6219 / 0.6291	65.22 / 57.75
[]	(a)	$s = 0.0$	24.80 / 28.18	0.7097 / 0.7562	0.3105 / 0.3316	0.6176 / 0.6224	64.86 / 57.31
		$s = 2.5$	24.41 / 27.76	0.6972 / 0.7383	0.3168 / 0.3417	0.6306 / 0.6422	66.02 / 59.21
		$s = 5.0$	23.96 / 27.21	0.6829 / 0.7188	0.3267 / 0.3583	0.6356 / 0.6558	66.84 / 61.07
		$s = 7.5$	23.53 / 26.68	0.6673 / 0.7003	0.3399 / 0.3774	0.6323 / 0.6621	67.26 / 62.41
[]	(b)	$s = 0.0$	24.77 / 28.13	0.7067 / 0.7520	0.3100 / 0.3317	0.6184 / 0.6239	64.81 / 57.27
		$s = 2.5$	24.46 / 27.90	0.7017 / 0.7467	0.3170 / 0.3371	0.6303 / 0.6409	66.29 / 58.97
		$s = 5.0$	24.13 / 27.61	0.6958 / 0.7391	0.3240 / 0.3467	0.6377 / 0.6490	67.43 / 60.69
		$s = 7.5$	23.78 / 27.30	0.6894 / 0.7310	0.3320 / 0.3578	0.6421 / 0.6583	68.13 / 62.12

prompts, *i.e.*, (a) "`3d, cartoon, anime, sketches, (worst quality:2), (low quality:2)`", and (b) "`Bad photo.`". As shown in Table 5, different prompts lead to different metric scores. Specifically, the classifier-free guidance with negative prompts shows a significant influence on the metrics, *i.e.*, higher guidance scales lead to higher CLIP-IQA and MUSIQ scores, indicating sharper results. Similar phenomena can also be observed in Fig. 12. However, an overly strong guidance, *e.g.*, $s = 7.5$ can result in oversharpening.

5.2 Pre-cleaning for Severe Degradations

It is observed that StableSR may yield suboptimal results when LR images are severely degraded with pronounced levels of blur or noise, as shown in the first column of Fig. 13. Drawing inspiration from RealBasicVSR (Chan et al., 2022b), we incorporate an auxiliary pre-cleaning phase preceding StableSR to address scenarios under severe degradations. Specifically, we first adopt an existing SR approach *e.g.*, RealESRGAN+ (Wang et al., 2021c) for general SR and CodeFormer (Zhou et al., 2022) for face SR[8] to mitigate the aforementioned severe degradations. To suppress the amplification of artifacts originating from the pre-cleaning phase, a subsequent 2× bicubic downsampling operation is further adopted after pre-cleaning. Subsequently, StableSR is used to generate the final outputs. As shown in Fig. 13, such a pre-cleaning stage substantially improves the robustness of StableSR.

[8] For face SR, we further finetune our StableSR model for 50 epochs on FFHQ (Karras et al., 2019) using the same degradations as CodeFormer (Zhou et al., 2022).

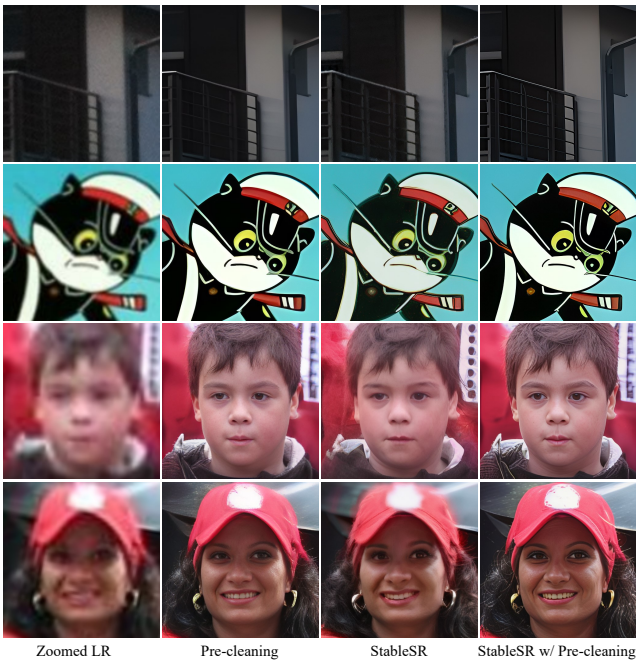

Fig. 13: StableSR may generate suboptimal results when the inputs have severe degradations. Adopting a simple pre-cleaning with a pre-trained SR model during sampling can effectively improve the performance of StableSR under such circumstances. (**Zoom in for details**)

6 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.

Acknowledgement: This study is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2022-01-033[T]), RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). We sincerely thank Yi Li for providing valuable advice and building the WebUI implementation[9] of our work.

Appendix A Details of Time-aware Encoder

As mentioned in the main paper, the architecture of the time-aware encoder is similar to the contracting path of the denoising U-Net in Stable Diffusion (Rombach et al., 2022) with much fewer parameters (∼105M, including SFT layers) by reducing the number of channels. The detailed settings are listed in Table 6.

Table 6: 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

Appendix B 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. 14. 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 (Zhang et al., 2021b)/Real-ESRGAN+ (Wang et al., 2021c)/LDM (Rombach et al., 2022). There are 48 questions in total with random orders for each subject.

[9] https://github.com/pkuliyi2015/sd-webui-stablesr

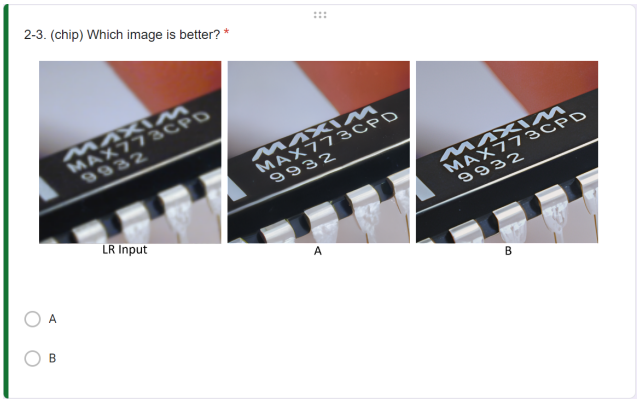

Fig. 14: Screenshot of the user interface in the user study.

Appendix C Additional Visual Results

C.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 obtain LR images with 128 × 128 resolution. As shown in Fig. 15, StableSR successfully produces outputs with finer details and sharper edges, significantly outperforming state-of-the-art methods.

C.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. 16, StableSR is capable of generating high-quality SR images beyond 4x resolution, indicating its practical use in real-world applications. Moreover, the results in Fig. 17 indicate that 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.

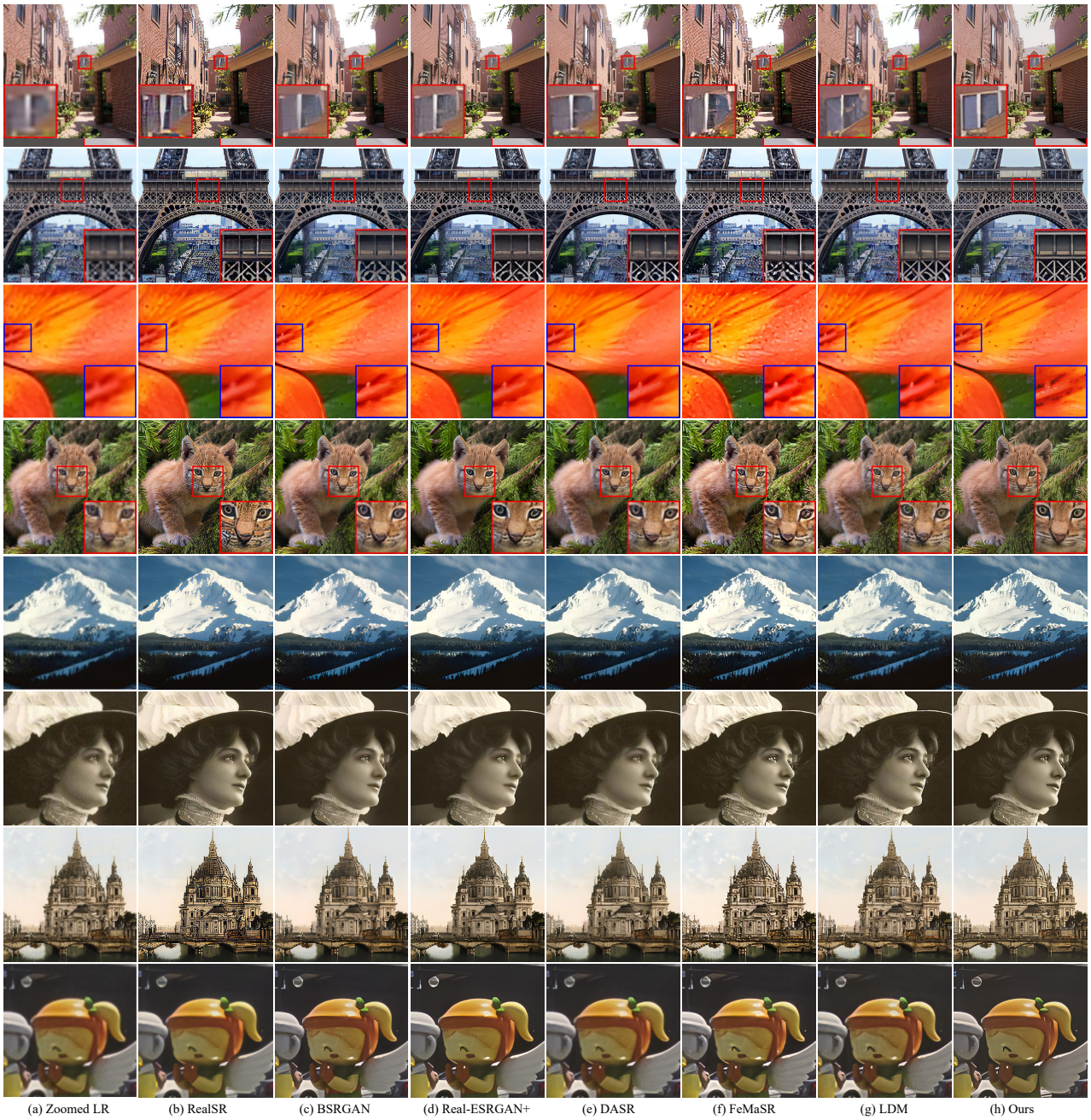

Fig. 15: More qualitative comparisons on real-world images (128 → 512). While existing methods typically fail to restore realistic textures under complicated degradations, our StableSR outperforms these methods by a large margin. (**Zoom in for details**)

References

Agustsson E, Timofte R (2017) 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)

Avrahami O, Lischinski D, Fried O (2022) Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 18208–18218

Balaji Y, Nah S, Huang X, Vahdat A, Song J, Kreis K, Aittala M, Aila T, Laine S, Catanzaro B, Karras T,

Fig. 16: A 4x StableSR result on AIGC content beyond 4K resolution. (**Zoom in for details**)

Liu MY (2022) ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. arXiv preprint arXiv:221101324

Barbero Jiménez Á (2023) Mixture of diffusers for scene composition and high resolution image generation. arXiv e-prints pp arXiv–2302

Cai J, Zeng H, Yong H, Cao Z, Zhang L (2019) 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)

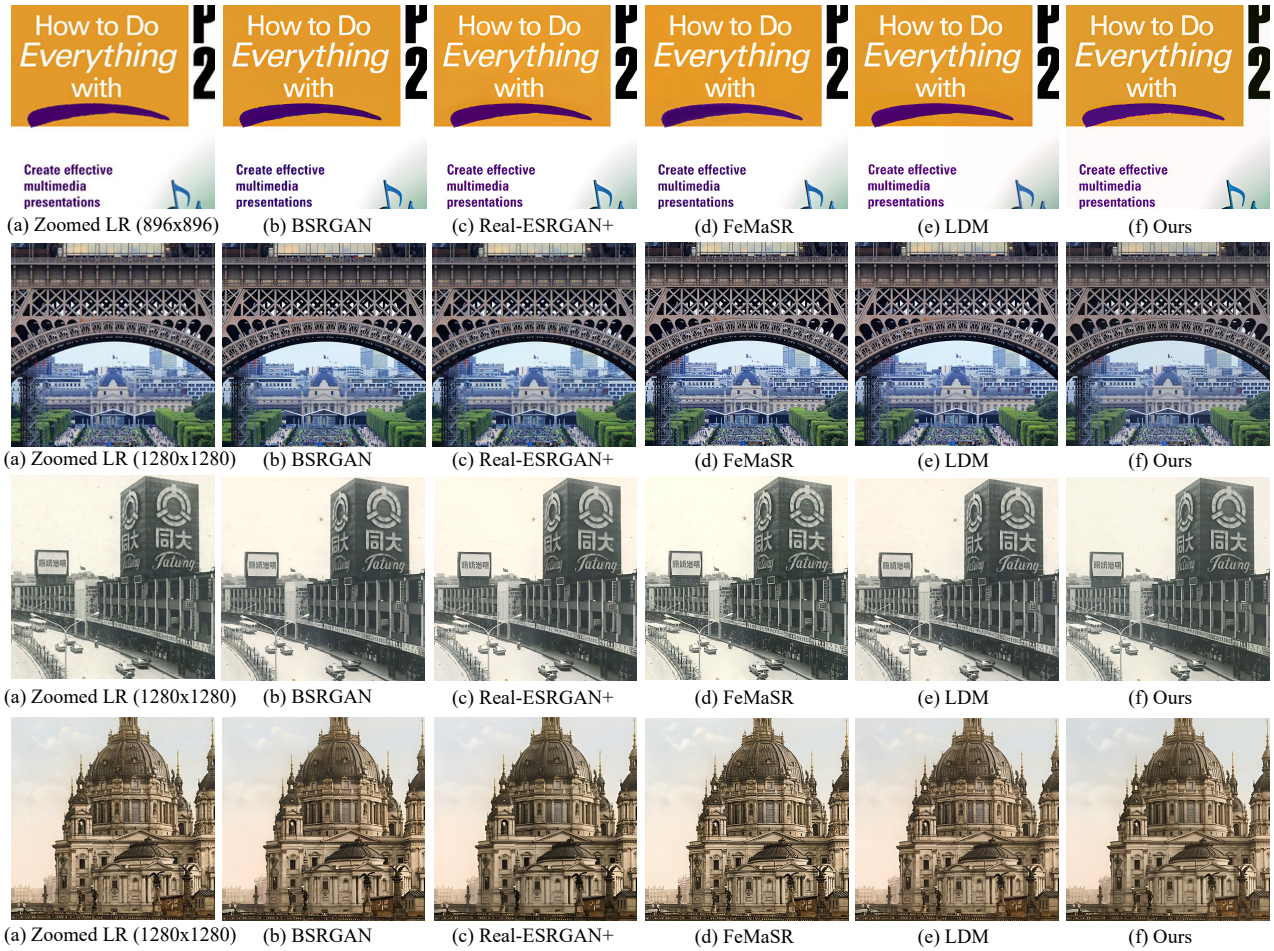

Fig. 17: More qualitative comparisons on original real-world images with diverse resolutions. Our StableSR is capable of generating vivid details without annoying artifacts. (**Zoom in for details**)

Chan KC, Wang X, Xu X, Gu J, Loy CC (2021) GLEAN: Generative latent bank for large-factor image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Chan KC, Wang X, Xu X, Gu J, Loy CC (2022a) GLEAN: Generative latent bank for large-factor image super-resolution and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

Chan KC, Zhou S, Xu X, Loy CC (2022b) Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Chen C, Shi X, Qin Y, Li X, Han X, Yang T, Guo S (2022) Real-world blind super-resolution via feature matching with implicit high-resolution priors. In: Proceedings of the ACM International Conference on Multimedia (ACM MM)

Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, Ma S, Xu C, Xu C, Gao W (2021) Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Choi J, Kim S, Jeong Y, Gwon Y, Yoon S (2021) Ilvr: Conditioning method for denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Choi J, Lee J, Shin C, Kim S, Kim H, Yoon S (2022) Perception prioritized training of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 11472–11481

Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Chung H, Sim B, Ryu D, Ye JC (2022) Improving diffusion models for inverse problems using manifold constraints. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Dai T, Cai J, Zhang Y, Xia ST, Zhang L (2019) Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Deep-floyd (2023) If. https://github.com/deep-floyd/IF

Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV)

Dong C, Loy CC, He K, Tang X (2015) Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: Proceedings of the European Conference on Computer Vision (ECCV)

Fang G, Ma X, Wang X (2023) Structural pruning for diffusion models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Feng W, He X, Fu TJ, Jampani V, Akula A, Narayana P, Basu S, Wang XE, Wang WY (2023) Training-free structured diffusion guidance for compositional text-to-image synthesis. Proceedings of International Conference on Learning Representations (ICLR)

Fritsche M, Gu S, Timofte R (2019) Frequency separation for real-world super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV-W)

Gal R, Arar M, Atzmon Y, Bermano AH, Chechik G, Cohen-Or D (2023) Designing an encoder for fast personalization of text-to-image models. arXiv preprint arXiv:230212228

Gu J, Shen Y, Zhou B (2020) Image processing using multi-code gan prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Gu S, Lugmayr A, Danelljan M, Fritsche M, Lamour J, Timofte R (2019) Div8k: Diverse 8k resolution image dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV-W)

Gu S, Chen D, Bao J, Wen F, Zhang B, Chen D, Yuan L, Guo B (2022) Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10696–10706

He X, Mo Z, Wang P, Liu Y, Yang M, Cheng J (2019) Ode-inspired network design for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Hertz A, Mokady R, Tenenbaum J, Aberman K, Pritch Y, Cohen-Or D (2022) Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:220801626

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Ho J, Salimans T (2021) Classifier-free diffusion guidance. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), vol 33, pp 6840–6851

Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, et al. (2019) Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Hu EJ, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W, et al. (2022) Lora: Low-rank adaptation of large language models. In: Proceedings of International Conference on Learning Representations (ICLR)

Ignatov A, Kobyshev N, Timofte R, Vanhoey K, Van Gool L (2017) Dslr-quality photos on mobile devices with deep convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Ji X, Cao Y, Tai Y, Wang C, Li J, Huang F (2020) Real-world super-resolution via kernel estimation and noise injection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (CVPR-W)

Jiang Y, Chan KC, Wang X, Loy CC, Liu Z (2021) Robust reference-based super-resolution via c2-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Karras T, Aittala M, Aila T, Laine S (2022) Elucidating the design space of diffusion-based generative models. In: Proceedings of Advances in Neural Information

Processing Systems (NeurIPS)
Ke J, Wang Q, Wang Y, Milanfar P, Yang F (2021) Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al. (2017) 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)
Li H, Yang Y, Chang M, Chen S, Feng H, Xu Z, Li Q, Chen Y (2022) SRDiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing 479:47–59
Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R (2021) SwinIR: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV-W)
Liang J, Zeng H, Zhang L (2022) Efficient and degradation-adaptive network for real-world image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV)
Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
Lu C, Zhou Y, Bao F, Chen J, Li C, Zhu J (2022) Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)
Luo S, Tan Y, Huang L, Li J, Zhao H (2023) Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:231004378
Maeda S (2020) Unpaired image super-resolution using pseudo-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
Meng X, Kabashima Y (2022) Diffusion model based posterior sampling for noisy linear inverse problems. arXiv preprint arXiv:221112343
Menon S, Damian A, Hu S, Ravi N, Rudin C (2020) 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)
Molad E, Horwitz E, Valevski D, Acha AR, Matias Y, Pritch Y, Leviathan Y, Hoshen Y (2023) Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:230201329
Mou C, Wang X, Xie L, Zhang J, Qi Z, Shan Y, Qie X (2023) T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:230208453
Nichol AQ, Dhariwal P, Ramesh A, Shyam P, Mishkin P, Mcgrew B, Sutskever I, Chen M (2022) Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: Proceedings of International Conference on Machine Learning (ICML)
Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:180703748
Pan X, Zhan X, Dai B, Lin D, Loy CC, Luo P (2021) Exploiting deep generative prior for versatile image restoration and manipulation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
Qi C, Cun X, Zhang Y, Lei C, Wang X, Shan Y, Chen Q (2023) Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:230309535
Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I (2021) Zero-shot text-to-image generation. In: Proceedings of International Conference on Machine Learning (ICML)
Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:220406125
Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, pp 234–241
Sahak H, Watson D, Saharia C, Fleet D (2023) Denoising diffusion probabilistic models for robust image super-resolution in the wild. arXiv preprint arXiv:230207864
Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, Ghasemipour SKS, Gontijo-Lopes R, Ayan BK, Salimans T, et al. (2022a) Photorealistic text-to-image diffusion models with deep language understanding. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Saharia C, Ho J, Chan W, Salimans T, Fleet DJ, Norouzi M (2022b) Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

Sajjadi MS, Scholkopf B, Hirsch M (2017) Enhancenet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Salimans T, Ho J (2021) Progressive distillation for fast sampling of diffusion models. In: Proceedings of International Conference on Learning Representations (ICLR)

Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S (2015) Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of International Conference on Machine Learning (ICML)

Song J, Meng C, Ermon S (2020) Denoising diffusion implicit models. In: Proceedings of International Conference on Learning Representations (ICLR)

Song J, Vahdat A, Mardani M, Kautz J (2023a) Pseudoinverse-guided diffusion models for inverse problems. In: Proceedings of International Conference on Learning Representations (ICLR)

Song Y, Dhariwal P, Chen M, Sutskever I (2023b) Consistency models. arXiv preprint arXiv:230301469

Timofte R, Agustsson E, Van Gool L, Yang MH, Zhang L (2017) 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)

Wan Z, Zhang B, Chen D, Zhang P, Chen D, Liao J, Wen F (2020) Bringing old photos back to life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Wang J, Chan KC, Loy CC (2023) Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence

Wang L, Wang Y, Dong X, Xu Q, Yang J, An W, Guo Y (2021a) Unsupervised degradation representation learning for blind super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Wang X, Yu K, Dong C, Loy CC (2018a) 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)

Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Change Loy C (2018b) Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision Workshops (ECCV-W)

Wang X, Li Y, Zhang H, Shan Y (2021b) Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Wang X, Xie L, Dong C, Shan Y (2021c) 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)

Wang Y, Yu J, Zhang J (2022) Zero-shot image restoration using denoising diffusion null-space model. Proceedings of International Conference on Learning Representations (ICLR)

Wei P, Xie Z, Lu H, Zhan Z, Ye Q, Zuo W, Lin L (2020) Component divide-and-conquer for real-world image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV)

Wei Y, Gu S, Li Y, Timofte R, Jin L, Song H (2021) 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)

Wu JZ, Ge Y, Wang X, Lei SW, Gu Y, Hsu W, Shan Y, Qie X, Shou MZ (2022) Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:221211565

Xu X, Sun D, Pan J, Zhang Y, Pfister H, Yang MH (2017) Learning to super-resolve blurry face and text images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Xu X, Ma Y, Sun W (2019) Towards real scene super-resolution with raw images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Yang F, Yang H, Fu J, Lu H, Guo B (2020) Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Yang S, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B (2021a) Score-based generative modeling through stochastic differential equations. In: Proceedings of International Conference on Learning Representations (ICLR)

Yang T, Ren P, Xie X, Zhang L (2021b) 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)

Yu K, Dong C, Lin L, Loy CC (2018) Crafting a toolchain for image restoration by deep reinforcement learning. In: Proceedings of the IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition (CVPR)

Yue Z, Wang J, Loy CC (2023) Resshift: Efficient diffusion model for image super-resolution by residual shifting. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Zhang J, Lu S, Zhan F, Yu Y (2021a) Blind image super-resolution via contrastive representation learning. arXiv preprint arXiv:210700708

Zhang K, Liang J, Van Gool L, Timofte R (2021b) Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Zhang L, Rao A, Agrawala M (2023) Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018a) The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018b) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV)

Zhang Z, Wang Z, Lin Z, Qi H (2019) Image super-resolution by neural texture transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Zhao Y, Su YC, Chu CT, Li Y, Renn M, Zhu Y, Chen C, Jia X (2022) Rethinking deep face restoration. In: cvpr

Zheng H, Ji M, Wang H, Liu Y, Fang L (2018) Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In: Proceedings of the European Conference on Computer Vision (ECCV)

Zhou S, Zhang J, Zuo W, Loy CC (2020) Cross-scale internal graph neural network for image super-resolution. Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Zhou S, Chan KC, Li C, Loy CC (2022) Towards robust blind face restoration with codebook lookup transformer. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)