StructSR: Refuse Spurious Details in Real-World Image Super-Resolution

Yachao Li[1], **Dong Liang**[1*], **Tianyu Ding**[2], **Sheng-Jun Huang**[1]

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211106, China.

[2]Microsoft, Washington, USA

{liyachao, liangdong, huangsj}@nuaa.edu.cn, tianyuding@microsoft.com

Abstract

Diffusion-based models have shown great promise in real-world image super-resolution (Real-ISR), but often generate content with structural errors and spurious texture details due to the empirical priors and illusions of these models. To address this issue, we introduce StructSR, a simple, effective, and plug-and-play method that enhances structural fidelity and suppresses spurious details for diffusion-based Real-ISR. StructSR operates without the need for additional fine-tuning, external model priors, or high-level semantic knowledge. At its core is the Structure-Aware Screening (SAS) mechanism, which identifies the image with the highest structural similarity to the low-resolution (LR) input in the early inference stage, allowing us to leverage it as a historical structure knowledge to suppress the generation of spurious details. By intervening in the diffusion inference process, StructSR seamlessly integrates with existing diffusion-based Real-ISR models. Our experimental results demonstrate that StructSR significantly improves the fidelity of structure and texture, improving the PSNR and SSIM metrics by an average of 5.27% and 9.36% on a synthetic dataset (DIV2K-Val) and 4.13% and 8.64% on two real-world datasets (RealSR and DRealSR) when integrated with four state-of-the-art diffusion-based Real-ISR methods.

Code — https://github.com/LYCEXE/StructSR

Introduction

Image super-resolution aims to reconstruct high-resolution (HR) images from their low-resolution (LR) counterparts. Traditional image super-resolution methods often rely on simplistic assumptions about degradation (e.g., Gaussian noise and bicubic downsampling) and design methods tailored to these degradation models (Chen et al. 2021, 2023a; Dong et al. 2014; Liang et al. 2021). However, their ability to generalize to real-world images with complex degradation is limited. To address the challenges of real-world image super-resolution (Real-ISR), methods (Zhang et al. 2021; Wang et al. 2021) have employed Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) to generate HR images from LR images collected from the real world. However, these methods rely heavily on specific paired training data and generally lack the capability to generate realistic texture details for Real-ISR.

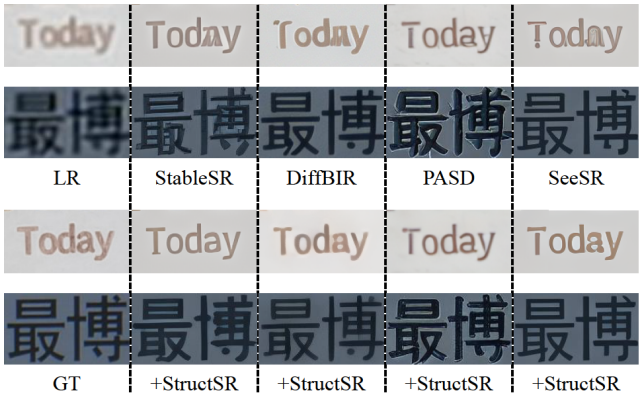


Figure 1: Comparison of diffusion-based Real-ISR methods with and without StructSR integration. The original methods generate spurious details in both English letters and Chinese characters. Integration with StructSR significantly reduces these artifacts, resulting in more accurate reconstruction.

Current diffusion-based Real-ISR methods typically utilize pre-trained text-to-image models, e.g., Stable Diffusion (Rombach et al. 2022) as priors and fine-tune them using real-world image super-resolution datasets. These methods (Wang et al. 2023a; Lin et al. 2023b; Yang et al. 2023; Wu et al. 2024) have shown remarkable proficiency in generating realistic image details. However, they often struggle with maintaining structural fidelity, and the generated details may be spurious with the real semantics due to the empirical priors and illusions of these models, as illustrated in Fig. 1.

We study the changes in the image structure during the inference process of the diffusion-based Real-ISR method by using structural similarity (SSIM) (Wang et al. 2004) between the temporal reconstructed image and the input LR image (Details see Methodology Section). Our analysis reveals that, the temporal reconstructed images with high SSIM values can be used to guide the inference process for eliminating structural errors and suppressing spurious details. Based on this finding, we propose a structure-aware Real-ISR (StructSR) method, a simple, effective, and plug-and-play method that enhances structural fidelity for

*Corresponding author.

diffusion-based Real-ISR. StructSR consists of three modules: Structure-Aware Screening (SAS), Structure Condition Embedding (SCE), and Image Details Embedding (IDE). SAS identifies the image with the highest structural similarity to the LR image in the early inference stage. SCE uses structural embedding from SAS to guide the inference in conjunction with the LR image, promoting the generation of high-fidelity structural information. IDE inserts the structural embedding into the clean latent image at each timestep, according to the degradation degree of the LR image, to suppress possible spurious details. Extensive experiments demonstrate the effectiveness of StructSR in enhancing the structural fidelity of diffusion-based Real-ISR methods while effectively suppressing potential spurious details.

In summary, we make the following contributions:

- We propose StructSR, to fully leverage the temporal reconstructed images during the inference process to enhance the structural fidelity of the diffusion-based Real-ISR methods, without introducing any excess fine-tuning, external models' prior, or high-level semantic knowledge.
- We introduce SAS, SCE, and IDE to interactively update the predicted noise and clean images according to the degradation degree of the LR image during inference, enabling a plug-and-play intervention generation process for diffusion-based Real-ISR methods while suppressing potential spurious structure and texture details.
- We demonstrate through experiments that the proposed StructSR consistently improves the structural fidelity of diffusion-based and GAN-based Real-ISR methods.

Related Work

Real-ISR. Starting with SRCNN (Dong et al. 2014), learning-based ISR has gained significant popularity. Numerous methods (Chen et al. 2021, 2023a,b; Dai et al. 2019; Lim et al. 2017) focusing on deep model design have been proposed. Some methods have explored more complex degradation models to approximate real-world degradation. BSRGAN (Zhang et al. 2021) introduces a randomly shuffled degradation modeling, while Real-ESRGAN (Wang et al. 2021) adopts a high-order degradation modeling. Both BSRGAN and Real-ESRGAN employ the Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) to reconstruct desired HR images using training samples with more realistic degradations. Although these methods generate high-fidelity structures, the training of GANs often suffers from instability, leading to unnatural artifacts in the Real-ISR outputs. Subsequent approaches like FeMaSR (Chen et al. 2022) and LDL (Liang, Zeng, and Zhang 2022) have been developed to mitigate this issue of artifacts. However, they still rely heavily on specific paired training data and lack the capability to generate realistic details.

Diffusion Probabilistic Models. The diffusion model is based on non-equilibrium thermodynamic theory (Jarzynski 1997) and the Monte Carlo principle (Neal 2001). It utilizes a diffusion process to iteratively sample from the data distribution, allowing it to capture underlying structures and features in high-dimensional spaces. Dhariwal and Nichol demonstrated that the diffusion model has generation capabilities beyond GAN. Research on accelerated samplers has significantly improved the efficiency of diffusion models, such as DDIM (Rombach et al. 2022). Large-scale latent space-based pre-trained text-to-image (T2I) diffusion models further improved the performance of the diffusion model by moving it from pixel space to latent space, such as Imagen (Saharia et al. 2022a). Meanwhile, the T2I diffusion model is gradually used in image restoration (Meng et al. 2021; Zhang, Rao, and Agrawala 2023), video generation (Singer et al. 2022; Wu et al. 2023), 3D content generation (Lin et al. 2023a; Wang et al. 2023b), etc.

Diffusion-based Real-ISR. Most early attempts (Kawar et al. 2022; Saharia et al. 2022b; Wang, Yu, and Zhang 2022; Fei et al. 2023) to utilize diffusion models for ISR were based on the assumption of simplistic degradations. Recently, researchers have turned to pre-trained text-to-image (T2I) models like Stable Diffusion (Rombach et al. 2022) (SD) to tackle the challenges of Real-ISR, which is trained on large-scale image-text pairs datasets. StableSR (Wang et al. 2023a) refines SD through fine-tuning with a time-aware encoder. DiffBIR (Lin et al. 2023b) employs a two-stage strategy, initially preprocessing the image as an initial estimate and subsequently fine-tuning SD to enhance image details. PASD (Yang et al. 2023) extracts text prompts from LR images and combines them with a pre-trained SD model using a pixel-aware cross-attention module. To further enhance the semantic perception of Real-ISR, SeeSR (Wu et al. 2024) introduces degradation-aware sematic prompts, combining with soft labels to jointly guide the diffusion process.

The above diffusion-based Real-ISR methods ignore the negative impact of the model's empirical prior, resulting in structural errors and spurious details. Since diffusion models are often criticized for their training and inference efficiency, instead of introducing additional degradation-aware sematic prompts (like SeeSR), an additional pre-processing stage (like DiffBIR), or better training materials for fine-tuning, we plan to explore the characteristics of the diffusion model during the inference process to solve this problem.

Methodology

Basic Definition in Diffusion-based Real-ISR

The diffusion-based inference defines the total inference timesteps T and randomly samples a noisy latent image Z_T from a normal distribution $Z_T \sim \mathcal{N}(0, \mathbf{I})$ as the initialization. Given the LR image I_{LR}, it is mapped to the latent space through an encoder $\mathcal{E}$ and uses $\mathcal{E}(I_{LR})$ as the control condition. According to $\mathcal{E}(I_{LR})$, the pre-trained denoiser ϵ_θ predicts the noise ϵ_t at each timestep $t \in [0, T]$ to denoise the noisy latent image Z_t and generates the clean latent image $Z_{0|t}$. The next noisy latent image Z_{t-1} is obtained by adding noise to $Z_{0|t}$ according to the specific sampler. After performing the above process T times, the clean latent image Z_0 is generated and the reconstructed image I_{HR} is the output of decoder $\mathcal{D}$, denoted as $\mathcal{D}(Z_0)$. A plug-and-play framework can directly guide ϵ_t and $Z_{0|t}$ to enhance the diffusion-based Real-ISR methods.

Role of Structural Similarity in Real-ISR

We then investigate the changes in structural fidelity of the reconstructed SR image during inference. We utilize the structural similarity (SSIM) (Wang et al. 2004) to measure the structural fidelity of the reconstructed images:

$$\mathbf{SSIM}(x, y) = \frac{(2\mu_x\mu_y + o)(2\sigma_{xy} + o)}{(\mu_x^2 + \mu_y^2 + o)(\sigma_x^2 + \sigma_y^2 + o)} \quad (1)$$

where μ_x and μ_y are the means, σ_x and σ_y are the standard deviations respectively, σ_{xy} is the covariance, and o is constant used to avoid denominators being zero. Specifically, we first prepare LR images with varying degradation degrees by applying a combination of downsampling, Gaussian Kernel blur, and JPEG compression to real-world images. We use StableSR (Wang et al. 2023a) for Real-ISR and set the total inference timesteps $T = 200$. At each timestep t, the decoder $\mathcal{D}$ generates the reconstructed image according to the clean latent image $Z_{0|t}$, denoted as $\mathcal{D}(Z_{0|t})$. We then resize the LR image I_{LR} by bicubic interpolation to maintain the same size with $\mathcal{D}(Z_{0|t})$, denoted as $SR(I_{LR})$, and calculate the SSIM value between $\mathcal{D}(Z_{0|t})$ and $SR(I_{LR})$ by Eq. 1:

$$S_t = \mathbf{SSIM}(\mathcal{D}(Z_{0|t}), SR(I_{LR})) \quad (2)$$

where S_t represents the calculated SSIM value at timestep t.

As shown in Fig. 2, by comparing the SSIM values and the reconstructed images throughout the inference process, we find that the reconstructed images show the most consistent and clearer structure compared with the LR images in the early inference stage. However, the model cannot maintain this clear structure, resulting in decreasing SSIM, indicating the generation of spurious structure and texture details. Another recent work DiffBIR (Lin et al. 2023b), uses an additional pre-processing stage to provide clear guidance images by introducing an additional model. In our work, we consider screening out the one with the most consistent structure with the LR image to guide the inference process. We define the initial T_{SAS} inference timesteps as the early inference stage, and define a structural embedding Z_{SE} to intervene in the clean latent image $Z_{0|t}$. Since S_t calculated in the early inference stage reflects the degradation degree of the LR image, it can be used in the structural embedding Z_{SE} to control the guidance strength of the reconstructed image after the inference timesteps T_{SAS}.

Based on the above observations and considerations, our proposed framework is shown in Fig. 3(a). The structure-aware screening (SAS) works in the early inference stage and screens out the structural embedding Z_{SE} according to S_t. The structure condition embedding (SCE) uses the structural embedding Z_{SE} to guide the prediction of ϵ_t. The image details embedding (IDE) inserts the structural embedding Z_{SE} into the clean latent image at each timestep to guide $Z_{0|t}$.

Structure-Aware Screening

Structure-Aware Screening (SAS) is implemented by adding operations based on SSIM in the original inference. As shown in Fig. 3(b), at each timestep t in the initial T_{SAS} inference timesteps, we use the decoder $\mathcal{D}$ to obtain the

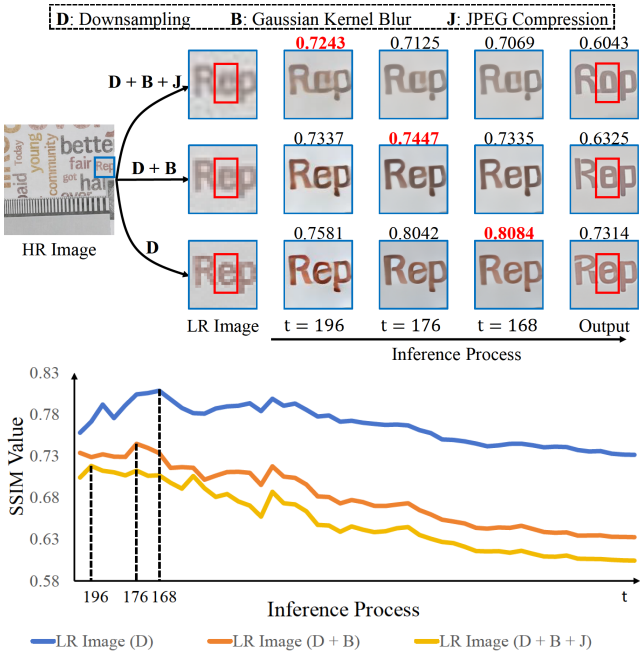


Figure 2: Comparison of the structural similarity (SSIM) between LR images with different degradation degrees and their temporal reconstructed images during the inference process. The calculated SSIM values are shown on the top of the reconstructed images, with the maximum SSIM value in red. The red boxes show the issues of structural errors and spurious details. It cannot maintain a stable SSIM, indicating the generation of spurious structure and texture details in the later stage of the inference.

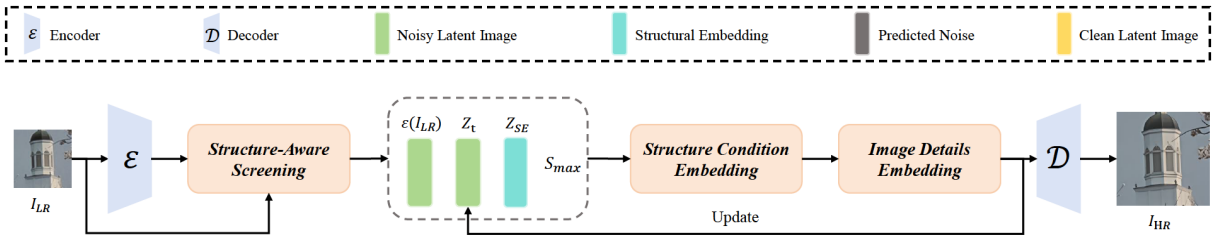
reconstructed image $\mathcal{D}(Z_{0|t})$ before updating Z_t to Z_{t-1}. Then, we use Eq. 2 to calculate the SSIM value S_t between $\mathcal{D}(Z_{0|t})$ and the LR image I_{LR}. We screen out the reconstructed image with the structure that is most consistent with the LR image to ensure the best guidance. Specifically, we use a buffer to store the calculated SSIM values and identify the maximum one from the buffer by $S_{max} = \max\{S_t\}$. If $S_t = S_{max}$, we assign this clean latent image $Z_{0|t}$ as the structural embedding Z_{SE},

$$Z_{SE} = \begin{cases} Z_{0|t}, & S_{t\in[T-T_{SAS},T]} = S_{max} \\ Z_{SE}, & \text{otherwise} \end{cases} \quad (3)$$

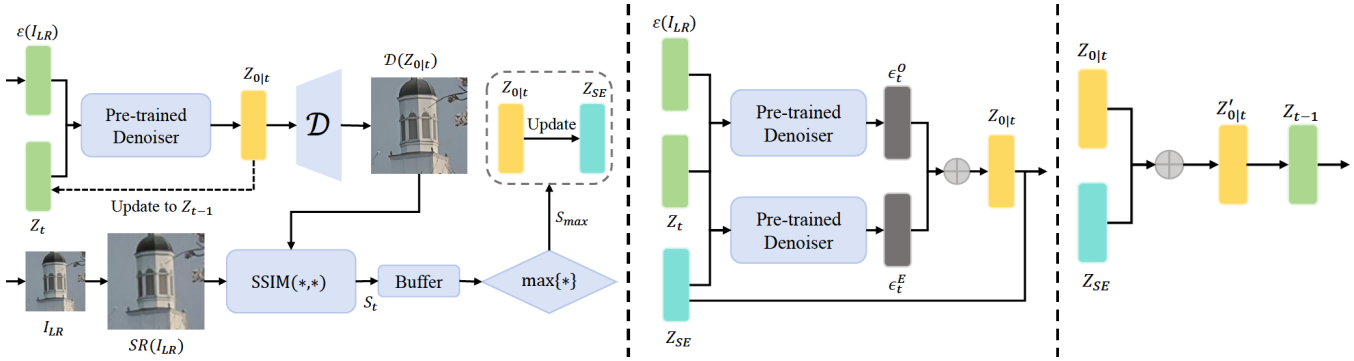
SAS provides the structural embedding Z_{SE} and the maximum SSIM value S_{max} for intervention in the subsequent inference process. S_{max} occurs in the early inference stage $t \in [T - T_{SAS}, T]$, which is typically proportional to the original image quality. Better original image quality can ensure that S_{max} occurs later, as shown in Fig. 2.

Structure Condition Embedding

Structural Conditional Embedding (SCE), as shown in Fig. 3(c), uses Z_{SE} as the condition of noise prediction to guide ϵ_t with clearer structural information for addressing the issue of structural errors. Specifically, in the original inference process, the pre-trained denoiser ϵ_θ uses $\mathcal{E}(I_{LR})$ as



(a) Overview of the proposed StructSR



(b) Structure-Aware Screening (c) Structure Condition Embedding (d) Image Details Embedding

Figure 3: In the proposed StructSR, the Structure-Aware Screening (SAS) works in the early inference stage and screens out the structural embedding Z_{SE} with the most consistent and clearer structure compared to the LR image. In the later inference stage, The Structure Condition Embedding (SCE) uses Z_{SE} to guide ϵ_t in conjunction with the LR image. The Image Details Embedding (IDE) inserts Z_{SE} into the clean latent image $Z_{0|t}$ at each timestep according to the degradation degree.

the control condition for the noise prediction. We defined this predicted noise as ϵ_t^O. SCE uses Z_{SE} as the additional control condition for ϵ_θ and predicts the extra noise, denoted as ϵ_t^E. SCE uses ϵ_t^O to suppress the impact of illusions by constraining ϵ_t^E and defines it as follows:

$$\hat{\epsilon}_t = S_{max}\epsilon_t^E + (1 - S_{max})\epsilon_t^O \tag{4}$$

where S_{max} is the maximum SSIM value provided by SAS.

The constraint strength provided by ϵ_t^O depends on the structural clarity of the reconstructed image corresponding to the control condition Z_{SE}. It is directly decided by the degradation degree of the LR image which is reflected by S_{max}. Hence, SCE uses S_{max} to control the constraint strength to provide more accurate structural guidance.

Image Details Embedding

Image Details Embedding (IDE), as shown in Fig. 3(d), works in the remaining $T - T_{SAS}$ inference timesteps. Compared with the subsequently reconstructed image, the image corresponding to Z_{SE} contains insufficient details. But it also has the advantage of fewer spurious details and can be used to suppress subsequent spurious details caused by the model's illusions. IDE inserts Z_{SE} into the clean latent image $Z_{0|t}$ at each timestep t, where $t \in [0, T - T_{SAS}]$. The insertion process can be expressed as follows:

$$Z'_{0|t} = w_t Z_{SE} + (1 - w_t) Z_{0|t} \tag{5}$$

Where w_t is the factor that controls the insertion ratio.

The image reconstructed from a severely degraded LR image is severely missing details. Continuously and intensively inserting it into subsequent reconstructed images will cause the final SR image to be too smooth. To avoid this problem, IDE determines the insertion rate according to the degradation degree of the LR image and gradually weakens it. Specifically, the control factor w_t is weakened according to the current timestep t, and the remaining inference timesteps $T - T_{SAS}$ as follows:

$$w_t = \frac{S_{max} t}{T - T_{SAS}} \tag{6}$$

Where S_{max} is the maximum SSIM value provided by SAS and reflects the degradation degree of the LR image.

StructSR exploits the structural information via S_{max} and Z_{SE} during the inference process and interactively updates the predicted noise ϵ_t and clean latent images $Z_{0|t}$ according to the degradation degree. This enables a plug-and-play intervention inference process for existing diffusion-based Real-ISR methods to suppress potential spurious structure and texture details, while avoiding oversmoothing the real details. Algorithm 1 summarizes the proposed StructSR.

Experiments

Experimental Settings

Test Datasets. We use one synthetic dataset and two real-world datasets to comprehensively evaluate StructSR. Fol-

Algorithm 1: StructSR Inference Process

Input: I_{LR} ▷ LR image
Parameter: T ▷ Total inference timesteps
T_{SAS} ▷ Inference timesteps with SAS
$\mathcal{E}$ ▷ Encoder
$\mathcal{D}$ ▷ Decoder
ϵ_θ ▷ Pre-trained denoiser
Output: I_{HR}

$Z_T \sim \mathcal{N}(0, \mathbf{I})$
for $t \in [T, ..., 0]$ **do**
  $\epsilon_t^O = \epsilon_\theta(Z_t, \mathcal{E}(I_{LR}), t)$
  if $t > T - T_{SAS}$ **then**
    $Z_{0|t} = \text{Sampler}(Z_t, \epsilon_t^O, t)$
    $S_t = \mathbf{SSIM}(\mathcal{D}(Z_{0|t}), SR(I_{LR}))$
    Add S_t to Buffer
    $S_{max} = \max\{S_t\}$
    $Z_{SE} = \begin{cases} Z_{0|t}, & S_t = S_{max} \\ Z_{SE}, & \text{otherwise} \end{cases}$
    $Z_{t-1} = \text{Sampler}(Z_{0|t}, \epsilon_t^O, t)$
  else
    $\epsilon_t^E = \epsilon_\theta(Z_t, Z_{SE}, t)$
    $\hat{\epsilon}_t = S_{max}\epsilon_t^E + (1 - S_{max})\epsilon_t^O$
    $Z_{0|t} = \text{Sampler}(Z_t, \hat{\epsilon}_t, t)$
    $w_t = \frac{S_{max} t}{T - T_{SAS}}$
    $Z'_{0|t} = w_t Z_{SE} + (1 - w_t) Z_{0|t}$
    $Z_{t-1} = \text{Sampler}(Z'_{0|t}, \hat{\epsilon}_t, t)$
  end if
end for
$I_{HR} = \mathcal{D}(Z_0)$
return I_{HR}

lowing the pipeline in StableSR (Wang et al. 2023a), first, we randomly crop 3K patches (resolution: 512×512) from the DIV2K validation set (Agustsson and Timofte 2017) and degrade them as synthetic images, named DIV2K-Val. Then, we use two real-world datasets, RealSR (Cai et al. 2019) and DRealSR (Wei et al. 2020), to center-crop LR images (128×128) as real-world images.

Evaluation Metrics. In order to provide a comprehensive and holistic assessment of the performance of different methods, we employ a range of reference and no-reference metrics. PSNR and SSIM (Wang et al. 2004) (calculated on the Y channel in YCbCr space) are reference-based structural fidelity measures, while LPIPS (Zhang et al. 2018) is the reference-based perceptual quality metric. MUSIQ (Ke et al. 2021) and CLIPIQA (Wang, Chan, and Loy 2023) are two no-reference image quality metrics.

Compared Methods and Implementation Details. We adopt four state-of-the-art diffusion-based Real-ISR methods StableSR (Wang et al. 2023a), DiffBIR (Lin et al. 2023b), PASD (Yang et al. 2023), and SeeSR (Wu et al. 2024) as baselines and test them using publicly released codes and models. The model parameters and samplers are set according to the original paper.

For StructSR, we set $T_{SAS} = 0.3\ T$ as the inference timesteps for screening and intervene in the inference process according to the sampler. We present the **ablation study** on T_{SAS}, SCE, and IDE in the supplementary material.

To provide a comprehensive assessment, we also show the comparison with the GAN-based Real-ISR methods. We choose BSRGAN (Zhang et al. 2021), Real-ESRGAN (Wang et al. 2021), LDL (Liang, Zeng, and Zhang 2022), and FeMaSR (Chen et al. 2022). For them, we directly use the publicly released codes and models for testing.

Comparison with the State-of-the-Art

Quantitative Comparisons. We present quantitative comparisons on three synthetic and real datasets in Table 1. We can find the following phenomena. (1) Compared with the original diffusion-based methods without integrating with our StructSR, GAN-based methods have significant advantages in PSNR, SSIM, and LPIPS metrics while showing disadvantages in no-reference metrics MUSIQ and CLIP-IQA. This demonstrates that the GAN-based methods generate images with higher structural fidelity but suffer from the problem of insufficient realistic details. (2) Compared with the four diffusion-based baselines, the methods integrating with our StructSR achieved the best results in PSNR and SSIM on all datasets. Although LPIPS did not achieve optimal results, they are better than the original. (3) After StructSR enhancement, StableSR, PASD, and SeeSR achieved better results on MUSIQ and CLIP-IQA metrics. Although DiffBIR slightly decreased on the MUSIQ and CLIP-IQA, it significantly improved on PSNR and SSIM, especially PSNR, achieving the best results among all methods. In summary, our StructSR significantly enhances the performance of diffusion-based Real-ISR methods across full-reference metrics and marginally improves their performance in no-reference metrics, which demonstrates the effectiveness of StructSR for enhancing structural fidelity and suppressing spurious details.

Qualitative Comparisons. Fig. 4 shows the visual comparison of different Real-ISR methods. From the first example, we can observe that BSRGAN, Real-ESRGAN, and LDL preserve the structures in LR images well. FeMaSR produces many artifacts due to the weakness of denoising ability. StableSR, PASD, and SeeSR produce a few spurious details. DiffBIR changes the structural information and generates spurious details. Compared with the baselines, StructSR suppresses the generation of spurious details and corrects the structural errors. From the second example, we can observe that the GAN-based methods lack the ability to generate details. Although the diffusion-based methods generate more details, they change the structural information and generate details inconsistent with GT. StructSR suppresses the spurious information compared to baselines. The last example shows similar conclusions. These results demonstrate that StructSR can effectively enhance the structural fidelity and suppress the spurious details for diffusion-based Real-ISR. More comparisons are shown in the supplementary material.

User Study. To further validate the effectiveness of our approach, we conducted user studies on both synthetic and real-world data. Considering that the LR-HR image pairs in

Datasets	Methods	PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ ↑	CLIP-IQA ↑
DIV2K-Val (synthetic)	BSRGAN	23.26	0.5907	0.3351	61.20	0.5247
	Real-ESRGAN	22.97	0.5986	0.3112	61.07	0.5281
	FeMaSR	21.74	0.5536	0.3126	60.83	0.5998
	LDL	22.17	0.5798	0.3256	60.04	0.5179
	StableSR / +StructSR	21.94 / 23.47 ↑	0.5335 / **0.6117** ↑	0.3122 / **0.3004** ↓	65.87 / 66.72 ↑	0.6776 / 0.6816 ↑
	DiffBIR / +StructSR	22.31 / **23.60** ↑	0.5272 / 0.5835 ↑	0.3520 / 0.3286 ↓	65.78 / 65.02 ↓	0.6696 / 0.6558 ↓
	PASD / +StructSR	22.31 / 23.35 ↑	0.5656 / 0.6029 ↑	0.3728 / 0.3072 ↓	64.52 / 67.93 ↑	0.6180 / 0.6951 ↑
	SeeSR / +StructSR	22.36 / 23.18 ↑	0.5673 / 0.5986 ↑	0.3194 / 0.3066 ↓	68.67 / **69.26** ↑	0.6934 / **0.6992** ↑
RealSR (real-world)	BSRGAN	25.06	0.7401	**0.2656**	63.28	0.5115
	Real-ESRGAN	24.32	0.7352	0.2726	60.45	0.4521
	FeMaSR	23.74	0.7087	0.2937	59.06	0.5408
	LDL	23.76	0.7198	0.2750	59.28	0.4430
	StableSR / +StructSR	23.41 / 24.31 ↑	0.6824 / **0.7447** ↑	0.3001 / 0.2915 ↓	65.05 / 67.68 ↑	0.6210 / 0.6624 ↑
	DiffBIR / +StructSR	23.67 / **25.09** ↑	0.6245 / 0.6938 ↑	0.3626 / 0.3610 ↓	64.66 / 63.93 ↓	0.6545 / 0.6487 ↓
	PASD / +StructSR	23.86 / 24.49 ↑	0.6946 / 0.7328 ↑	0.2960 / 0.2904 ↓	65.13 / 69.31 ↑	0.5760 / **0.6995** ↑
	SeeSR / +StructSR	23.83 / 24.25 ↑	0.6947 / 0.7159 ↑	0.3007 / 0.2989 ↓	69.81 / **70.76** ↑	0.6703 / 0.6868 ↑
DRealSR (real-world)	BSRGAN	27.38	0.7740	0.2858	57.16	0.5091
	Real-ESRGAN	27.29	0.7768	0.2819	54.27	0.4515
	FeMaSR	25.55	0.7246	0.3156	53.71	0.5639
	LDL	25.97	0.7667	**0.2791**	53.95	0.4474
	StableSR / +StructSR	26.88 / 27.86 ↑	0.7252 / **0.7937** ↑	0.3182 / 0.2967 ↓	57.81 / 60.96 ↑	0.6029 / 0.6524 ↑
	DiffBIR / +StructSR	25.40 / **27.98** ↑	0.6226 / 0.7566 ↑	0.4381 / 0.3640 ↓	60.37 / 59.76 ↓	0.6379 / 0.6176 ↓
	PASD / +StructSR	26.48 / 27.14 ↑	0.7321 / 0.7662 ↑	0.3327 / 0.3151 ↓	58.90 / 65.33 ↑	0.5909 / **0.7243** ↑
	SeeSR / +StructSR	26.74 / 27.17 ↑	0.7405 / 0.7754 ↑	0.3173 / 0.3028 ↓	65.09 / **66.80** ↑	0.6912 / 0.7013 ↑

Table 1: Quantitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks. BSRGAN (ICCV2021), Real-ESRGAN (ICCV2021), FeMaSR(ACM Multimedia2022) ,and LDL (CVPR2022) are GAN-based methods. StableSR (IJCV2024), DiffBIR (arXiv2023), PASD (ECCV2024), and SeeSR (CVPR2024) are diffusion-based methods. The best results of each metric are highlighted. Increasing and decreasing metrics are indicated by ↑ and ↓, respectively.

real-world data cannot be exactly matched like those in synthetic data, we use different settings for them. On the synthetic data, followed the way in SR3 (Saharia et al. 2022c) and SeeSR (Wu et al. 2024), participants were shown an LR image placed between two HR images: one was GT and the other was the Real-ISR output of a certain method. They were asked to decide, "Which HR image has the structure that better corresponds to the LR image?" When making their decision, participants were asked to consider the structural similarity to the LR image. A confusion rate was then calculated, which indicated whether participants preferred the GT or the Real-ISR output. On the real-world data, participants were shown an LR image along with all Real-ISR outputs and asked to answer "Which image has the structure that best corresponds to the LR image?" Then, the best rate was calculated to represent the probability of a model being selected.

We invite 30 participants to test representative Real-ISR methods. There are 12 synthetic test sets and 36 real-world test sets. Synthetic data are randomly sampled from DIV2K-Val, and real-world data are randomly sampled from DRealSR and RealSR. Each of the 30 participants is asked to make 144 (12x12) choices on synthetic data and 36 choices on real-world data. As shown in Table 2, methods Integrated with our StructSR significantly outperform other methods in terms of selection rate on both synthetic and real-world data. In the user study on synthetic data, SeeSR achieves the best results among all methods without applying StructSR, while applying StructSR achieves a 58.3% confusion rate, 25% higher than the original. In the user study on real-world data, SeeSR applying our StructSR achieves the best selection rate of 16.7%.

Methods	Confusion rates	Best rates
BSRGAN	8.3%	5.6%
Real-ESRGAN	16.7%	5.6%
FeMaSR	8.3%	2.8%
LDL	8.3%	2.8%
StableSR / +StructSR	25.0% / 33.3%	8.3% / 13.9%
DiffBIR / +StructSR	16.7% / 25.0%	2.8% / 5.6%
PASD / +StructSR	25.0% / 41.67%	11.1% / 13.9%
SeeSR / +StructSR	33.3% / 58.3%	11.1% / 16.7%

Table 2: User study on synthetic and real-world data. Confusion rates are calculated on synthetic data and best rates are calculated on real-world data.

Conclusion

We propose StructSR, a novel plug-and-play method that enhances structural fidelity and suppresses spurious details for diffusion-based Real-ISR without any forms of model fine-

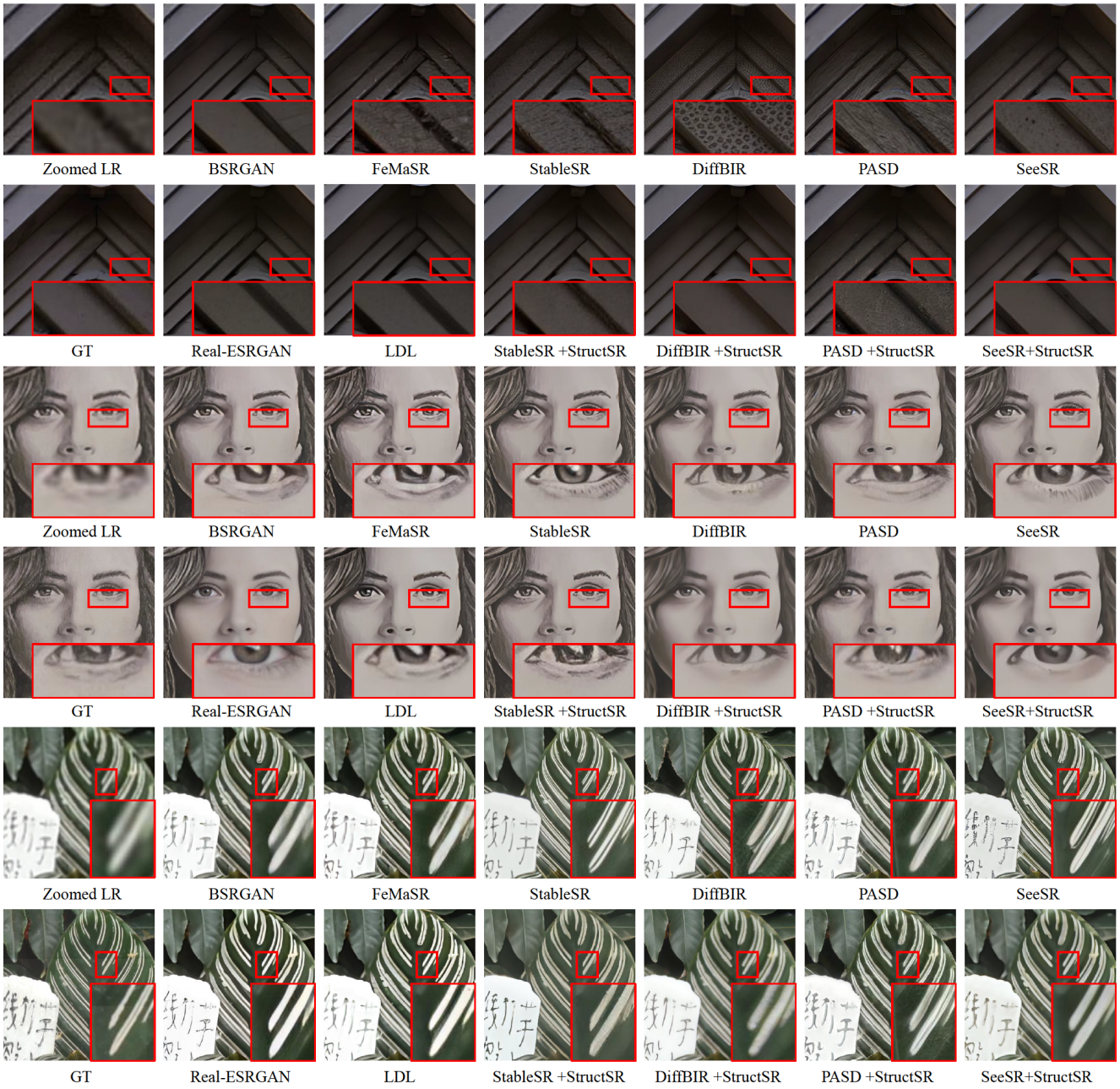


Figure 4: Qualitative comparisons of different Real-ISR methods. Integration with StructSR significantly reduces spurious details of diffusion-based methods, resulting in high fidelity.

tuning, external model priors, or high-level semantic knowledge. By exploring the structural similarity between reconstructed images and the original low-resolution image during the inference process, we found that the reconstructed images with high structural similarity values can be used to guide the inference process. Our work makes significant progress in effectively leveraging the internal characteristics of the diffusion-based Real-ISR models to synthesize high-resolution images with higher structural fidelity and fewer spurious details, as demonstrated through extensive experimental evaluations. More details can be found in the supplementary material.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under grant 62272229 and U2441285, the Natural Science Foundation of Jiangsu Province under grant BK20222012, and Shenzhen Science and Technology Program JCYJ20230807142001004.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126–135.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3086–3095.

Chen, C.; Shi, X.; Qin, Y.; Li, X.; Han, X.; Yang, T.; and Guo, S. 2022. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1329–1338.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12299–12310.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023a. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22367–22377.

Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yang, X.; and Yu, F. 2023b. Dual Aggregation Transformer for Image Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12312–12321.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and 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*, 11065–11074.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, 184–199. Springer.

Fei, B.; Lyu, Z.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; and Dai, B. 2023. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9935–9946.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63: 139–144.

Jarzynski, C. 1997. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56: 5018.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35: 23593–23606.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liang, J.; Zeng, H.; and Zhang, L. 2022. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5657–5666.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136–144.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023a. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 300–309.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Fei, B.; Dai, B.; Ouyang, W.; Qiao, Y.; and Dong, C. 2023b. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. *arXiv preprint arXiv:2308.15070*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Neal, R. M. 2001. Annealed importance sampling. *Statistics and computing*, 11: 125–139.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022a. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 4713–4726.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022c. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45: 4713–4726.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2555–2563.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023a. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv preprint arXiv:2305.07015*.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1905–1914.

Wang, Y.; Yu, J.; and Zhang, J. 2022. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13: 600–612.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv preprint arXiv:2305.16213*.

Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; and Lin, L. 2020. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 101–117. Springer.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7623–7633.

Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; and Zhang, L. 2024. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25456–25467.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2023. Pixel-Aware Stable Diffusion for Realistic Image Super-resolution and Personalized Stylization. *arXiv preprint arXiv:2308.14469*.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4791–4800.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.

Supplementary Material

In this supplementary material, we first present the ablation studies on Structure Condition Embedding (SCE) and Image Details Embedding (IDE). Then we present further analysis of the definition of the early inference stage T_{SAS}.

Ablation Study

Effectiveness of SCE and IDE

To demonstrate the effectiveness of the proposed Structure Condition Embedding (SCE) and Image Details Embedding (IDE) in improving the structural fidelity of the reconstructed image and suppressing spurious details, we perform several ablation studies. We used StableSR (Wang et al. 2023a), DiffBIR (Lin et al. 2023b), PASD (Yang et al. 2023), and SeeSR (Wu et al. 2024) as baselines. The processed real-world datasets RealSR (Cai et al. 2019) and DRealSR (Wei et al. 2020) were used for testing. The quantitative evaluation results of the ablation experiments are shown in Table **??**, and the corresponding visual performance is shown in Fig. 1.

Specifically, 1) **w/o StructSR**: Images are reconstructed only through the original inference process. **2) w/o SCE**: Images do not go through SCE, and only IDE is used for suppressing spurious details. **3) w/o IDE**: Images do not go through IDE, and only SCE is used for guiding noise prediction. **4) w StructSR**: Images are reconstructed by integrating StructSR into the original inference process.

As shown in Table **??** and Fig. 1, for w/o SCE, our proposed IDE effectively suppresses spurious details in the reconstructed image but causes the problem of over-smoothing. The results of w/o IDE show that our proposed SCE provides clearer structural guidance for the inference process and helps the model produce more realistic details. In terms of metrics, compared with the baselines, w/o SCE improves in PSNR and SSIM but decreases in LPIPS, MUSIQ, and CLIP-IQA. This shows that the results of w/o SCE suppress spurious details while causing insufficient details in the reconstructed image because IDE suppresses the model's ability to generate details. Compared with the baselines, w/o IDE improves in all metrics, but the improvement in PSNR is lower than those of w/o SCE. This shows that the clear structural guidance provided by SCE guides the model in producing more details while guiding high-fidelity structural reconstruction, but some details are spurious.

Datasets	Metrics	StableSR				DiffBIR			
		w/o StructSR	w/o SCE	w/o IDE	w StructSR	w/o StructSR	w/o SCE	w/o IDE	w StructSR
RealSR	PSNR ↑	23.41	24.29	24.07	**24.31**	23.47	24.95	24.62	**25.09**
	SSIM ↑	0.6824	0.7001	0.7393	**0.7447**	0.6245	0.6502	0.6846	**0.6938**
	LPIPS ↓	0.3001	0.3213	**0.2907**	0.2915	0.3626	0.3942	**0.3342**	0.3610
	MUSIQ ↑	65.05	61.94	**70.29**	67.68	64.66	56.05	**65.74**	63.93
	CLIP-IQA ↑	0.621	0.5698	**0.6967**	0.6624	0.6545	0.5336	**0.6815**	0.6487
	Metrics	PASD				SeeSR			
		w/o StructSR	w/o SCE	w/o IDE	w StructSR	w/o StructSR	w/o SCE	w/o IDE	w StructSR
	PSNR ↑	23.86	24.26	23.91	**24.49**	23.83	24.17	23.92	**24.25**
	SSIM ↑	0.6946	0.7022	0.7223	**0.7328**	0.6947	0.7060	0.7114	**0.7159**
	LPIPS ↓	0.296	0.3071	**0.2901**	0.2904	0.3007	0.3065	**0.2971**	0.2989
	MUSIQ ↑	65.13	68.55	**70.71**	69.31	69.81	65.23	**71.59**	70.76
	CLIP-IQA ↑	0.5760	0.6730	**0.7077**	0.6995	0.6703	0.6409	**0.7054**	0.6868
DRealSR	Metrics	StableSR				DiffBIR			
		w/o StructSR	w/o SCE	w/o IDE	w StructSR	w/o StructSR	w/o SCE	w/o IDE	w StructSR
	PSNR ↑	26.88	27.42	27.00	**27.86**	25.40	27.17	26.61	**27.98**
	SSIM ↑	0.7252	0.7541	0.7827	**0.7937**	0.6226	0.7396	0.7415	**0.7566**
	LPIPS ↓	0.3182	0.3259	**0.2958**	0.2967	0.4381	0.4649	**0.3556**	0.364
	MUSIQ ↑	57.81	51.85	**64.25**	60.96	60.37	56.99	**61.06**	59.76
	CLIP-IQA ↑	0.6029	0.5067	**0.7026**	0.6524	0.6379	0.4779	**0.6737**	0.6176
	Metrics	PASD				SeeSR			
		w/o StructSR	w/o SCE	w/o IDE	w StructSR	w/o StructSR	w/o SCE	w/o IDE	w StructSR
	PSNR ↑	26.48	26.91	26.54	**27.14**	26.74	26.95	26.81	**27.17**
	SSIM ↑	0.7321	0.7449	0.7575	**0.7662**	0.7405	0.7661	0.7697	**0.7754**
	LPIPS ↓	0.3327	0.3418	**0.3122**	0.3151	0.3173	0.3232	**0.3006**	0.3028
	MUSIQ ↑	58.9	65.12	**67.75**	65.33	65.09	60.44	**67.11**	66.8
	CLIP-IQA ↑	0.5909	0.7171	**0.7320**	0.7243	0.6912	0.6523	**0.7332**	0.7013

Table 1: Ablation study on SCE and IDE with state-of-the-art diffusion-based Real-ISR baselines.

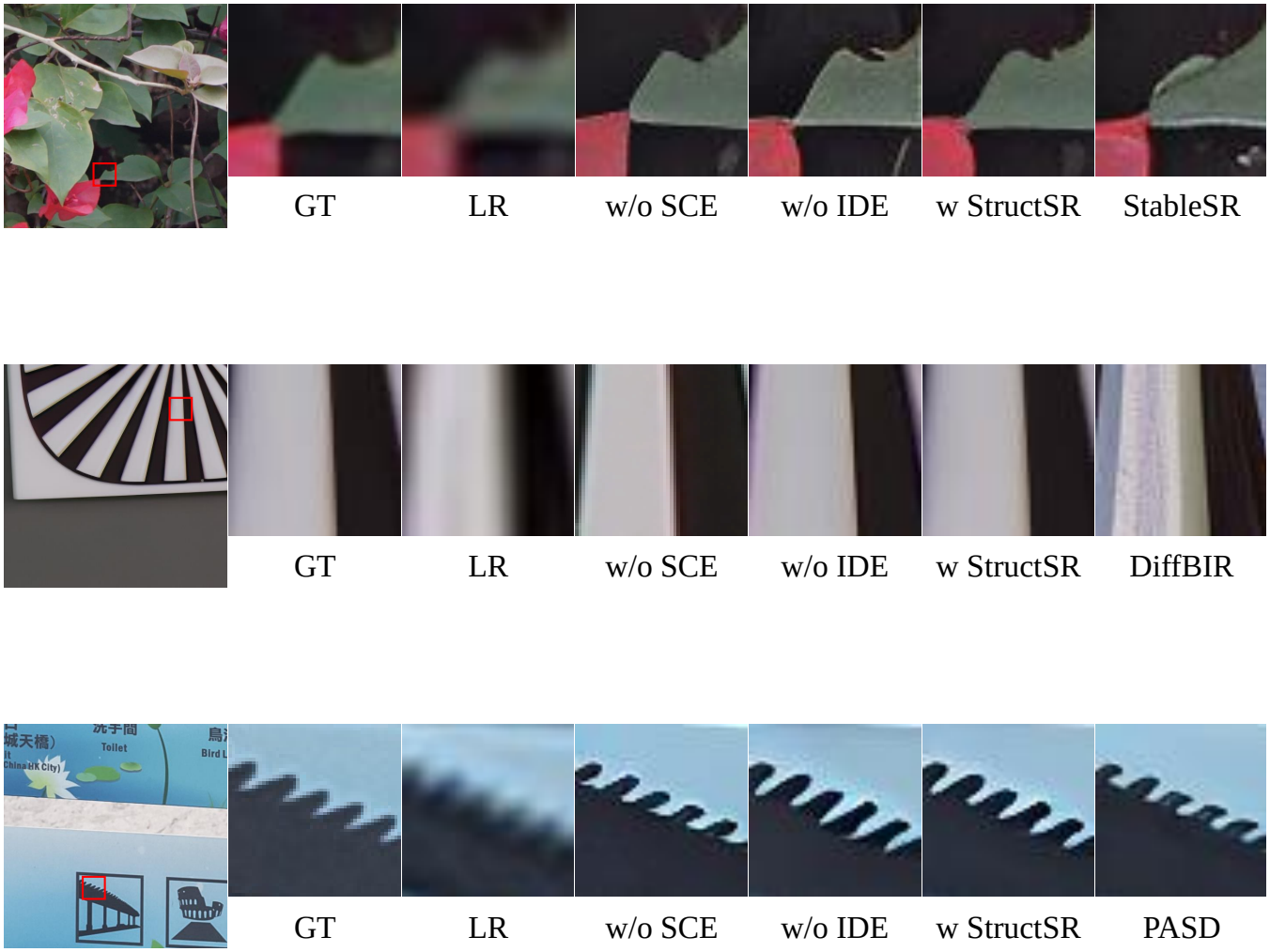


Figure 1: Ablation study on SCE and IDE with state-of-the-art diffusion-based Real-ISR baselines. Integration with StructSR generates high-fidelity structures by combining the clear structural guidance provided by SCE and the suppression of spurious details by IDE.

Definition of early inference stage

The definition of T_{SAS} in the early inference stage is determined by the time when the maximum SSIM value between the reconstructed images and the LR image occurs. In the main text, we obtain three images with different degradation degrees based on a real image and study the changes in SSIM between the reconstructed images and them during the inference process.

To obtain more generalized conclusions, we use more real-world images for research. Specifically, we perform center-crop on the HR images in the RealSR dataset to obtain **100** real-world images. Following the settings in the main text, we apply a combination of downsampling (D), Gaussian Kernel blur (D + B), and JPEG compression (J) to obtain three sets of LR images with different degradation degrees.

As shown in Figure 2, by calculating the average SSIM value, we find that when $t \in [0.9\ T, T]$, the results of different degradation degrees all contain the maximum value. This shows that, in most cases, setting $T_{SAS} = 0.1\ T$ can ensure that the reconstructed image screened by SAS is most consistent with the LR image in structure.

Based on the above findings, we set five sets of T_{SAS} to verify the impact of longer early inference time steps on image quality. They are $T_{SAS} = 0.1\ T$, $T_{SAS} = 0.2\ T$, $T_{SAS} = 0.3\ T$, $T_{SAS} = 0.4\ T$, and $T_{SAS} = 0.5\ T$. We use the processed RealSR (Cai et al. 2019) and DRealSR (Wei et al. 2020) datasets as test datasets and StableSR (Wang et al. 2023a), DiffBIR (Lin et al. 2023b), PASD (Yang et al. 2023), and SeeSR (Wu et al. 2024) as baselines. As shown in Table **??** and Fig. 3, for $T_{SAS} = 0.3\ T$, integration with StructSR generates the most consistent structure with GT and the least spurious details. In terms of metrics, $T_{SAS} = 0.3\ T$ achieves the best performance on PSNR, SSIM, and LPIPS. Compared with $T_{SAS} = 0.3\ T$, the results of $T_{SAS} = 0.4\ T$ and $T_{SAS} = 0.5\ T$ contain more realistic but spurious details that do not match GT. In terms of metrics, $T_{SAS} = 0.5\ T$ achieves the best performance on MUSIQ and CLIP-IQA, but is inferior to $T_{SAS} = 0.3\ T$ in terms of PSNR, SSIM, and LPIPS. This indicates that a too-long early stage sacrifices the structural fidelity of the reconstructed images. By balancing structural fidelity and realistic details, we propose to adopt $T_{SAS} = 0.3\ T$ as the early inference timesteps.

Datasets	Metrics	StableSR					DiffBIR				
		$0.1\ T$	$0.2\ T$	$0.3\ T$	$0.4\ T$	$0.5\ T$	$0.1\ T$	$0.2\ T$	$0.3\ T$	$0.4\ T$	$0.5\ T$
RealSR	PSNR ↑	24.08	24.15	**24.31**	24.10	24.06	24.59	24.86	**25.09**	24.91	24.83
	SSIM ↑	0.7347	0.7411	**0.7447**	0.7387	0.7215	0.6847	0.6891	**0.6938**	0.6891	0.6835
	LPIPS ↓	0.3012	0.2974	**0.2915**	0.2948	0.2967	0.3683	0.3636	**0.3610**	0.3627	0.3676
	MUSIQ ↑	66.89	67.31	67.68	67.74	**67.86**	63.28	63.69	63.93	64.05	**64.39**
	CLIP-IQA ↑	0.6559	0.6605	0.6624	0.6632	**0.6643**	0.6419	0.6468	0.6487	0.6489	**0.6494**
	Metrics	PASD					SeeSR				
		$0.1\ T$	$0.2\ T$	$0.3\ T$	$0.4\ T$	$0.5\ T$	$0.1\ T$	$0.2\ T$	$0.3\ T$	$0.4\ T$	$0.5\ T$
	PSNR ↑	23.97	24.13	**24.49**	24.15	24.03	23.90	24.04	**24.25**	24.07	23.91
	SSIM ↑	0.7259	0.7285	**0.7328**	0.7243	0.7161	0.7034	0.7124	**0.7159**	0.7114	0.7079
	LPIPS ↓	0.2966	0.2935	**0.2904**	0.2915	0.2937	0.3033	0.3023	**0.2989**	0.3004	0.3027
	MUSIQ ↑	67.98	68.94	69.31	69.38	**69.49**	70.64	70.66	70.76	70.83	**70.90**
	CLIP-IQA ↑	0.6948	0.6973	0.6995	0.6996	**0.7001**	0.6837	0.6858	0.6868	0.6887	**0.6893**
DRealSR	Metrics	StableSR					DiffBIR				
		$0.1\ T$	$0.2\ T$	$0.3\ T$	$0.4\ T$	$0.5\ T$	$0.1\ T$	$0.2\ T$	$0.3\ T$	$0.4\ T$	$0.5\ T$
	PSNR ↑	27.10	27.40	**27.86**	27.67	26.46	27.03	27.45	**27.98**	27.57	27.39
	SSIM ↑	0.7833	0.7895	**0.7937**	0.7877	0.7851	0.7451	0.7494	**0.7566**	0.7475	0.7428
	LPIPS ↓	0.3081	0.3002	**0.2967**	0.3037	0.3043	0.3787	0.3718	**0.3640**	0.3642	0.3643
	MUSIQ ↑	60.18	60.57	60.96	61.14	**61.37**	59.54	59.61	59.76	59.83	**59.91**
	CLIP-IQA ↑	0.6275	0.6396	0.6524	0.6558	**0.6582**	0.6132	0.6165	0.6176	0.6197	**0.6209**
	Metrics	PASD					SeeSR				
		$0.1\ T$	$0.2\ T$	$0.3\ T$	$0.4\ T$	$0.5\ T$	$0.1\ T$	$0.2\ T$	$0.3\ T$	$0.4\ T$	$0.5\ T$
	PSNR ↑	26.70	26.93	**27.14**	27.05	26.89	26.83	27.09	**27.17**	27.03	26.95
	SSIM ↑	0.7568	0.7604	**0.7662**	0.7615	0.7572	0.7681	0.7716	**0.7754**	0.7726	0.7709
	LPIPS ↓	0.3365	0.3266	**0.3151**	0.3213	0.3221	0.3182	0.3083	**0.3028**	0.3054	0.3073
	MUSIQ ↑	64.28	64.69	65.33	65.52	**65.56**	65.95	66.03	66.80	66.87	**66.94**
	CLIP-IQA ↑	0.7203	0.7213	0.7243	0.7251	**0.7256**	0.6969	0.7004	0.7013	0.7057	**0.7093**

Table 2: Quantitative experiments on different setting of T_{SAS} with state-of-the-art diffusion-based Real-ISR baselines.

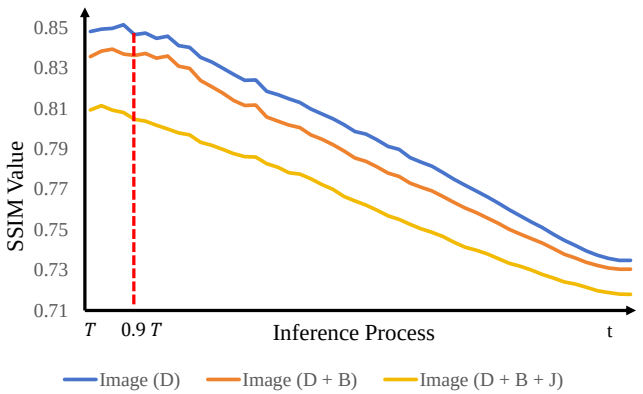


Figure 2: Comparison of the SSIM between LR images with different degradation degrees and their temporal reconstructed images during the StableSR inference process with total inference timesteps $T = 200$.

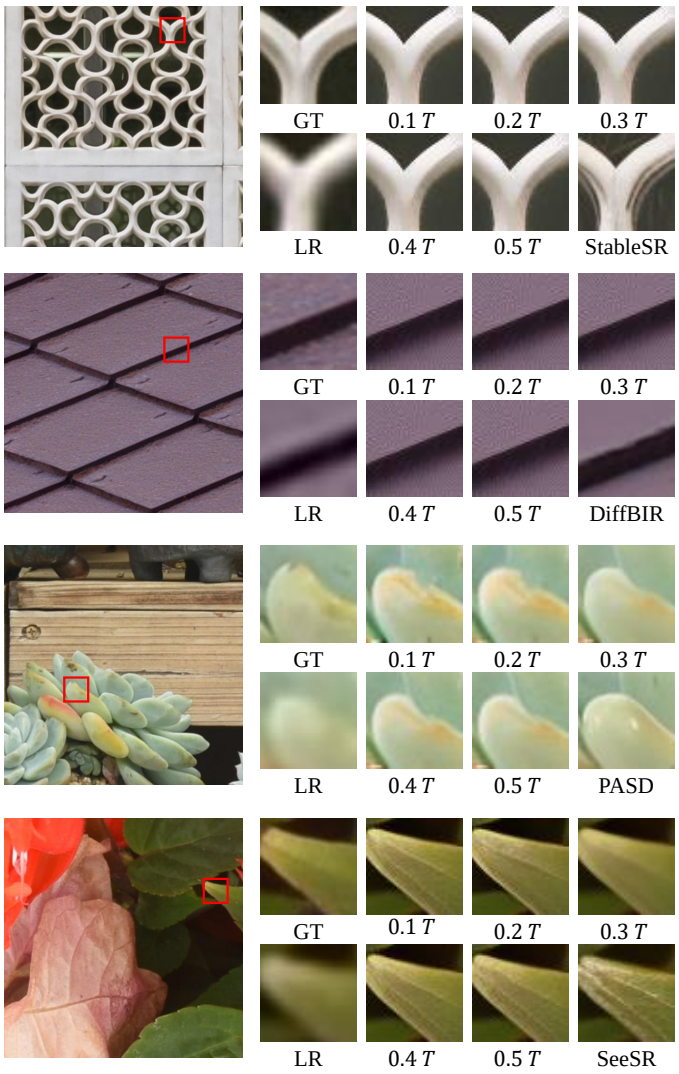


Figure 3: Qualitative comparisons on different setting of T_{SAS} with state-of-the-art diffusion-based Real-ISR baselines. Too-long early stage sacrifices the structural fidelity of reconstructed images. (Zoom in for details)