FIDELITY-PRESERVED DIFFUSION-BASED IMAGE SUPER-RESOLUTION WITH GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

While diffusion-based image super-resolution models have demonstrated remarkable performance in generating visually appealing images, they are prone to hallucination, often producing image details that appear deceptively realistic but are, in fact, spurious. We attribute this to the probabilistic generative nature of diffusion models, yet we contend that generation can be guided. To this end, this paper proposes a conditioning strategy that guides the diffusion process to preserve the fidelity with the given low-resolution (LR) observation, and consequently prevents hallucinatory detail artifacts. Specifically, we derive a theoretical formulation, enabling a unified conditional posterior score to guide the diffusion generation process, including both the training of the denoising autoencoder and the inference of super-resolved images. During training, the LR observation serves as a condition input to guide the denoising autoencoder, while the noise predicted by the autoencoder is used to calculate the consistency loss against the high-resolution (HR) image, steering the training to be well-aligned with the generation of super-resolved images. During inference, where the HR ground truth is unavailable, we instead pair the LR observation with itself. One serves as an input to the trained autoencoder, whereas the other is utilized to compute the residual against the inferred super-resolved image, thereby guiding the inference process. Experimental results demonstrate that our method effectively suppresses hallucinations, achieving significantly more faithful super-resolved results on digits, text characters, and texture details compared to competing methods.

1 INTRODUCTION

Image super-resolution (ISR) Lepcha et al. (2023) tries to recover the original High-Resolution (HR) image $x \in \mathbb{R}^l$ from its degraded Low-Resolution (LR) observation $y \in \mathbb{R}^m$ that is given by

$$y = \mathcal{A}(x) + n, \tag{1}$$

where $\mathcal{A}$: $\mathbb{R}^l \to \mathbb{R}^m$ refers to a degradation operator and n is Gaussian degradation noise. Since $m \ll l$, $\mathcal{A}$ is generally not invertible, making the recovery of the original x from y a highly ill-posed inverse problem. A wide spectrum of ISR methods, ranging from traditional interpolation to convolutional architectures and transformers Mei et al. (2025); Long et al. (2025); Tarasiewicz & Kawulok (2025), have been investigated to address this inherently ill-posedness. Recently, diffusion probabilistic models have gained widespread attention by leveraging their generative capabilities. Diffusion-based ISR methods Yue et al. (2025); Jeong et al. (2025) adopt a stochastic Bayesian viewpoint, aiming to generate super-resolved images by sampling from the posterior distribution $p(x|y)$, where Eq. (1) serves as the likelihood model for $p(y|x)$. Diffusion-based ISR typically comprises two stages: training and inference. During training, a denoising autoencoder ϵ_θ for $p(x|y)$ is learned using the LR image as an additional input to constrain its output space. At inference time, trained ϵ_θ is capable of generating plausible super-resolved images from the corresponding LR inputs following an inference algorithm such as the popular DDPM or DDIM Song et al. (2021).

Diffusion-based ISR methods have been shown to yield visually attractive super-resolved images. However, the inherent stochasticity of the diffusion model often leads to spurious image details that are essentially hallucinated by the model. In an effort to remedy this, recent approaches propose concatenating LR image features into the denoising U-Net backbone during the inference stage Wang et al. (2024); Wan et al. (2024); Li et al. (2025). This guides the image super-resolving process,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

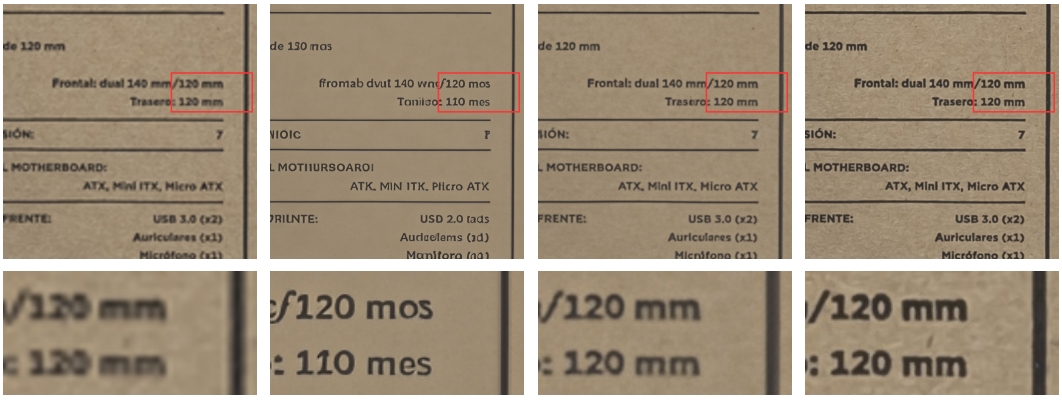

Figure 1: Hallucinatory artifacts. Diffusion-based ISR method StableSR has a significantly better visual effect than that of Bicubic interpolation, with character clarity that even surpasses the Ground Truth. But in the finer details, it exhibits obvious hallucinations, with "120 mm" being incorrectly imagined as "120 mos" and "110 mes".

trying to preserve fidelity to the original LR observations. Unfortunately, despite the various efforts made in existing studies, they still remain prone to hallucinations and produce fabricated content. As can be seen in Fig. 1, diffusion-based StableSR Wang et al. (2024) produces substantially superior visual results compared to the Bicubic interpolation approach. However, it also exhibits noticeable hallucination, incorrectly generating the digit and text characters "120 mm" (millimeter) as "120 mos" and "110 mes". This difficulty stems from insufficient guidance strength: injecting LR features into the U-Net backbone and then relying on the U-Net to guide super-resolving constitutes an implicit conditional guidance, which is indirect and thus inherently less effective.

In this paper, we propose to mitigate hallucination in super-resolved images by integrating direct and explicit conditional guidance throughout the entire diffusion process. Conditional guidance approaches Chung et al. (2023); Rout et al. (2023); Wang et al. (2025) are originally introduced to enable general diffusion models to adapt to various tasks, such as image super-resolving, by adding task-specific conditions during inference, thus eliminating the need to train dedicated diffusion models. While the original motivation was to save training costs, this paper instead harnesses conditional guidance to prevent hallucinations. We formalize the preservation of fidelity between super-resolved images and their LR observations as a specific task for the diffusion model, and then guide diffusion-based ISR to accomplish this fidelity-oriented task. We extend the explicit guiding strategy from the inference stage to both stages of training and inference, while retaining indirect U-Net guidance at each stage. This steers the diffusion process to ensure that the super-resolved images remain faithful to the LR observation, thereby effectively suppressing hallucinations.

The main contributions of this paper are summarized as follows:

1. We derive, from a Bayesian perspective, an analytical formulation of the conditional posterior score function under unified conditions, establishing a general theoretical foundation for incorporating both implicit and explicit conditional guidance in diffusion-based ISR.
2. We propose a conditioning mechanism that applies this posterior score to guide the entire diffusion pipeline, encompassing both denoising autoencoder training and super-resolving inference, thereby effectively curbing hallucinatory artifacts.
3. The proposed method is inherently plug-and-play, as it builds upon general diffusion-based ISR methods, thus leveraging their intrinsic super-resolving capabilities while supplementing them with explicit guidance.
4. Extensive experiments across various datasets demonstrate that the proposed method produces super-resolved images with significantly improved fidelity preservation in fine-grained details, such as digits, text, and textures.

2 Related Work

ISR aims to recover original HR images from their LR observations with the highest possible recovery accuracy Lepcha et al. (2023); Zamfir et al. (2023). Early ISR approaches relied on filtering techniques such as Bicubic interpolation, which are computationally efficient but yield limited accuracy. In recent years, deep learning models, including convolutional neural networks (CNNs) and transformers, have been extensively employed in ISR, achieving relatively higher accuracy Conde et al. (2022); Long et al. (2025). As a supervised learning paradigm, CNN- and transformer-based ISR networks typically set the minimization of the mean squared error between the super-resolved image and the corresponding HR ground truth as the training objective. Nevertheless, because the mean squared error is computed on a pixel-wise basis, they struggle to capture perception-critical discrepancies, such as fine texture details, ultimately leading to an unsatisfactory visual effect. To enhance the visual quality of super-resolved outputs, generative adversarial networks (GANs) have been adopted in ISR approaches Wang et al. (2021). While GAN-based ISR methods excel at generating super-resolved results with visually more attractive textures, they are nevertheless susceptible to mode collapse, a common training instability.

Diffusion models have been demonstrated to have generation capabilities beyond GAN and stationary training objectives Rombach et al. (2022). For diffusion models targeting image super-resolving task Saharia et al. (2022); Yue et al. (2024), an autoencoder with LR images as input is first trained to learn the corresponding LR information when predicting the noise vector, and then the trained denoising autoencoder is used to generate the super-resolved image with an inference algorithm. Diffusion-based ISR, which depend on generative priors, tend to produce outputs whose content often deviates from that of the input LR images. To mitigate this issue, StableSR Wang et al. (2024) introduces a feature transform module that transforms the LR image into representations compatible with the U-Net architecture within the autoencoder, trying to guide image generation to prevent deviation from the LR image. ControlSR Wan et al. (2024) incorporates a detail-preserving module that uses the latent LR embedding, enabling the generated images to retain the detailed information of the LR observations. Similarly, StructSR Li et al. (2025) adopts a structure-aware screening mechanism to identify the generated super-resolved images with highest structural similarity to their LR counterparts. However, all these methods rely on an indirect cascade, where the LR observation first guides the autoencoder, which in turn guides the generation of the SR image. Such an implicit guidance paradigm frequently leads to the generation of spurious texture details that do not align with the original LR observation. This discrepancy is widely referred to as hallucination.

Another line of work on diffusion models is conditional guidance Chung et al. (2023); Rout et al. (2023); Wang et al. (2025), which does not train dedicated diffusion models, but instead guides general diffusion models during inference towards specific tasks, such as image synthesis and super-resolving, thereby saving the cost of model training. In this paper, we abstract the fidelity between the generated super-resolved image and the given LR observation as a target task for the diffusion model, and then adopt the conditioning mechanism to guide the model toward this task. While conditional guidance was originally introduced to save training cost, our goal is to mitigate the hallucination issue commonly observed in diffusion-based ISR.

3 Method

This section begins with an introduction to the foundational unconditional diffusion model, followed by the diffusion model incorporating implicit guidance for ISR. We then define an explicit condition and consolidate both implicit and explicit guidance into a unified Bayesian framework. Leveraging this framework, we formulate the training algorithm for the denoising autoencoder and the inference algorithm for generating super-resolved images.

3.1 Diffusion Model and Diffusion-Based ISR

Diffusion models are a class of probabilistic generative models that sample from a given data distribution by learning to reverse a forward noising process. By utilizing trained compression models composed of encoder $\mathcal{E}(\cdot)$ and decoder $\mathcal{D}(\cdot)$, diffusion models can be transferred from the data space to a computationally much more efficient latent space. In the diffusion model, the forward noising process can be formulated as a Stochastic Differential Equations (SDE) $\mathrm{d}z_t =$

$f(z_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}w_t$, where z_t denotes the noisy version of the latent representation z_0 at time step t (with $t = 0, 1, \cdots, T$), $f(z_t, t)$ is the drift term, $g(t)$ is the diffusion coefficient, and w_t is the standard Wiener process. The reverse generating process proceeds in the opposite direction, gradually recovering z_0 from a noise. This can be written as the reverse-time SDE:

$$\mathrm{d}z_t = \left[f(z_t, t) - g(t)^2 \nabla_{z_t} \log p(z_t)\right] \mathrm{d}t + g(t)\mathrm{d}\tilde{w}_t, \tag{2}$$

where time flows in the reverse direction, i.e., $t = T, T-1, \cdots, 0$, and $d\tilde{w}_t$ denotes the associated reverse-time Wiener process. Note that the score function, $\nabla_{z_t} \log p(z_t)$, can be estimated by training a denoising autoencoder $\epsilon_\theta(z_t, t)$.

Diffusion-based ISR adapts the foundational unconditional diffusion model to the task of generating an HR image from an LR input. Compared to Eq. (2), the reverse-time SDE for the diffusion-based ISR is given by

$$\mathrm{d}z_t = \left[f(z_t, t) - g(t)^2 \nabla_{z_t} \log p(z_t|y)\right] \mathrm{d}t + g(t)\mathrm{d}\tilde{w}_t. \tag{3}$$

The key difference lies in the score function, $\nabla_{z_t} \log p(z_t|y)$, where LR observation y acts as an additional condition, and $p(z_t|y)$ is the conditional distribution of z_t given y. Accordingly, the general autoencoder $\epsilon_\theta(z_t, t)$ is extended to a conditional $\epsilon_\theta(z_t, y, t)$.

For diffusion-based ISR, the reverse generative process consists of two stages: (i) training, where a denoising autoencoder $\epsilon_\theta(z_t, y, t)$ is learned with LR observation y serving as a conditional input; and (ii) inference, where the trained autoencoder is used in a step-wise denoising inference algorithm to produce a super-resolved image. Observe that during inference, condition y is conveyed via $\epsilon_\theta(z_t, y, t)$, making the guiding indirect, and we refer to it as implicit guidance.

3.2 Conditional Guiding for Training and Inference

3.2.1 Conditional Guidance

We first define a condition pair $\{y, \tilde{y}\}$, where y and $\tilde{y}$ represent the implicit and explicit conditions, respectively. y is an LR image that is consistent with the general diffusion-based ISR shown in Eq. (3), let x_0 denote an HR ground-truth image, and $\tilde{y}$ is set to

$$\tilde{y} = \tilde{\mathcal{A}}(x_0) + n, \tag{4}$$

where $\tilde{\mathcal{A}}$ is some measurement operator. Note that the specific form of $\tilde{\mathcal{A}}$ depends on the context, namely, whether it is used for denoising autoencoder training or for inference of the super-resolved image. With the condition pair $\{y, \tilde{y}\}$, Eq. (3) can be formulated into a more general form:

$$\mathbf{d}z_t = \left[f(z_t, t) - g(t)^2 \nabla_{z_t} \log p(z_t \mid y, \tilde{y})\right] \mathrm{d}t + g(t)\,\mathrm{d}\tilde{w}_t. \tag{5}$$

We then derive score function $\nabla_{z_t} \log p(z_t|y, \tilde{y})$ from a Bayesian perspective as follows.

$$\nabla_{z_t} \log p(z_t|y, \tilde{y}) = \nabla_{z_t} \log \frac{p(\tilde{y}|z_t, y)\, p(z_t|y)}{p(\tilde{y}|y)} \tag{6}$$

$$= \nabla_{z_t} \log p(\tilde{y}|z_t, y) + \nabla_{z_t} \log p(z_t|y) \tag{7}$$

$$\simeq \nabla_{z_t} \log p(\tilde{y}|z_t, y) - \frac{\epsilon_\theta(z_t, y, t)}{\sqrt{1-\bar{\alpha}_t}}. \tag{8}$$

Because $\nabla_{z_t} \log p(\tilde{y}|y)$ does not involve z_t, the gradient is zero, and thus Eq. (7) follows directly from Eq. (6). $\nabla_{z_t} \log p(z_t|y)$ can be identified as the conditional posterior score that can be approximated by denoising autoencoder $\epsilon_\theta(z_t, y, t)$, and the approximation is $-\frac{\epsilon_\theta(z_t, y, t)}{\sqrt{1-\bar{\alpha}_t}}$, leading to the derivation of Eq. (8) from Eq. (7). Here, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative signal-retention coefficient, $\alpha_s = 1 - \beta_s$, and $\beta_s \in (0, 1)$ is the noise scheduling coefficient at the s-th time step.

We next derive the gradient $\nabla_{z_t} \log p(\tilde{y}|z_t, y)$ in Eq. (8), and the derivation proceeds as follows.

$$\nabla_{z_t} \log p(\tilde{y} \mid z_t, y) = \nabla_{z_t} \log \int p(\tilde{y} \mid x_0)\, p(x_0 \mid z_t, y)\, \mathrm{d}x_0 \tag{9}$$

$$\simeq \nabla_{z_t} \log p(\tilde{y}|\hat{x}_0(z_t)) \tag{10}$$

$$\simeq \nabla_{z_t} \log \exp(-\frac{1}{2\sigma_t^2} \|\tilde{y} - \tilde{\mathcal{A}}(\hat{x}_0(z_t))\|_2^2) \tag{11}$$

$$\simeq -\eta_t \nabla_{z_t} \|\tilde{y} - \tilde{\mathcal{A}}(\hat{x}_0(z_t))\|_2^2. \tag{12}$$

The derivation of $\nabla_{z_t} \log p(\tilde{y}|z_t, y)$ starts by applying the law of total probability to decompose $p(\tilde{y}|z_t, y)$ via marginalization over x_0, yielding Eq. (9). To make the problem tractable, the conditional expectation is approximated by replacing the posterior distribution $p(x_0|z_t, y)$ with a point mass at its mean estimate $\hat{x}_0(z_t)$, commonly referred to as the Tweedie estimate Efron (2011), which leads to Eq. (10). Considering a Gaussian observation likelihood at time step t, i.e., $n \sim \mathcal{N}(0, \sigma_t^2 I)$, one has $p(\tilde{y}|\hat{x}_0(z_t)) \propto \exp\left(-\frac{1}{2\sigma_t^2}\|\tilde{y} - \tilde{\mathcal{A}}(\hat{x}_0(z_t))\|_2^2\right)$. Taking the logarithm of both sides gives Eq. (11). Finally, computing the gradient with respect to z_t yields Eq. (12), where η_t is defined as the step size at time step t in the diffusion process, which is proportional to $\frac{1}{2\sigma_t^2}$.

In this way, the score function $\nabla_{z_t} \log p(z_t|y, \tilde{y})$ is ultimately expressed as:

$$\nabla_{z_t} \log p(z_t|y, \tilde{y}) \simeq -\eta_t \nabla_{z_t} \|\tilde{y} - \tilde{\mathcal{A}}(\hat{x}_0(z_t))\|_2^2 - \frac{\epsilon_\theta(z_t, y, t)}{\sqrt{1 - \bar{\alpha}_t}}, \tag{13}$$

where the step size η_t varies along with time step t. The first term in Eq. (13), i.e., -$\eta_t \nabla_{z_t} \|\tilde{y} - \tilde{\mathcal{A}}(\hat{x}_0(z_t))\|_2^2$ indicates the explicit observation constraint, which guides the Tweedie estimate of the super-resolved image toward consistency with condition $\tilde{y}$. The second term, predicted by the diffusion denoising autoencoder, represents the generative prior implicitly conditioned on y.

3.2.2 Training and Inference with Conditional Guidance

The processes of training a denoising autoencoder and inferring a super-resolved image are outlined in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Training a Denoising Autoencoder

Require: $\{x_0, y\}, \mathcal{E}(\cdot), \mathcal{D}(\cdot), T$;
Ensure: Autoencoder model parameters θ;
1: **repeat**
2: $\tilde{y} \leftarrow \mathcal{V}(x_0), \tilde{\mathcal{A}}(\cdot) \leftarrow \mathcal{V}(\cdot)$;
3: $z_0 \leftarrow \mathcal{E}(x_0)$;
4: $t \sim \text{Uniform}(\{1, \ldots, T\})$;
5: $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$;
6: $z_t \leftarrow \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$;
7: $\hat{z}_0(z_t) \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} z_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \epsilon_\theta(z_t, y, t)$;
8: $\hat{x}_0(z_t) \leftarrow \mathcal{D}(\hat{z}_0(z_t))$;
9: $\mathcal{L}_{\text{visual}} \leftarrow \|\tilde{y} - \tilde{\mathcal{A}}(\hat{x}_0(z_t))\|_2^2$;
10: $\hat{\epsilon}_\theta \leftarrow \eta_t \sqrt{1 - \bar{\alpha}_t} \nabla_{z_t} \mathcal{L}_{\text{visual}} + \epsilon_\theta(z_t, y, t)$;
11: Take gradient descent step on: $\nabla_\theta \|\epsilon - \hat{\epsilon}_\theta\|_2^2$;
12: **until** convergence

Algorithm 2 Inferring a Super-Resolved Image

Require: $y, \mathcal{D}(\cdot), T$;
Ensure: Generated HR image $\mathcal{D}(z_0)$;
1: $z_T \sim \mathcal{N}(0, \boldsymbol{I})$;
2: **for** $t = T, T-1, \ldots, 1$ **do**
3: $\tilde{y} \leftarrow \phi_t(y), \tilde{\mathcal{A}}(\cdot) \leftarrow \phi_t(\mathcal{A}(\cdot))$;
4: $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$;
5: $\hat{\epsilon}_\theta \leftarrow \epsilon_\theta(z_t, y, t)$;
6: $\hat{z}_0(z_t) \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} z_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \hat{\epsilon}_\theta$;
7: $z'_{t-1} \leftarrow \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} z_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \hat{z}_0(z_t)$;
8: $z_{t-1} \leftarrow z'_{t-1} - \eta_t \nabla_{z_t} \|\tilde{y} - \tilde{\mathcal{A}}(\mathcal{D}(\hat{z}_0(z_t)))\|_2^2$;
9: **end for**
10: **return** $\mathcal{D}(z_0)$.

In the training stage, the training objective is to allow autoencoder $\epsilon_\theta(z_t, y, t)$ to approximate posterior score $\nabla_{z_t} \log p(z_t|y, \tilde{y})$ with condition pair $\{y, \tilde{y}\}$. Notice that during the supervised training, HR image x_0 is available and can serve as explicit condition $\tilde{y}$ alongside LR observation y. We propose to carry out the guidance in the frequency domain, and thus instantiate $\tilde{y}$ as $\mathcal{V}(x_0)$, where $\mathcal{V}(\cdot)$ represents a frequency-domain operator. The corresponding measurement operator $\tilde{\mathcal{A}}(\cdot)$ is instantiated as $\mathcal{V}(\cdot)$ that is further defined as $\mathcal{V}(\cdot) = \text{Norm}(\log(1 + |\text{Shift}(\mathcal{F}(\cdot))|))$. Here, $\text{Shift}(\cdot)$ denotes spectrum centering, $\text{Norm}(\cdot)$ denotes statistical normalization, and the two-dimensional Fourier transform $\mathcal{F}(\cdot)$ is given by $\mathcal{F}(u, v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h, w) \exp\left(-2\pi i \left(\frac{uh}{H} + \frac{vw}{W}\right)\right)$ where H and W are the height and width of image x_0, h and w are spatial indices, and u and v are frequency-domain coordinate indices. The operator $\mathcal{V}(\cdot)$ then performs spectrum centering, logarithmic magnitude compression, and zero-mean unit-variance normalization.

In Algorithm 1, Step 6 applies the general forward noising operation to obtain z_t, the noisy version of z_0 at time step t. The corresponding Tweedie approximation, represented as $\hat{z}_0(z_t)$, is achieved through Step 7. This approximation is subsequently decoded into the pixel space using the decoder $\mathcal{D}(\cdot)$ at Step 8. At Step 9, we formulate the residual between the Tweedie approximation $\hat{x}_0(z_t)$ and the HR ground truth image as the visual consistency loss $\mathcal{L}_{\text{visual}}$, defined as

$$\mathcal{L}_{\text{visual}} = \left\|\tilde{y} - \tilde{\mathcal{A}}(\hat{x}_0(z_t))\right\|_2^2. \tag{14}$$

This consistency loss directly regularizes the frequency-domain representation, and is expected to be able to suppress texture shifts and thus mitigate hallucinated textures. We further construct a new noise prediction $\hat{\epsilon}_\theta$ to exactly match the posterior score $\nabla_{z_t} \log p(z_t|y, \tilde{y})$. At this time, one has $-\frac{\hat{\epsilon}_\theta}{\sqrt{1-\bar{\alpha}_t}} \simeq \nabla_{z_t} \log p(z_t|y, \tilde{y})$, where the factor $\sqrt{1-\bar{\alpha}_t}$ is used to match the gradient injection strength with the noise level at time step t. According to Eq. (13), one has the following conditional noise estimate:

$$\hat{\epsilon}_\theta \simeq \epsilon_\theta(z_t, y, t) + \eta_t \sqrt{1-\bar{\alpha}_t}\, \nabla_{z_t} \mathcal{L}_{\text{visual}}, \tag{15}$$

as shown in Step 10, where step size η_t is used as the trade-off hyper-parameter that balances visual consistency loss and the predicted noise. Different from directly adding $\mathcal{L}_{\text{visual}}$ as an auxiliary loss term, its gradient with respect to z_t is injected into each training iteration of the noise prediction process. Consequently, the final training objective is defined as $\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_t, \boldsymbol{\epsilon}, t}\left[\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta\|_2^2\right]$, as indicated in Step 11. During the training stage, the encoder $\mathcal{E}(\cdot)$ and decoder $\mathcal{D}(\cdot)$ are frozen, and only the parameter set θ of the denoising autoencoder is updated.

During inference, HR images are not available and the inference process relies solely on LR observation y. We then leverage y and itself to construct a condition pair. Notably, we apply this conditioning to the frequency-domain representation to explicitly exploit the frequency structure of the inference procedure. In this way, condition $\tilde{y}$ is instantiated as $\phi_t(y)$, while the measurement operator $\tilde{\mathcal{A}}(\cdot)$ is instantiated as $\phi_t(\mathcal{A}(\cdot))$, where $\phi_t(\cdot)$ represents time-varying frequency-domain low-pass filter at the t-th iteration and $\mathcal{A}(\cdot)$ denotes the degradation operation as defined in Eq. (1).

In Algorithm 2, Steps 4 through 7 follow the popular DDPM inference process. Specifically, at Step 5, the denoising autoencoder outputs the predicted noise $\epsilon_\theta(z_t, y, t)$, where y is the implicit conditional input. At Step 6, estimated posterior mean $\hat{z}_0$ at t-th iteration, $\hat{z}_0(z_t)$, is calculated by Tweedie's formula, and at Step 7, the intermediate variable of z_{t-1}, z'_{t-1}, is obtained using the standard DDPM inference update. The guidance with condition $\tilde{y}$ is realized in Step 8 of the algorithm. Recall that the score function $\nabla_{z_t} \log p(\tilde{y}|z_t, y)$ characterizes the residual between the posterior mean conditioned on $\{z_t, y\}$ and the LR observation $\tilde{y}$. According to Eq. (12), this score is approximated by $-\frac{1}{2\sigma_t^2} \nabla_{z_t} \left\| \tilde{y} - \tilde{\mathcal{A}}(\hat{x}_0(z_t)) \right\|_2^2$, where $\hat{x}_0(z_t) = \mathcal{D}(\hat{z}_0(z_t))$ is the Tweedie estimate in pixel space at the current time step t. Intermediate latent variable z'_{t-1} is then updated to z_{t-1} at Step 8, where η_t is used to express the step size that is proportional to $\frac{1}{2\sigma_t^2}$.

As demonstrated in Algorithm 2, since HR images are unavailable during inference, we reuse the LR observation as the conditions y and $\tilde{y}$. These implicit and explicit conditions are respectively leveraged to steer the denoising autoencoder in predicting the noise, and to guide the iterative inference procedure. This design ensures that the super-resolved output remains consistent with the LR observation, trying to prevent the occurrence of hallucinations.

3.3 Discussion

To mitigate hallucinatory artifacts in diffusion-based ISR, we introduce a novel explicit condition $\tilde{y}$ that, together with the commonly-used indirect condition y in ISR, constitutes a condition pair $\{y, \tilde{y}\}$. It should be noted that we are not imposing an additional constraint in this context. For autoencoder training, $\tilde{y}$ corresponds to the HR images already present in the training set, and for inference, it is the LR observation, i.e., the reuse of condition y. In our framework, y is used for implicit guidance through denoising autoencoder $\epsilon_\theta(z_t, y, t)$ to guide the noise prediction, while $\tilde{y}$ is used for explicit guidance that is incorporated at every iteration step in both Algorithms 1 and 2. By pervading guidance throughout the entire diffusion process, this design aims to ensure that the super-resolved image remains faithful to the LR observation and the HR ground truth.

As can be seen from Algorithms 1 and 2, our guidance is performed in the frequency domain rather than in the pixel domain. The underlying rationale is to enforce consistency between the super-resolved image generated by the diffusion model and the corresponding LR observation in both low- and high-frequency components. Specifically, this ensures that the overall style of the super-resolved image aligns with that of the LR input, while simultaneously preventing hallucination artifacts in high-frequency details.

Our method trades computational cost for improved accuracy, as the guidance step in the training and inference algorithm introduces additional overhead. Fortunately, this guidance is plug-and-play.

When sufficient computational resources are available, incorporating the guidance step into Algorithms 1 and 2 effectively mitigates hallucination artifacts. In use-case scenarios such as video-on-demand, where end-users may not be particularly sensitive to fine image details, the guidance step can be safely omitted without affecting the overall algorithm functionality.

4 EXPERIMENTS

This section evaluates the proposed method and compares it with competing ones, including GAN-based RealESRGAN+ Wang et al. (2021), and diffusion-based LDM Rombach et al. (2022), StableSR Wang et al. (2024), ControlSR Wan et al. (2024), and StructSR Li et al. (2025). All these competing methods are implemented with their open-source codes and official pre-trained model weights.

4.1 EXPERIMENTAL SETUP

The training sets are DIV2K, Flickr2K, DIV8K, and OutdoorSceneTraining, which follow the same training set setup as StableSR Wang et al. (2024). During training, these datasets are used to construct the original HR image pool, and LR images are generated online using the two-stage RealESRGAN degradation strategy. During testing, RealSR Cai et al. (2019) and DRealSR Wei et al. (2020) datasets are used. To ensure fair comparison under the same input condition, 512×512 patches are cropped from the center of HR images, and anti-aliased bicubic downsampling with a scale factor of $\times 4$ is used to generate corresponding LR images of size 128×128.

All training and testing experiments are conducted on a hardware platform equipped with an Intel Core i9-14900KF CPU and a single NVIDIA GeForce RTX 4090D GPU with 24 GB memory. The system runs Windows 11. The software environment is Python 3.10.18 and CUDA 11.8, and the deep learning framework is implemented with PyTorch 2.1.0 and PyTorch Lightning 2.5.6. In the training procedure outlined in our Algorithm 1 of Section 3.2.2, the denoising autoencoder ϵ_θ is initialized using the pre-trained weights from StableSR. AdamW is used as the optimizer with a learning rate of 5×10^{-5}. The batch size is set to 2, and four gradient accumulation steps are used to obtain an effective batch size of 8. The model is trained for 15,000 iterations where the time step T is set to 1,000. In Algorithm 1, parameter η_t is fixed to 1. In Algorithm 2, however, parameter η_t varies with the iteration step t. During inference, the diffusion model adopts the DDPM scheduling scheme, where the number of diffusion steps is reduced from 1,000 to 200 with a resampling strategy. The downsampling operator $\mathcal{A}(\cdot)$ is implemented as differentiable anti-aliased bicubic downsampling, and the random seed is 42.

The evaluation is performed in terms of both objective quantitative metrics and subjective visual effect. Quantitative evaluation uses full-reference and no-reference metrics. Full-reference metrics include peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS) Zhang et al. (2018). PSNR and SSIM are calculated on the Y channel in the YCbCr color space to evaluate luminance fidelity and structural consistency. LPIPS measures the perceptual difference between the super-resolved images and the reference ones based on features of a pre-trained neural network. No-reference metrics include CLIP image quality assessment (CLIP) Wang et al. (2023), multi-scale image quality transformer (MUSIQ) Ke et al. (2021), and natural image quality evaluator (NIQE), which evaluate perceptual quality from the perspectives of vision-language semantic priors, multi-scale quality prediction, and natural scene statistics, respectively.

4.2 PERFORMANCE EVALUATION

4.2.1 QUANTITATIVE COMPARISON

Table 1 compares our method with the competing ones on RealSR and DRealSR datasets, where symbols $\uparrow$ and $\downarrow$ indicate that higher and lower scores correspond to better performance on the corresponding evaluation metric, respectively.

Across all full-reference metrics including PSNR, SSIM, and LPIPS scores, our proposed method consistently achieves the best performance scores on both RealSR and DRealSR datasets, significantly outperforming GAN-based RealESRGAN+, as well as all diffusion-based LDM, StableSR,

ControlSR, and StructSR. It should be noted that, for the no-reference metrics CLIP, MUSIQ, and NIQE, competing methods achieve better performance scores. It should also be noticed that, these methods yield numerical scores that even exceed those of the ground truth, which serves as the ideal reference. This suggests that, from a purely visual perspective, the super-resolved images generated by these methods may appear even sharper than the ground truth. However, it also indicates that these super-resolved images may constitute hallucinated content that is inconsistent with the ground truth. We will further verify this using the visual examples shown in Fig. 2.

Table 1: Quantitative results on RealSR and DRealSR. Note that we include the performance scores for the ground truth as the ideal reference for comparison.

Method	RealSR						DRealSR					
	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	MUSIQ↑	NIQE↓	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	MUSIQ↑	NIQE↓
RealESRGAN+	27.196	0.8024	0.2093	0.5376	66.947	5.7595	31.906	0.8768	0.2002	0.5340	61.458	6.7209
LDM	28.803	0.8105	0.1776	0.6020	65.783	6.4238	33.648	0.8835	0.1749	0.5977	59.857	6.7645
StableSR	25.014	0.7417	0.2646	0.6665	70.039	5.8895	28.876	0.7987	0.2830	0.6886	65.803	6.7183
StructSR	25.543	0.7490	0.2990	**0.7156**	69.755	6.7630	29.348	0.8225	0.3023	**0.7187**	65.464	7.9381
ControlSR	26.356	0.7255	0.2918	0.6877	**71.033**	**5.2505**	30.256	0.8046	0.2927	0.6941	**67.285**	**6.4165**
Ours	**31.198**	**0.8526**	**0.1389**	0.4293	54.843	6.5288	**35.447**	**0.8988**	**0.1474**	0.4566	47.972	7.3861
Ground Truth	∞	*1.0000*	*0.0000*	*0.4586*	*57.454*	*6.1168*	∞	*1.0000*	*0.0000*	*0.4743*	*50.560*	*6.7913*

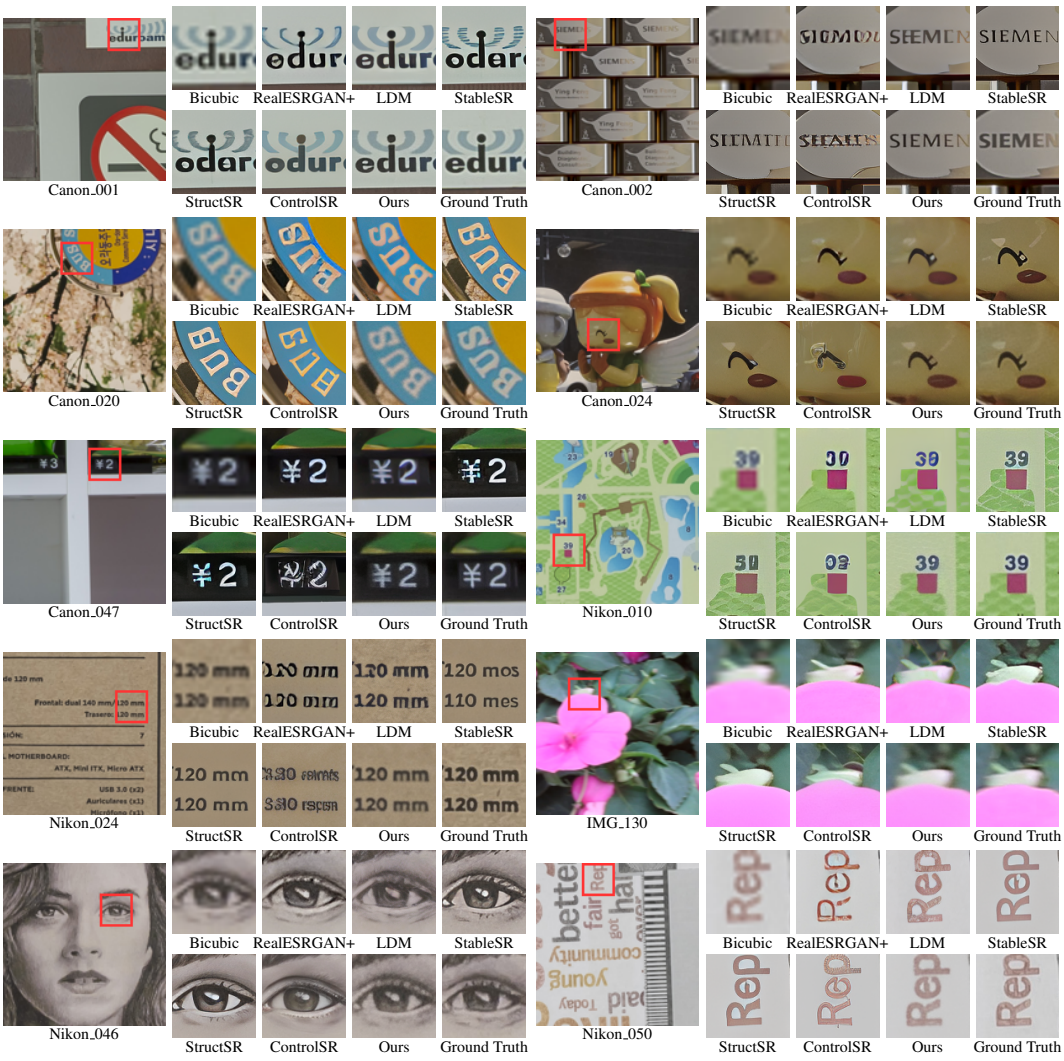

Figure 2: Visual comparison of the super-resolved images with various ISR methods: focusing on fidelity.

Table 2: Ablation Study. '✓' and '–' denote the explicit $\tilde{y}$ guidance is enabled and disabled, respectively. The best results are highlighted in bold.

Training	Inference	RealSR			DRealSR		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
–	–	25.014	0.7417	0.2646	28.876	0.7987	0.2830
✓	–	25.589	0.7466	0.2555	29.179	0.7973	0.2701
–	✓	30.805	0.8423	0.1562	35.064	0.8889	0.1658
✓	✓	**31.198**	**0.8526**	**0.1389**	**35.447**	**0.8988**	**0.1474**

4.2.2 Visual Comparison

Fig. 2 compares the visual effects of super-resolved images using the proposed method and the competing ones, with a particular focus on fine-grained details such as digits, characters, and textures.

Our method exhibits not only perceptual naturalness and favorable visual effects but also, more importantly, it mitigates the hallucination issue commonly observed in GAN- and diffusion-based ISR methods. In comparison to competing ones, the super-resolved images produced by our method show obviously better fidelity to the ground truth. Taking the image 'Canon_001' from the RealSR dataset, shown in the top-left panel of Fig. 2, as an example, Bicubic interpolation yields noticeably blurred results, RealESRGAN+ introduces conspicuous distortions, and both StableSR and StructSR also exhibit unnatural texture artifacts. Note that LDM and ControlSR obtain relatively better visual quality. However, they suffer from hallucinations. Specifically, LDM misgenerates the character "e" as "θ", while ControlSR produces "o" instead. By contrast, our method delivers the best performance in preserving visual fidelity.

4.3 Ablation Study

Table 2 reports the experimental results of the ablation study, while also testing the plug-and-play capability of explicit $\tilde{y}$ guidance within our method. It can be observed that, when explicit condition $\tilde{y}$ is used in both stages of training and inference, the super-resolved images achieve the best PSNR, SSIM, and LPIPS scores on both RealSR and DRealSR datasets, indicating that the dual-stage guidance can effectively improve the fidelity of the super-resolved image. It can also be observed that enabling the explicit guidance only during training improves performance scores, showing that it can help enhance the perceptual similarity between the super-resolved image and the ground-truth one. When explicit guidance only during inference is enabled, the performance score improvement is more obvious, indicating that it has a more direct effect on achieving texture fidelity at the inference time. Overall, the dual-stage conditional guidance is more conducive to obtaining high-fidelity super-resolved results. The ablation study also provides a test of the plug-and-play property of our proposed conditioning module. Specifically, moving from the fourth row to the first row in Table 2 corresponds to gradually removing the conditioning module, until the method in the first row reduces to a general diffusion-based ISR approach, namely, StableSR in this ablation study.

5 Conclusion

Diffusion-based ISR delivers globally plausible super-resolved images, but the local details are prone to hallucination and substantial deviation from the LR observations. In this paper, we propose to provide continuous guidance throughout the entire diffusion process, covering both training and inference, by applying implicit and explicit guidance at the respective stages, thereby aligning the super-resolved image to prevent hallucinatory artifacts. We perform this conditioning guidance in the frequency domain rather than in the pixel domain, which is more conducive to preserving fidelity in fine-grained image details. Furthermore, our guiding strategy is plug-and-play. It can be progressively removed from both the training procedure and the inference algorithm, thereby enabling flexible adaptation to available computational resources. Experimental results demonstrate that, compared to existing methods, our method achieves significantly higher visual fidelity and effectively mitigates hallucinations.

REFERENCES

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3086–3095, 2019. doi: 10.1109/ICCV.2019.00318.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *European conference on computer vision*, pp. 669–687. Springer, 2022.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Jinho Jeong, Sangmin Han, Jinwoo Kim, and Seon Joo Kim. Latent space super-resolution for higher-resolution image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2355–2365, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021. doi: 10.1109/ICCV48922.2021.00510.

Dawa Chyophel Lepcha, Bhawna Goyal, Ayush Dogra, and Vishal Goyal. Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 91: 230–260, 2023.

Yachao Li, Dong Liang, Tianyu Ding, and Sheng-Jun Huang. StructSR: Refuse spurious details in real-world image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5022–5030, 2025. doi: 10.1609/aaai.v39i5.32532.

Wei Long, Xingyu Zhou, Leheng Zhang, and Shuhang Gu. Progressive focused transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2279–2288, 2025.

Kangfu Mei, Hossein Talebi, Mojtaba Ardakani, Vishal M Patel, Peyman Milanfar, and Mauricio Delbracio. The power of context: How multimodality improves image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23141–23152, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj"orn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49960–49990, 2023. doi: 10.52202/075280-2174.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Tomasz Tarasiewicz and Michal Kawulok. A graph attention network for real-world multi-image super-resolution. *Information Fusion*, 124:103325, 2025.

Yuhao Wan, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, Ming-Ming Cheng, and Bo Li. ControlSR: Taming diffusion models for consistent real-world image super resolution. *arXiv preprint arXiv:2410.14279*, 2024.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. doi: 10.1007/s11263-024-02168-7.

Xiaodong Wang, Ping Wang, Zhangyuan Li, and Xin Yuan. Plug-and-play diffusion models for inverse problems with data consistency projection. In *Proceedings of the IEEE International Conference on Image Processing Workshops*, pp. 350–355, 2025. doi: 10.1109/ICIPW68931.2025.11386346.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021. doi: 10.1109/ICCVW54120.2021.00217.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pp. 101–117, 2020. doi: 10.1007/978-3-030-58598-3_7.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 116–130, 2024.

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23153–23163, 2025.

Eduard Zamfir, Marcos V Conde, and Radu Timofte. Towards real-time 4k image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1522–1532, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

A APPENDIX

In the appendix, we provide additional quantitative and qualitative comparisons to complement the experimental results in the main paper.

We further evaluate the performance of the proposed method on the additional DIV2K validation set. The experimental results are presented in Table 3. It can be observed that our proposed method achieves a performance on the DIV2K validation set comparable to that on the RealSR and DRealSR datasets shown in Table 1. In terms of PSNR and SSIM, the most representative metrics for image similarity, our method has significantly higher scores than all competing methods. Metric LPIPS and no-reference metrics CLIP, MUSIQ, and NIQE are more indicative of perceptual quality and are much less capable of reflecting the fidelity of super-resolved images. As shown in Table 3, existing methods such as ControlSR and StructSR even outperform the ground truth in terms of CLIP and MUSIQ. This also suggests that existing methods strive to generate more visually appealing images rather than images consistent with the ground truth, and thus fall short in terms of fidelity.

Table 3: Quantitative results on DIV2K validation set, where the performance scores corresponding to the ground truth is given as a reference for comparison.

Method	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	MUSIQ↑	NIQE↓
RealESRGAN+	26.454	0.7326	0.2291	0.5341	64.213	4.8928
LDM	27.190	0.7321	**0.2070**	0.5971	63.840	5.2807
StableSR	24.346	0.6647	0.2357	0.7008	68.713	4.7258
StructSR	24.880	0.6705	0.2706	**0.7521**	68.136	6.1609
ControlSR	25.403	0.6502	0.2705	0.7040	**69.295**	**4.6508**
Ours	**28.302**	**0.7482**	0.2275	0.4579	51.759	5.6841
Ground Truth	∞	*1.0000*	*0.0000*	*0.6009*	*63.757*	*4.4268*

Figs. 3 and 4 present additional visual comparisons of more super-resolved images produced by various ISR methods. The super-resolved images reconstructed by the conventional bicubic method are noticeably blurry and clearly inferior to those produced by other ISR methods. Visually, RealESRGAN+, LDM, StableSR, StructSR, and ControlSR sometimes appear to produce sharper results than our method, even surpassing the ground truth. This is consistent with the numerical results for the performance metrics in Tables 1 and 3, where existing methods achieve higher CLIP and MUSIQ scores than our method and the ground truth. However, upon closer inspection, this apparent advantage is often illusory, since the ground truth itself is sometimes not sharp, whereas our method is closer to the ground truth and thus achieves better fidelity.

Taking Canon_045 in the sixth row of Fig. 3 as an example, we analyze the super-resolved images generated by the different methods from two perspectives: visual quality and fidelity. If we ignore the ground truth and consider only visual quality, all methods except Bicubic, which produces noticeably blurry results, yield relatively clear super-resolved images, including our method. In particular, StableSR and StructSR appear to produce even sharper images. However, when fidelity is evaluated against the ground truth, our method generates the super-resolved image that is more similar to the ground truth, from the background grayscale to the foreground characters "36", which means that our method has higher fidelity.

Fig. 4 further compares the visual quality of several sub-regions within the same image. When only the super-resolved images are examined, without reference to the ground truth, existing methods, including LDM, StableSR, ControlSR, and StructSR, often exhibit strong visual quality. This also explains why they achieve high scores on no-reference performance metrics such as CLIP and MUSIQ. However, when their local details are compared with the ground truth, these methods are found to produce hallucinations, i.e., they appear visually pleasing but are not faithful.

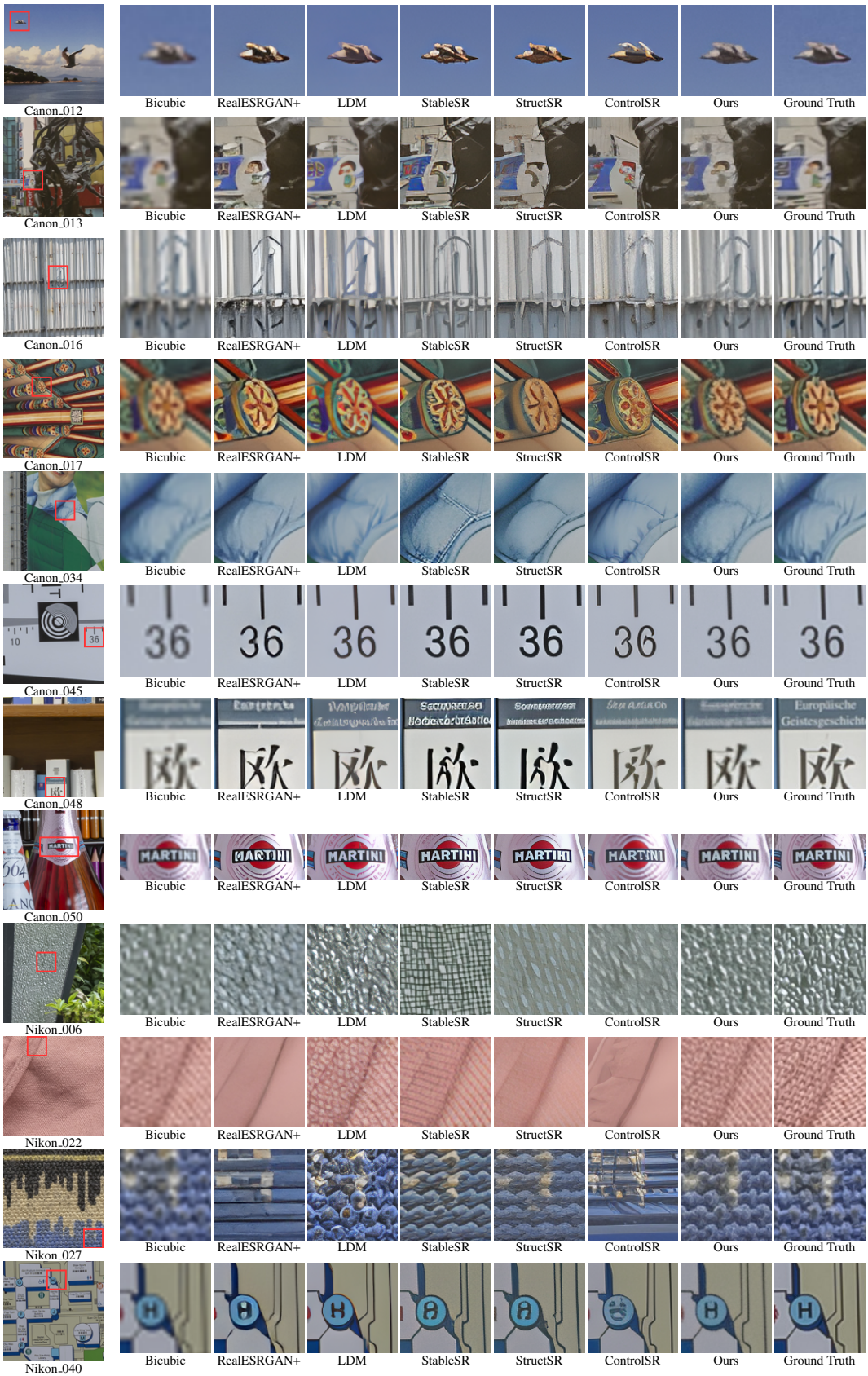

Figure 3: Visual effect of super-resolved images. Please pay particular attention to the fidelity comparison between each super-resolved image and its corresponding ground truth.

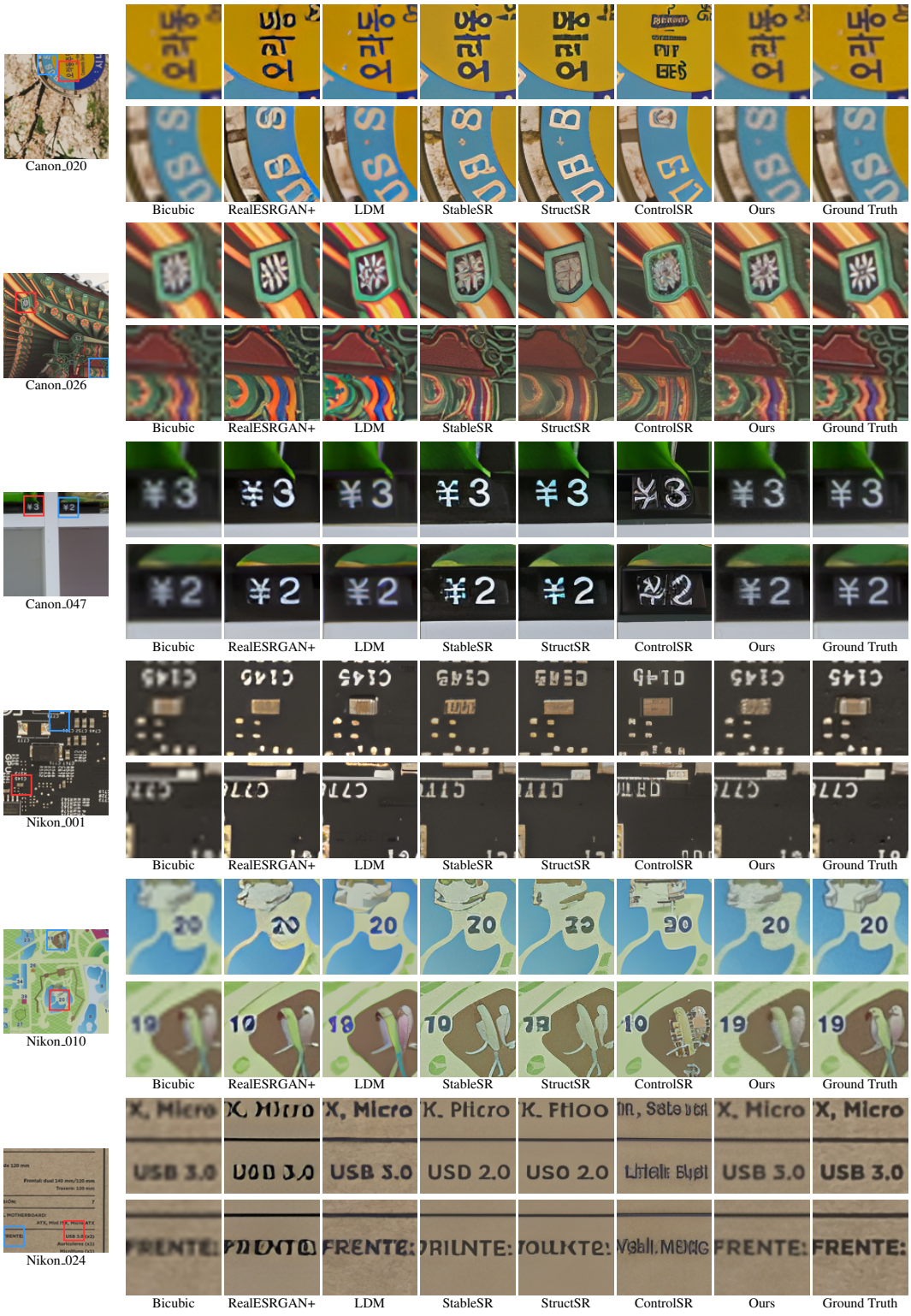

Figure 4: Visual comparison of more super-resolved images, with two sub-regions examined for fidelity per image.