

Hero-SR: One-Step Diffusion for Super-Resolution with Human Perception Priors

Jiangang Wang^{1,2}, Qingnan Fan^{2†}, Qi Zhang²,

Haigen Liu^{1,2}, Yuhang Yu², Jinwei Chen², Wenqi Ren^{1†}

¹School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University

²vivo Mobile Communication Co. Ltd

{wangjg33, liuhg6}@mail2.sysu.edu.cn, {fqnchina, npwpuqzhang}@gmail.com
yuyuhang@vivo.com, chenjinwei_1987@126.com, renwq3@mail.sysu.edu.cn

Project Page: <https://github.com/W-JG/Hero-SR>



Figure 1. Performance and Visual Comparison. (1) Performance Comparison: Compared to one-step and multi-step methods, Hero-SR achieves superior performance with just a single diffusion step. Tested on the DRealSR benchmark, all metrics are normalized using min-max scaling, with ‘S’ denoting the number of diffusion steps. (2) Visual Comparison: Hero-SR restores more realistic textures and aligns better with human perception, outperforming both one-step and multi-step methods. **Zoom in for details.**

Abstract

Owing to the robust priors of diffusion models, recent approaches have shown promise in addressing real-world super-resolution (Real-SR). However, achieving semantic consistency and perceptual naturalness to meet human perception demands remains difficult, especially under conditions of heavy degradation and varied input complexities. To tackle this, we propose Hero-SR, a one-step diffusion-based SR framework explicitly designed with human perception priors. Hero-SR consists of two novel modules: the Dynamic Time-Step Module (DTSM), which adaptively selects optimal diffusion steps for flexibly meeting human perceptual standards, and the Open-World Multi-modality Supervision (OWMS), which integrates guidance from both image and text domains through CLIP to improve semantic consistency and perceptual naturalness. Through these

modules, Hero-SR generates high-resolution images that not only preserve intricate details but also reflect human perceptual preferences. Extensive experiments validate that Hero-SR achieves state-of-the-art performance in Real-SR. The code will be publicly available upon paper acceptance.

1. Introduction

Image super-resolution (SR) reconstructs high-resolution (HR) images from low-resolution (LR) inputs and is critical in fields such as computational photography, video surveillance, and media entertainment, where perceptually accurate visuals are essential [5, 8, 20]. In these applications, perceptual quality directly affects user interpretation and interaction with the content, impacting usability and user experience. However, achieving SR that aligns with high perceptual quality remains a challenge, particularly in real-world SR (Real-SR) tasks with complex degradations like

† Corresponding author.

This work was completed during an internship at vivo.

noise and compression [39].

Traditional pixel-based methods minimize pixel-level distortions but often result in overly smooth images [10, 11, 14]. GAN-based approaches enhance realism but introduce unnatural artifacts [23, 38, 39, 42]. Recently, diffusion-based [16] SR methods have gained attention for their strong priors. Approaches such as StableSR [37], Diff-BIR [24], and SeeSR [46] use pre-trained diffusion models along with guidance mechanisms such as ControlNet [53] to improve SR quality. As diffusion models often require hundreds of iterative steps, methods such as ADDSR [47], OSEDiff [45], and S3Diff [51] apply distillation techniques to reduce the computational cost by using the LR image as a starting point [29] and specialized losses to minimize the number of steps. Despite these improvements, these methods struggle to meet human perception demands for better photo-realistic image super-resolution effects.

In this paper, we interpret the concept of human perception [54] in SR from two core factors : semantic consistency and perceptual naturalness. Semantic consistency ensures that generated images maintain meaningful content. Methods like SeeSR [46], PASD [48], and SUPIR [49] apply various forms of semantic guidance, such as tags, high-level semantic cues, and multimodal textual descriptions. However, these methods often lack the explicit semantic supervision essential for diffusion models to align effectively with semantic consistency. Perceptual naturalness, on the other hand, requires that generated images not only follow general distribution but also align with human perceptual standards. Studies in image quality assessment, such as CLIP-IQA [36] and Q-Align [44], have shown that simply approximating statistical distributions is insufficient; Human-centered evaluations are crucial to align image quality with perceptual standards. However, current SR methods often overlook semantic consistency and perceptual naturalness, leading to images that fall short of human perception standards for coherence and realism.

To address these issues, we propose **Hero-SR**, a one-step diffusion-based super-resolution framework with **Human-perception priors**, specifically designed to improve semantic consistency and perceptual naturalness. Hero-SR consists of two novel modules: the Dynamic Time-Step Module (DTSM) and Open-World Multi-modality Supervision (OWMS). First, the DTSM dynamically selects the optimal time-step based on image-specific features, precisely restoring intricate details. Unlike previous methods [45, 47, 51] that use a fixed starting point from pure noise, DTSM adaptively chooses a starting step from a flexible range by analyzing image degradation and structural complexity. Leveraging a feature extraction network and the Gumbel-Softmax method, DTSM aligns the denoising process with visual details, flexibly meeting human perceptual standards. Second, OWMS improves semantic consistency and perceptual

naturalness by integrating CLIP multimodal guidance [31], aligning SR outputs with both text and image information. In the text domain, perceptual attribute prompts (e.g., quality, sharpness, clarity) guide the model toward criteria that reflect human preferences. In the image domain, the image encoder of CLIP [31] extracts contextual features, enforcing semantic consistency across generated outputs.

Hero-SR integrates DTSM and OWMS to apply human perception priors throughout the SR process, addressing key aspects such as semantic consistency and perceptual naturalness. As shown in Figure 1, extensive experiments demonstrate the effectiveness and flexibility of Hero-SR. The contributions of our work can be summarized as follows:

- We introduce Hero-SR, a one-step diffusion-based super-resolution framework with human perception priors. To the best of our knowledge, we are the first to incorporate multimodal models into the training of Real-SR tasks.
- Hero-SR integrates two novel modules, DTSM and OWMS, to enforce semantic consistency and perceptual naturalness throughout the SR process, ensuring perceptually accurate restorations.
- Hero-SR achieves state-of-the-art performance, outperforming existing one-step and multi-step methods in both quantitative and qualitative evaluations.

2. Related Work

2.1. Real-world Image Super-Resolution

Deep learning has driven advances in SR, beginning with methods like SRCNN [10], which introduced deep neural networks for SR. Subsequent architectures, such as ResNet and Transformer models [11, 14, 22, 42], emphasize fidelity through pixel-level losses. However, these methods often yield overly smooth images that lack the details essential for human perception alignment. Such artifacts can negatively impact the practical usability of SR models, especially in applications like media and surveillance, where fine details are crucial. These limitations are amplified in real-world super-resolution (Real-SR) tasks [4], where images are degraded by noise, compression, and other distortions. The challenge in Real-SR is to restore fine details while ensuring semantic consistency and perceptual naturalness, requiring models that can handle complex degradations and maintain visual fidelity. GAN-based methods [13, 39, 52] incorporate adversarial training to generate finer details. While effective in enhancing realism, GANs frequently introduce unnatural artifacts due to training instability, disrupting semantic consistency [23]. Additionally, GAN-based SR models struggle to preserve coherent global structures, limiting their ability to meet human perception standards [39]. These limitations underscore the need for generative models with stronger priors. Recently,

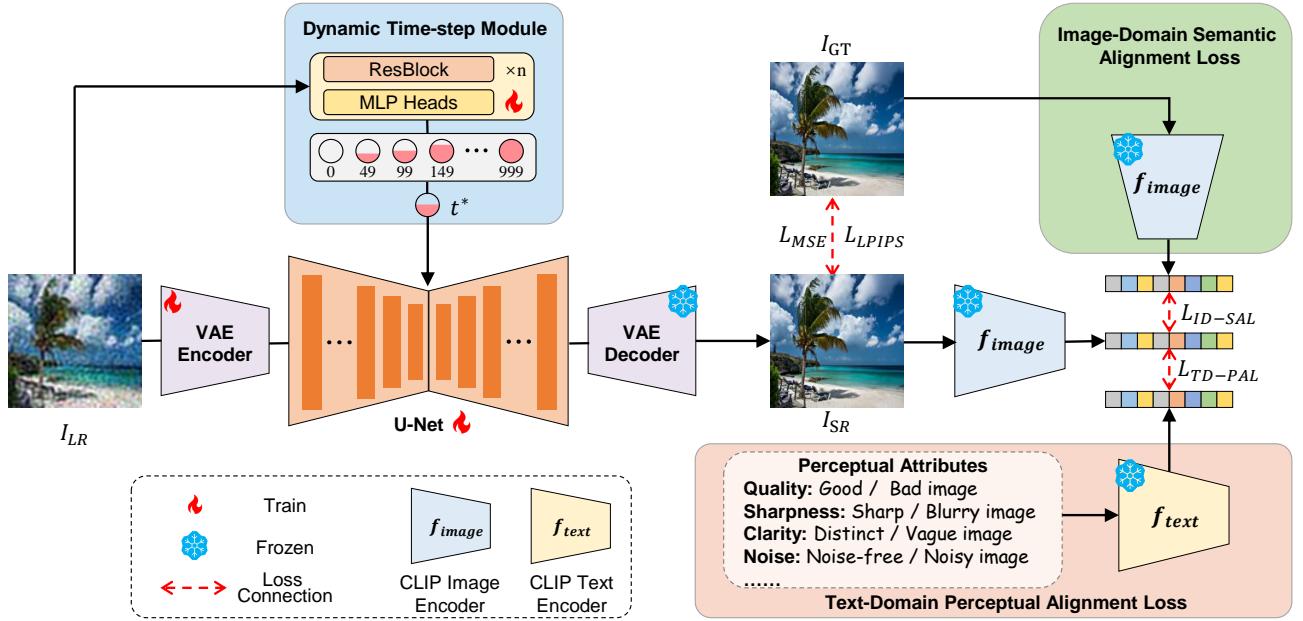


Figure 2. Training framework of Hero-SR. Hero-SR incorporates a Dynamic Time-step Module to adaptively determine the optimal time-step t^* based on the input image I_{LR} , flexibly meeting human perceptual standards. Both I_{LR} and t^* are then input to the diffusion network to generate the restored image I_{SR} . Text-Domain Perceptual Alignment Loss and Image-Domain Semantic Alignment Loss ensure semantic consistency and perceptual naturalness, aligning outputs with human perception.

diffusion models have shown strong potential in generating high-quality images with enhanced detail and coherence.

2.2. Diffusion-based Real-SR

Diffusion models employ Markov processes to generate complex data distributions, with foundational models like DDPM [16] and DDIM [34] establishing the groundwork. The Latent Diffusion Model [32] further improves computational efficiency, enabling large-scale pretrained models such as Stable Diffusion [30]. Extensions like ControlNet [53] provide added control over the generation process, enhancing applications of diffusion models in restoration and editing. In SR tasks, diffusion-based methods generally fall into three categories. The first approach [3, 9, 26, 28] modifies pretrained diffusion models with gradient descent but is constrained by reliance on predefined degradation models, limiting adaptability in real-world scenarios. The second approach, including methods like ResShift [50] and SinSR [40], trains models from scratch on paired data, but results are limited by data diversity and scale. Consequently, the adaptability of these models to complex, real-world degradation patterns remains limited, as they often struggle to adapt to challenging conditions. The third and most common approach leverages pretrained diffusion models with ControlNet [53] to generate high-quality SR outputs from LR inputs. Models like StableSR [37], SeeSR [46], DiffBIR [24], and others [48, 49] improve upon

this approach by incorporating architectural and semantic guidance, yielding visually enhanced outputs. Diffusion-based SR methods typically require numerous sampling steps, reducing practical efficiency. To address this limitation, recent diffusion-based SR methods like ADDSR [47], S3Diff [51], and OSEDiff [45] incorporate adversarial distillation and score-matching to accelerate inference. However, diffusion-based SR methods still fall short of fully meeting human perception standards, especially in semantic consistency, and perceptual naturalness. This highlights the need for SR methods better aligned with human visual expectations.

3. Methodology

3.1. Framework Overview

Hero-SR is a one-step diffusion-based super-resolution framework with human perception priors, with two core modules: the Dynamic Time-Step Module (DTSM) to flexibly meet human perceptual standards and the Open-World Multi-modality Supervision (OWMS) for perceptual and semantic alignment. Hero-SR is built on the Stable Diffusion model [32], comprising a VAE encoder \mathcal{E} , a U-Net \mathcal{U} , and a VAE decoder \mathcal{D} .

As shown in Figure 2, given a low-resolution input I_{LR} , DTSM adaptively selects an optimal time-step $t^* = \text{DTSM}(I_{LR})$. The VAE encoder encodes I_{LR} into a latent

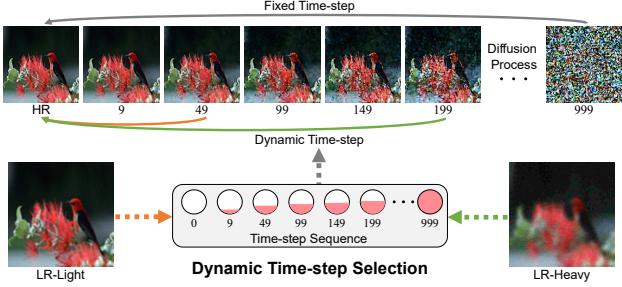


Figure 3. The time-step selection process of DTSM. Previous one-step methods use a fixed starting time-step from pure noise, while DTSM adaptively selects a dynamic starting time-step based on the input image to better align with the diffusion process.

representation $z_{LR} = \mathcal{E}(I_{LR})$, which is then processed by the U-Net at t^* to produce an enhanced latent representation $z_{SR} = \mathcal{U}(z_{LR}, t^*)$. Finally, the VAE decoder reconstructs the high-resolution output $I_{SR} = \mathcal{D}(z_{SR})$. Low-rank adaptation (LoRA) [17] is applied, with the VAE decoder frozen during training to maintain its generative capacity.

3.2. Dynamic Time-Step Module

In Real-SR tasks, the degradation of input images varies widely, leading to a range of structural complexities [4]. As shown in Figure 3, previous one-step diffusion-base SR methods [45, 47, 51] with a fixed starting time-step, such as step 999, fail to account for this variation, limiting the restoration of rich details. The Dynamic Time-Step Module (DTSM) addresses this by adaptively selecting the optimal time-step based on the degradation level and complexity of the input image, thus improving detail restoration to flexibly meet perceptual standards.

Diffusion models operate through progressive denoising [16], with each time-step corresponding to a noise level [27]. To allow DTSM to adapt to different levels of input complexity, we select a time-step candidate subset S from the original diffusion sequence:

$$S \subseteq \{x \in \mathbb{Z} \mid 0 \leq x \leq 999\},$$

where each element in S represents a specific noise level across the diffusion trajectory.

For a given low-resolution input I_{LR} , DTSM extracts features relevant to degradation and complexities to guide time-step selection. First, I_{LR} is processed through convolutional layers to capture localized feature patterns:

$$f_{shallow} = \text{Conv}(I_{LR}). \quad (1)$$

These features are further refined through a series of residual blocks [15] to model more complex characteristics:

$$f_{deep} = \text{ResBlocks}_n(f_{shallow}), \quad (2)$$

Table 1. Perceptual attributes and their corresponding prompts. We select key perceptual attributes closely aligned with human perception and apply respective positive and negative prompts.

Perceptual Attributes	Positive Prompts	Negative Prompts
Quality	Good image	Bad image
Sharpness	Sharp image	Blurry image
Edge Clarity	Sharp edges	Blurry edges
Resolution	High resolution image	Low resolution image
Noise	Noise-free image	Noisy image
Clarity	Distinct image	Vague image

where n is the number of residual blocks. This output f_{deep} is then flattened and passed through a multi-layer perceptron (MLP), yielding a compact feature vector v :

$$v = \text{MLP}(\text{Flatten}(f_{deep})). \quad (3)$$

To select the optimal time-step t^* , we use the Gumbel-Softmax trick [18], which enables differentiable selection during training. The time-step t^* is computed as:

$$t^* = \text{Gumbel-Softmax}(v, S). \quad (4)$$

By aligning the denoising process with both structural complexity and degradation characteristics of the input, DTSM effectively balances detail restoration with perceptual naturalness, flexibly aligning with human perception requirements.

3.3. Open-World Multi-modality Supervision

To address the limitations of traditional loss functions and better align with human perception, we propose an Open-World Multi-modality Supervision strategy (OWMS). This approach leverages the powerful multimodal capabilities of CLIP [31], a model pre-trained on large-scale datasets to establish strong visual-textual associations, achieving an open-world level of perceptual understanding [36]. This shared space enables effective alignment through two main components: the Text-Domain Perceptual Alignment Loss (TD-PAL), which guides perceptual alignment, and the Image-Domain Semantic Alignment Loss (ID-SAL), which enforces semantic consistency.

3.3.1. Text-Domain Perceptual Alignment Loss

Text-Domain Perceptual Alignment Loss (TD-PAL) aligns restored images with human-perceptual standards by focusing on n perceptual attributes, each represented by a positive and negative prompt pair [36], as shown in Table 1. By adjusting these attributes, TD-PAL enhances the perceptual quality of restored images, aligning them more closely with human expectations.

For a restored image I_{SR} , we compute its embedding e_{SR} using image encoder of CLIP f_{image} :

$$e_{SR} = f_{image}(I_{SR}). \quad (5)$$

Subsequently, we encode predefined prompts using the text encoder. Each attribute has a positive prompt T_i^p and a negative prompt T_i^n . Using text encoder of CLIP f_{text} , we obtain their embeddings:

$$\mathbf{e}_{\text{text}}^{(i,p)} = f_{\text{text}}(T_i^p), \quad \mathbf{e}_{\text{text}}^{(i,n)} = f_{\text{text}}(T_i^n), \quad (6)$$

where $\mathbf{e}_{\text{text}}^{(i,p)}$ and $\mathbf{e}_{\text{text}}^{(i,n)}$ denote the embeddings of the positive and negative prompts for the i -th attribute, respectively.

To assess alignment in the perceptual attributes, we compute the cosine similarity between the image embedding \mathbf{e}_{SR} and each text embedding, $\mathbf{e}_{\text{text}}^{(i,p)}$ and $\mathbf{e}_{\text{text}}^{(i,n)}$:

$$s_i^{(p)} = \frac{\mathbf{e}_{\text{SR}} \odot \mathbf{e}_{\text{text}}^{(i,p)}}{\|\mathbf{e}_{\text{SR}}\| \cdot \|\mathbf{e}_{\text{text}}^{(i,p)}\|}, \quad s_i^{(n)} = \frac{\mathbf{e}_{\text{SR}} \odot \mathbf{e}_{\text{text}}^{(i,n)}}{\|\mathbf{e}_{\text{SR}}\| \cdot \|\mathbf{e}_{\text{text}}^{(i,n)}\|}, \quad (7)$$

where $s_i^{(p)}$ and $s_i^{(n)}$ represent the cosine similarities between the image embedding \mathbf{e}_{SR} and each positive and negative prompt for the i -th attribute, respectively.

We apply softmax normalization for stability:

$$\hat{s}_i^{(p)} = \frac{e^{s_i^{(p)}}}{e^{s_i^{(p)}} + e^{s_i^{(n)}}}, \quad (8)$$

where $\hat{s}_i^{(p)}$ represents the normalized similarity score for the positive prompt of the i -th attribute, reflecting the alignment of I_{SR} with perceptual attributes and enabling stable comparisons between positive and negative prompts for consistent alignment across attributes.

TD-PAL is then defined as:

$$\mathcal{L}_{\text{TD-PAL}} = 1 - \frac{1}{n} \sum_{i=1}^n \hat{s}_i^{(p)}, \quad (9)$$

encouraging the alignment of I_{SR} with human-perceptual standards across each attribute. This alignment enhances the perceptual quality of the generated images, making them more attuned to human quality assessments.

3.3.2. Image-Domain Semantic Alignment Loss

Diffusion models differ from traditional SR approaches by relying on semantic information to guide image generation. However, existing methods [46, 48, 49] focus on semantic guidance, neglecting the importance of semantic supervision. To address this gap, we propose the Image-Domain Semantic Alignment Loss (ID-SAL) to improve the generative ability of the model through semantic-level alignment.

ID-SAL enforces semantic consistency by aligning restored images with ground truth (GT) images. For a restored image I_{SR} and its GT image I_{GT} , we use the CLIP image encoder to compute their embeddings in the semantic space. Since \mathbf{e}_{SR} , the embedding for I_{SR} , has already been computed in Equation (5), we compute only the embedding for I_{GT} , denoted as \mathbf{e}_{GT} , as follows:

$$\mathbf{e}_{\text{GT}} = f_{\text{image}}(I_{\text{GT}}). \quad (10)$$

Next, we calculate the cosine similarity between \mathbf{e}_{SR} and \mathbf{e}_{GT} to assess semantic alignment:

$$s = \frac{\mathbf{e}_{\text{SR}} \odot \mathbf{e}_{\text{GT}}}{\|\mathbf{e}_{\text{SR}}\| \cdot \|\mathbf{e}_{\text{GT}}\|}, \quad (11)$$

where $s \in [-1, 1]$ denotes the semantic alignment score, with values closer to 1 indicating higher alignment in semantic space. This score quantifies how well the restored image I_{SR} preserves the semantic content of its ground-truth counterpart I_{GT} .

ID-SAL is then defined as:

$$\mathcal{L}_{\text{ID-SAL}} = 1 - s, \quad (12)$$

which drives the restored image to maintain semantic fidelity with its GT counterpart. This alignment improves the ability of the model to produce semantically consistent outputs, enhancing both perceptual coherence and fidelity for diverse real-world SR inputs.

3.4. Total Loss Function

The total loss combines multiple objectives to balance fidelity, perceptual alignment, and semantic consistency:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{LPIPS}} + \lambda_3 \mathcal{L}_{\text{TD-PAL}} + \lambda_4 \mathcal{L}_{\text{ID-SAL}}, \quad (13)$$

where, λ_i corresponds to \mathcal{L}_i for $i = 1, 2, 3, 4$, representing \mathcal{L}_{MSE} , $\mathcal{L}_{\text{LPIPS}}$, $\mathcal{L}_{\text{TD-PAL}}$, and $\mathcal{L}_{\text{ID-SAL}}$ respectively. This combination ensures that Hero-SR meets human perception criteria, achieving high-quality detail restoration, semantic consistency, and perceptual naturalness in SR tasks.

4. Experiments

4.1. Experiments Setting

Training and Testing Datasets. We train the model on the LSDIR [21] dataset, using the Real-ESRGAN [39] degradation pipeline to generate LR-HR training pairs. Testing is conducted on the StableSR [37] test set, including synthetic and real data. The synthetic dataset consists of 3,000 images at 512×512 resolution, with GT images randomly cropped from DIV2K-val [2] and degraded using Real-ESRGAN. Real data is sourced from RealSR [4] and DRealSR [43], containing 128×128 and 512×512 LR-HR pairs. This combination of synthetic and real-world test sets assesses the model on both controlled and unpredictable degradations, ensuring its robustness and generalization.

Compared Methods. We compare our model with recent advanced diffusion model super-resolution methods, categorized into one-step (e.g., ADDSR [47], S3Diff [51], OSEDiff [45], SinSR [40]) and multi-step approaches (e.g., StableSR [37], DiffBIR [24], SeeSR [46], ResShift [50]). ResShift and its distilled one-step variant, SinSR, are

Table 2. Quantitative comparison with **one-step** diffusion methods on both synthetic and real-world benchmarks. The best and second best results of each metric are highlighted in **red** and **blue**, respectively.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	HyperIQA↑	TOPIQ↑	TRES↑	ARNIQA↑	Q-Align↑
DIV2K	ADDSSR	23.2604	0.5902	0.3623	63.3961	0.5829	0.5730	73.3587	0.7107	3.2480
	S3Diff	23.5164	0.5949	0.2581	68.0107	0.6376	0.6342	80.7641	0.7209	3.7949
	OSEDiff	23.3820	0.6009	0.3173	69.1473	0.6421	0.6459	82.4990	0.7209	4.0781
	SinSR	24.4111	0.6017	0.3239	62.7990	0.5797	0.5721	72.5458	0.6659	3.1895
	Hero-SR (Ours)	24.3663	0.6257	0.3111	69.8524	0.6711	0.6948	87.3938	0.7255	3.9968
DrealSR	ADDSSR	27.7707	0.7722	0.3196	60.8542	0.5797	0.5688	71.7246	0.6654	3.2578
	S3Diff	27.3852	0.7468	0.3130	64.1622	0.6053	0.6053	75.6122	0.6784	3.6094
	OSEDiff	27.6269	0.7740	0.3159	66.3766	0.6287	0.6220	79.4500	0.6833	3.6855
	SinSR	28.3578	0.7518	0.3659	55.6310	0.5182	0.5193	61.4332	0.5985	3.1191
	Hero-SR (Ours)	28.8962	0.8016	0.2933	66.4874	0.6434	0.6622	83.5888	0.6913	3.6302
RealSR	ADDSSR	24.7929	0.7077	0.3091	66.1849	0.6082	0.5991	79.9438	0.6923	3.4102
	S3Diff	25.1930	0.7315	0.2707	67.9144	0.6104	0.6137	78.7253	0.6969	3.6523
	OSEDiff	24.8520	0.7218	0.3115	69.9864	0.6469	0.6506	83.5311	0.7013	3.8047
	SinSR	26.3254	0.7364	0.3195	60.5987	0.5205	0.5184	67.8383	0.6435	3.1816
	Hero-SR (Ours)	25.8271	0.7439	0.2893	70.0254	0.6623	0.6881	88.5315	0.7170	3.8470

Table 3. Quantitative comparison with **multi-step** diffusion methods on both synthetic and real-world benchmarks. The best and second best results of each metric are highlighted in **red** and **blue**, respectively.

DataSet	Methods	Step	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	HyperIQA↑	TOPIQ↑	TRES↑	ARNIQA↑	Q-Align↑
DIV2K	StableSR	200	23.2613	0.5726	0.3113	65.9177	0.6130	0.5979	77.3719	0.6916	3.5273
	DiffBIR	50	23.4091	0.5732	0.3456	68.3954	0.6315	0.6344	80.9948	0.7002	3.7324
	SeeSR	50	23.6780	0.6043	0.3193	68.6721	0.6679	0.6857	85.8015	0.7302	4.1211
	ResShift	15	24.7538	0.6300	0.3649	60.0644	0.5564	0.5253	73.3462	0.6645	2.8613
	Hero-SR (Ours)	1	24.3663	0.6257	0.3111	69.8524	0.6711	0.6948	87.3938	0.7255	3.9968
DrealSR	StableSR	200	28.0297	0.7536	0.3284	58.5118	0.5482	0.5323	66.7321	0.6254	3.0613
	DiffBIR	50	25.9304	0.6526	0.4518	65.6740	0.6296	0.6149	77.8703	0.6546	3.5977
	SeeSR	50	28.0719	0.7684	0.3174	65.0907	0.6642	0.6574	84.7264	0.6883	3.5879
	ResShift	15	28.8071	0.8065	0.3207	53.3830	0.5031	0.4757	64.2898	0.6037	2.8145
	Hero-SR (Ours)	1	28.8962	0.8016	0.2933	66.4874	0.6434	0.6622	83.5888	0.6913	3.6302
RealSR	StableSR	200	24.6451	0.7080	0.3002	65.8833	0.5796	0.5748	74.2591	0.6756	3.2764
	DiffBIR	50	24.2406	0.6649	0.3469	68.3388	0.6121	0.6052	78.9864	0.6717	3.6328
	SeeSR	50	25.1477	0.7210	0.3007	69.8191	0.6748	0.6891	88.5903	0.7155	3.7148
	ResShift	15	26.5344	0.7636	0.2964	60.0152	0.5393	0.5215	73.7639	0.6647	3.0469
	Hero-SR (Ours)	1	25.8271	0.7439	0.2893	70.0254	0.6623	0.6881	88.5315	0.7170	3.8470

trained from scratch, while other methods rely on pre-trained SD models. GAN-based methods such as SwinIR [22], BSRGAN [52], FeMaSR [6] and RealESRGAN [39] are presented in the Appendix for comparison.

Evaluation Metrics. To comprehensively and accurately evaluate the performance of various methods, we employ a series of full-reference and no-reference metrics. PSNR and SSIM [41], calculated on the Y channel in YCbCr space, serve as full-reference fidelity metrics, while LPIPS [54] is utilized as a full-reference perceptual quality metric. For no-reference image quality assessment, we employ advanced metrics such as MUSIQ [19], HyperIQA [35], TOPIQ [7], TRES [12], ARNIQA [1], and Q-Align [44]. These no-reference IQA methods are SOTA metrics, closely

aligned with human subjective evaluations and perception. In particular, Q-Align, based on the LMM model, demonstrates exceptional evaluation capabilities.

Implementation Details. Model training is conducted with the AdamW [25] optimizer at a learning rate of 5×10^{-5} . Training is performed on 2 NVIDIA L40s GPUs for approximately 8 hours with a batch size of 2. SD-Turbo [33] is used as a pre-trained diffusion model. The VAE encoder and U-Net network are fine-tuned using LoRA [17] with a rank level of 16. The adaptive time-step module is trained from scratch with randomly initialized parameters. The weights of the losses λ_1 , λ_2 , λ_3 , and λ_4 are set to 2, 5, 1, and 0.5, respectively. In TD-PAL and ID-SAL, the

parameters of CLIP are frozen.

4.2. Comparison with State-of-the-Arts

4.2.1. Quantitative Comparisons.

One-Step Methods. Table 2 presents the quantitative comparison between Hero-SR and other one-step methods. Key observations include: (1) Hero-SR consistently outperforms other methods across nearly all metrics, particularly on real-world datasets like DRealSR and RealSR. (2) Hero-SR achieves leading results in full-reference metrics, surpassing other methods in PSNR, SSIM, and LPIPS. SinSR attains a higher PSNR, likely due to its scratch-trained diffusion model, but underperforms on no-reference perceptual metrics. S3Diff shows a better LPIPS score but worse results on other no-reference metrics, likely due to its heavier LPIPS loss weighting during training. (3) Hero-SR outperforms other methods across all datasets for no-reference perceptual metrics (e.g., MUSIQ, HyperIQA, TOPIQ, TRES, ARNIQA). For example, Hero-SR exceeds competitors by 7.0% on the TOPIQ metric. These advanced no-reference metrics emphasize overall perceptual quality and align closely with human perception standards, highlighting the ability of Hero-SR to generate high-quality reconstructions that meet human visual expectations.

Multi-Step Methods. Table 3 provides the quantitative comparison between Hero-SR and multi-step methods, with key findings as follows: (1) As a one-step diffusion model, Hero-SR achieves competitive results with multi-step approaches across multiple datasets. (2) ResShift, which does not use a pre-trained diffusion model, shows relatively better performance on full-reference metrics like PSNR and SSIM but lower scores on no-reference metrics. However, compared to pretrained diffusion-based methods, Hero-SR achieves superior results on almost all full-reference fidelity metrics. (3) Hero-SR consistently ranks first or second across nearly all datasets in no-reference perceptual metrics, highlighting its strong alignment with human perceptual standards. These results demonstrate the ability of Hero-SR to capture visual qualities aligned with human judgment, such as naturalness and semantic consistency.

4.2.2. Qualitative Comparisons.

Figures 1 and 4 present visual comparison results. (1) In terms of texture restoration in the fox and owl case, Hero-SR generates more realistic details compared to other approaches. Compared to single-step methods, Hero-SR demonstrates clear advantages by producing more natural and perceptually aligned results. Compared to multi-step methods, Hero-SR produces more realistic texture details. (2) In terms of semantic consistency in the leaf case, Hero-SR demonstrates superior semantic consistency by generating a coherent and complete leaf structure. Notably, Hero-SR not only preserves intricate details but also avoids in-

Table 4. Ablation study on the impact of different perceptual attributions.

Perceptual Attributes	MUSIQ↑	HyperIQA↑	TOPIQ ↑	Q-Align↑
w/o Quality	65.4624	0.6316	0.6319	3.5633
w/o Sharpness	66.4843	0.6203	0.6589	3.5664
w/o Edge Clarity	65.6499	0.6360	0.6460	3.5534
w/o Resolution	66.3280	0.6315	0.6533	3.5936
w/o Noise	65.2068	0.6417	0.6524	3.5892
w/o Clarity	66.3015	0.6438	0.6579	3.6109
All	66.4874	0.6434	0.6622	3.6302

troducing unnatural artifacts, thereby achieving a balance between local texture fidelity and global structural coherence. These results highlight the capabilities of Hero-SR across various scenarios, ranging from texture restoration to complex semantic alignment. **Additional visual results are provided in the appendix.**

4.3. Ablation Study

We first evaluate the effectiveness of the proposed DTSM and OWMS, with OWMS comprising the two components ID-SAL and TD-PAL, by testing Hero-SR with each module removed. Next, we analyze the impact of perceptual attributes within TD-PAL. Unless otherwise noted, all experiments are conducted on the DRealSR dataset with retrained models, while holding all other settings constant.

The Effectiveness of DTSM. As shown in Table 5, removing DTSM (Variant-1 vs. Hero-SR) results in a noticeable decline in no-reference perceptual metrics, including HyperIQA, TOPIQ, and Q-Align. These results underscore the critical function of DTSM in dynamically adjusting timestep to optimize perceptual quality. By adapting to image-specific features, DTSM effectively balances detail restoration with perceptual naturalness, aligning with human perceptual expectations.

The Effectiveness of ID-SAL. As shown in Table 5, the removal of ID-SAL (Variant-2 vs. Hero-SR) causes decreases in perceptual alignment metrics, highlighting the role of ID-SAL in maintaining semantic consistency. The reduction in Q-Align, a key metric for alignment with perceptual standards, emphasizes the contribution of ID-SAL to content coherence, ensuring generated images closely align with human perceptual expectations.

The Effectiveness of TD-PAL. As shown in Table 5, without TD-PAL (Variant-3 vs. Hero-SR), we observe noticeable declines in no-reference perceptual metrics, such as MUSIQ, TOPIQ, and TRES. These results suggest that TD-PAL is essential for enhancing perceptual naturalness, guiding the model to produce outputs that align well with perceptual standards.

The Impact of Different Perceptual Attributions. The ablation study in Table 4 demonstrates the impact of individual perceptual attributes on the performance of Hero-SR,

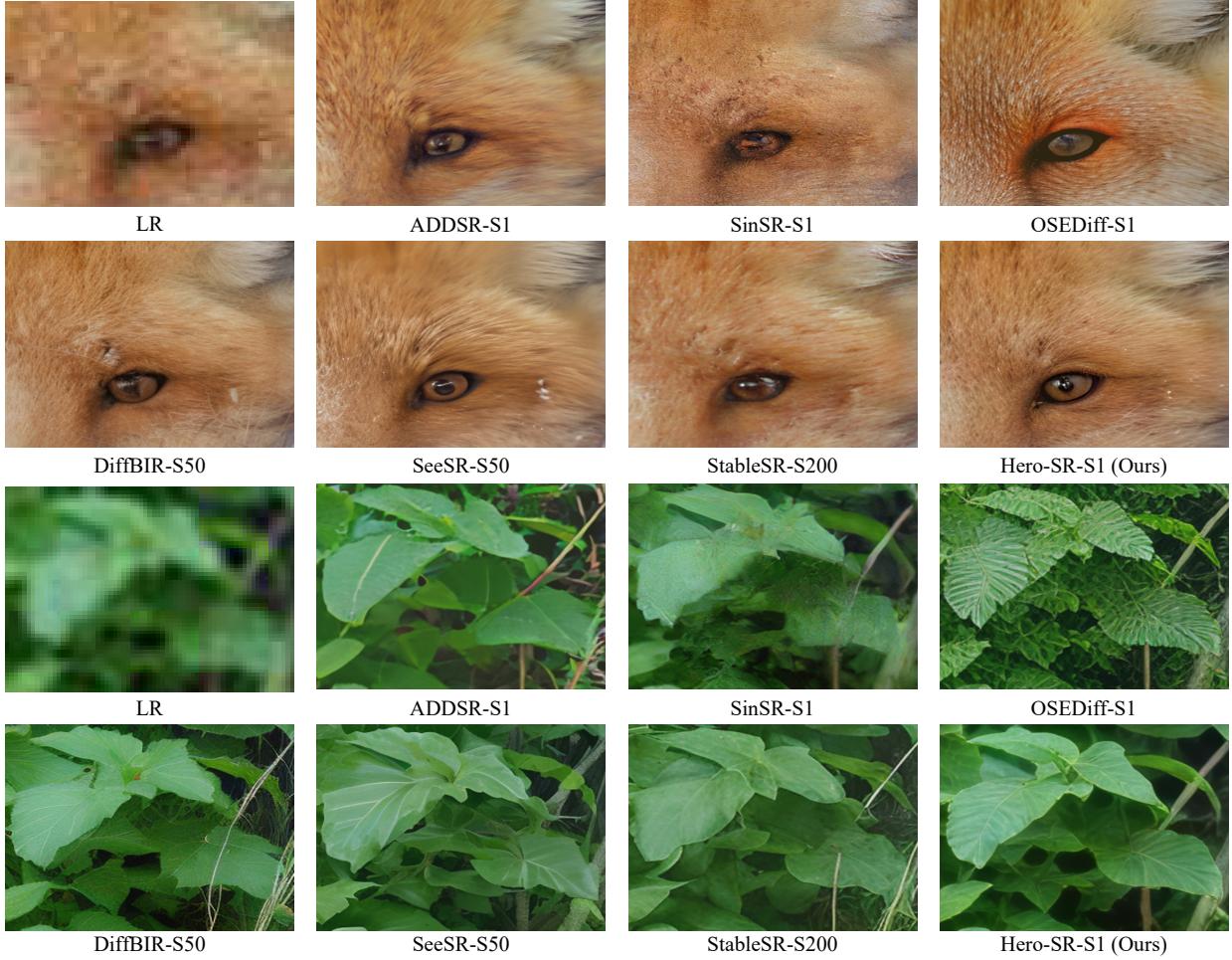


Figure 4. Qualitative comparison with one-step and multi-step methods. ‘S’ indicates the number of diffusion steps. **Zoom in for details.**

Table 5. Ablation study results on the effectiveness of the proposed DTSM, ID-SAL, and TD-PAL.

Methods	DTSM	ID-SAL	TD-PAL	MUSIQ↑	HyperIQA↑	TOPIQ↑	TRES↑	ARNIQA↑	Q-Align↑
Variant-1	✗	✓	✓	66.4019	0.6361	0.6528	82.4065	0.6875	3.5282
Variant-2	✓	✗	✓	66.0649	0.6385	0.6617	83.4267	0.6609	3.5877
Variant-3	✓	✓	✗	60.7671	0.5849	0.5629	73.7752	0.6302	3.1945
Hero-SR	✓	✓	✓	66.4874	0.6434	0.6622	83.5888	0.6913	3.6302

with some attributes contributing more than others. Excluding attributes like Quality and Noise led to marked declines in perceptual metrics; for example, removing Noise reduced HyperIQA, while omitting Quality notably lowered MUSIQ, underscoring the role of these attributes in achieving perceptual fidelity and naturalness. Including all perceptual attributes yields optimal performance, confirming that the combined use of all attributes is essential for aligning SR outputs with human perceptual standards and achieving high-quality, realistic results.

5. Conclusion and Limitation

We propose Hero-SR, a one-step diffusion-based super-resolution framework specifically designed with human perception priors to enhance semantic consistency and perceptual naturalness in real-world SR tasks. Hero-SR integrates two core modules: the Dynamic Time-Step Module (DTSM), which flexibly selects optimal diffusion steps to balance fidelity with perceptual standards, and the Open-World Multi-modality Supervision (OWMS), which leverages multimodal guidance from CLIP across image and text

domains to reinforce semantic alignment with human visual preferences. Through these modules, Hero-SR effectively captures fine details and produces high-resolution images closely aligned with human perceptual expectations. Extensive experiments demonstrate that Hero-SR achieves state-of-the-art performance across both real and synthetic datasets, surpassing existing one-step and multi-step methods in quantitative metrics and qualitative evaluation.

Hero-SR has certain limitations. Like other SD-based methods, it is constrained by the reconstruction capacity of the VAE, which restricts its ability to restore small structures, such as small-scale text and face. We aim to address these challenges in future work.

References

- [1] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. ARNIQA: learning distortion manifold for image quality assessment. In *IEEE/CVF Winter Conference on Applications of Computer Vision, CVPR 2024*, pages 188–197, 2024. [6](#)
- [2] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2017*, pages 1122–1131, 2017. [5](#), [11](#)
- [3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 18187–18197, 2022. [3](#)
- [4] 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 *IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 3086–3095, 2019. [2](#), [4](#), [5](#), [11](#)
- [5] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 1652–1660, 2019. [1](#)
- [6] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *30th ACM International Conference on Multimedia*, pages 1329–1338, 2022. [6](#), [11](#)
- [7] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Trans. Image Process.*, 33:2404–2418, 2024. [6](#)
- [8] Benjamin Naoto Chiche, Arnaud Woiselle, Joana Frontera-Pons, and Jean-Luc Starck. Stable long-term recurrent video super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 827–836, 2022. [1](#)
- [9] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contrac-
tion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 12403–12412, 2022. [3](#)
- [10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of European Conference on Computer Vision, Part IV*, pages 184–199, 2014. [2](#)
- [11] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Proceedings of European Conference on Computer Vision, Part II*, pages 391–407, 2016. [2](#)
- [12] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *IEEE/CVF Winter Conference on Applications of Computer Vision, CVPR 2022*, pages 3989–3999, 2022. [6](#)
- [13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. [2](#)
- [14] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 1604–1613, 2019. [2](#)
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 770–778, 2016. [4](#)
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. [2](#), [3](#), [4](#)
- [17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Tenth International Conference on Learning Representations, ICLR 2022*, 2022. [4](#), [6](#)
- [18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017*, 2017. [4](#)
- [19] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. In *IEEE/CVF International Conference on Computer Vision, CVPR 2021*, pages 5128–5137, 2021. [6](#), [11](#)
- [20] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 105–114, 2017. [1](#)
- [21] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. LSDIR: A large scale dataset for image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2023*, pages 1775–1787, 2023. [5](#)

- [22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021*, pages 1833–1844, 2021. 2, 6, 11
- [23] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 5647–5656, 2022. 2
- [24] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 2, 3, 5
- [25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. 6
- [26] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 11451–11461, 2022. 3
- [27] Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Image restoration with mean-reverting stochastic differential equations. In *International Conference on Machine Learning, ICLR 2023*, pages 23045–23066, 2023. 4
- [28] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jianjun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *Tenth International Conference on Learning Representations, ICLR 2022*, 2022. 3
- [29] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *ArXiv preprint, abs/2403.12036*, 2024. 2
- [30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. 3
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *38th International Conference on Machine Learning, ICLR 2021*, pages 8748–8763, 2021. 2, 4
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 10674–10685, 2022. 3
- [33] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Proceedings of European Conference on Computer Vision*, pages 87–103, 2024. 6
- [34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. 3
- [35] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 3664–3673, 2020. 6
- [36] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 2, 4
- [37] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *Int. J. Comput. Vis.*, 132(12):5929–5949, 2024. 2, 3, 5
- [38] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *Proceedings of European Conference on Computer Vision Workshops, ECCVW 2018*, pages 63–79, 2018. 2
- [39] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021*, pages 1905–1914, 2021. 2, 5, 6, 11
- [40] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 25796–25805, 2024. 3, 5
- [41] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 6
- [42] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3365–3387, 2021. 2
- [43] 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 European Conference on Computer Vision*, pages 101–117, 2020. 5, 11
- [44] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *Forty-first International Conference on Machine Learning, ICLR 2024*, 2024. 2, 6, 11
- [45] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 2024. 2, 3, 4, 5
- [46] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware

- real-world image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 25456–25467, 2024. [2](#), [3](#), [5](#)
- [47] Rui Xie, Ying Tai, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, Xiaoqian Ye, Qian Wang, and Jian Yang. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *ArXiv preprint*, abs/2404.01717, 2024. [2](#), [3](#), [4](#), [5](#)
- [48] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *Proceedings of European Conference on Computer Vision*, pages 74–91, 2024. [2](#), [3](#), [5](#)
- [49] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 25669–25680, 2024. [2](#), [3](#), [5](#)
- [50] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems*, 2023. [3](#), [5](#)
- [51] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *ArXiv preprint*, abs/2409.17058, 2024. [2](#), [3](#), [4](#), [5](#)
- [52] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE/CVF International Conference on Computer Vision, ICCV 2021*, pages 4771–4780, 2021. [2](#), [6](#), [11](#)
- [53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 3813–3824, 2023. [2](#), [3](#)
- [54] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 586–595, 2018. [2](#), [6](#)

Appendix

A. Comparison with GAN-based Methods

We compare Hero-SR with four representative GAN-based Real-SR methods: BSRGAN [52], Real-ESRGAN [39], SwinIR [22], and FeMaSR [6], using three synthetic and real-world datasets [2, 4, 43]. Quantitative and qualitative comparisons demonstrate that Hero-SR achieves superior perceptual consistency and generates more realistic textures, particularly in complex real-world scenarios.

Quantitative Comparisons. As shown in Table 6, two key observations can be made: (1) GAN-based methods achieve higher fidelity metrics: GANs perform better on PSNR and SSIM. However, GAN-based methods are limited by their generative capacity and often fail to maintain high perceptual quality, falling short of aligning with human perception standards. (2) Hero-SR significantly outperforms GAN methods in perceptual quality: On no-reference perceptual metrics, such as MUSIQ [19] and Q-Align [44], Hero-SR demonstrates a substantial advantage over all GAN-based methods. This improvement is attributed to the strong generative priors of diffusion models and the human perception design of Hero-SR, enabling exceptional perceptual alignment and naturalness.

Qualitative Comparisons. Figure 5 highlights the superiority of Hero-SR over GAN-based methods in texture restoration and semantic consistency. For instance, in the example of the blind, Hero-SR accurately restores high-frequency details and produces structured, natural textures. By contrast, while capturing some details, GAN methods fail to restore complex textures convincingly. In the leaf example, Hero-SR reconstructs a complete leaf structure with clearly defined vein patterns, achieving higher semantic consistency compared to GAN-based methods.

B. Additional Visual Comparisons

Figures 6, 7, 8, and 9 present additional visual comparisons between Hero-SR and other diffusion-based methods. Hero-SR consistently outperforms one-step methods across various scenarios, including architectural structures, animal fur, and text. It also achieves results comparable to or exceeding those of multi-step methods, demonstrating its capability to produce high-quality outputs efficiently. Notably, Hero-SR excels in balancing fine detail restoration and semantic consistency, making its outputs more aligned with human perception across diverse and challenging contexts.

Table 6. Quantitative comparison with **GAN-base** methods on both synthetic and real-world benchmarks. The best and second best results of each metric are highlighted in **red** and **blue**, respectively.

DataSet	Methods	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	HyperIQA↑	TOPIQ↑	TRES↑	ARNIQA↑	Q-Align↑
DIV2K	BSRGAN	24.5831	0.6269	0.3351	61.1928	0.5719	0.5460	74.0277	0.6605	2.8535
	RealESRGAN	24.2927	0.6372	0.3112	61.0570	0.5665	0.5297	70.1277	0.6734	3.0684
	FeMaSR	23.0587	0.5887	0.3126	60.8277	0.5591	0.5231	70.7251	0.6645	2.8828
	SwinIR	23.9314	0.6285	0.3160	60.2177	0.5504	0.5100	68.6045	0.6616	2.9727
	Hero-SR (Ours)	24.3663	0.6257	0.3111	69.8524	0.6711	0.6948	87.3938	0.7255	3.9968
DrealSR	BSRGAN	28.7021	0.8028	0.2858	57.1596	0.5304	0.5060	66.7613	0.6262	2.9551
	RealESRGAN	26.8655	0.7569	0.3157	53.7035	0.4877	0.4673	59.3529	0.6181	2.8711
	FeMaSR	28.6147	0.8051	0.2819	54.2777	0.4938	0.4623	58.7931	0.6101	2.8633
	SwinIR	28.4969	0.8044	0.2743	52.7369	0.4800	0.4424	58.0347	0.5948	2.8125
	Hero-SR (Ours)	28.8962	0.8016	0.2933	66.4874	0.6434	0.6622	83.5888	0.6913	3.6302
RealSR	BSRGAN	26.3793	0.7651	0.2656	63.2838	0.5617	0.5505	75.7009	0.6830	3.1797
	RealESRGAN	25.0632	0.7356	0.2937	59.0565	0.5215	0.5029	67.2148	0.6674	3.0117
	FeMaSR	25.6854	0.7614	0.2709	60.3697	0.5231	0.5148	67.6841	0.6751	3.1055
	SwinIR	26.3081	0.7729	0.2539	58.6948	0.4973	0.4787	64.7595	0.6609	2.9434
	Hero-SR (Ours)	25.8271	0.7439	0.2893	70.0254	0.6623	0.6881	88.5315	0.7170	3.8470

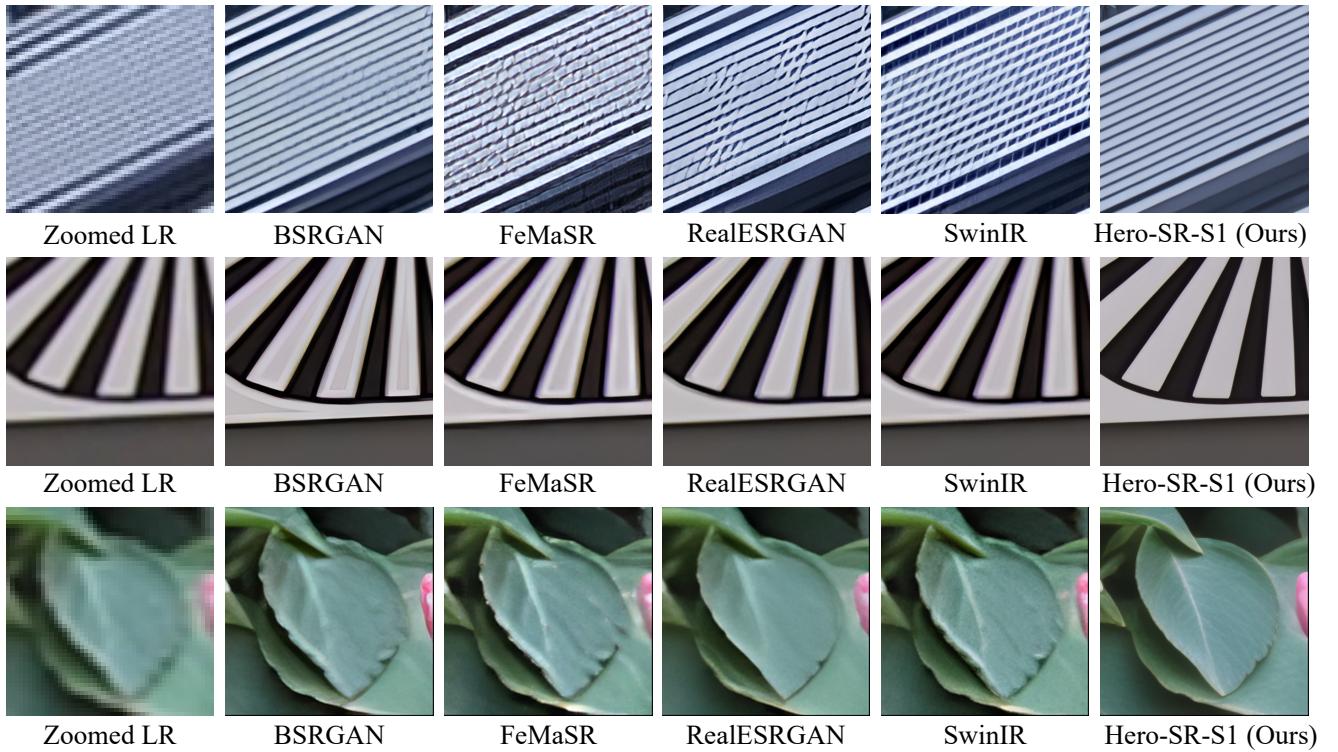


Figure 5. Qualitative comparison with GAN-base methods. **Zoom in for details.**

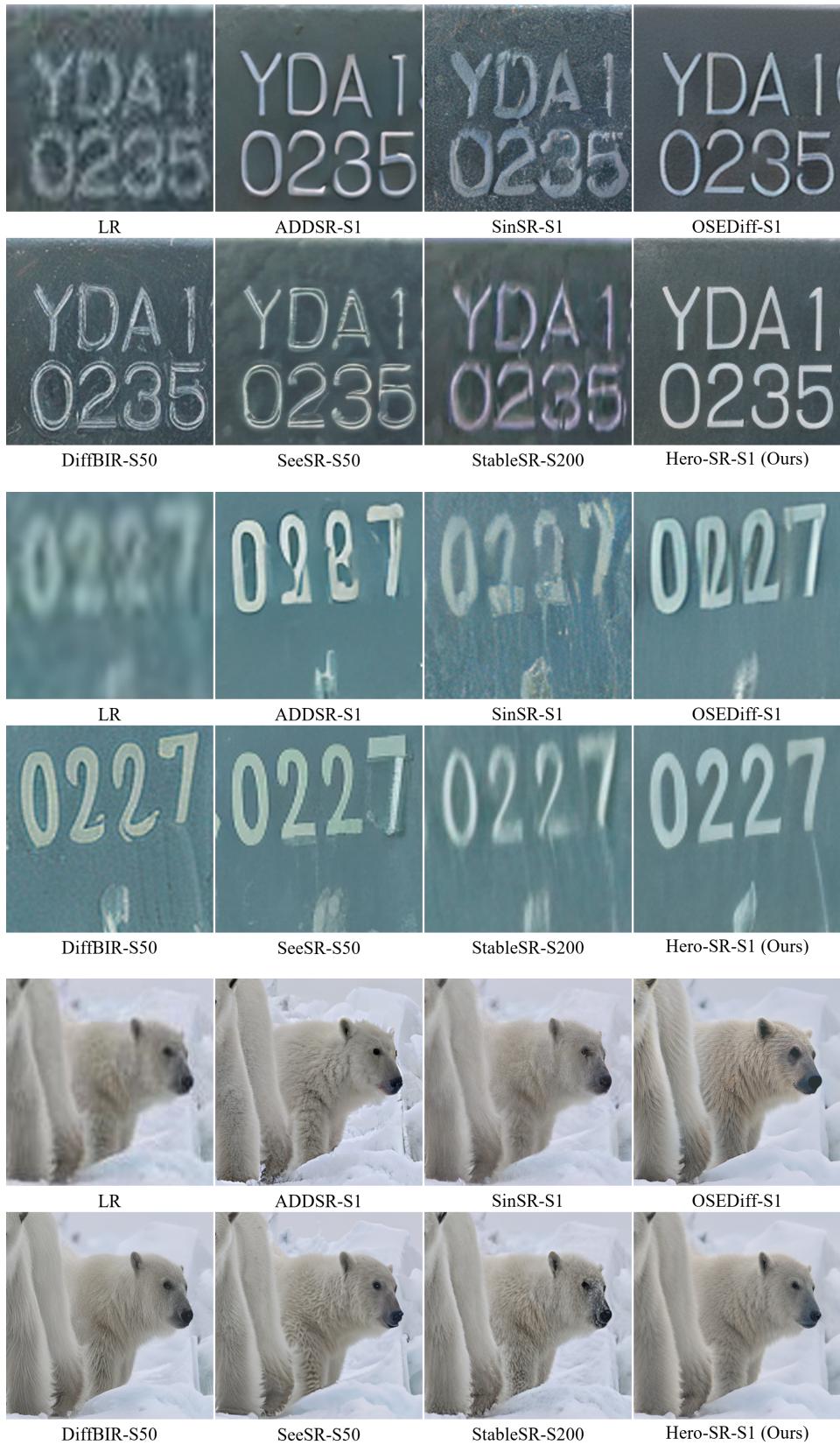


Figure 6. Qualitative comparison with one-step and multi-step methods. ‘S’ indicates the number of diffusion steps. **Zoom in for details.**

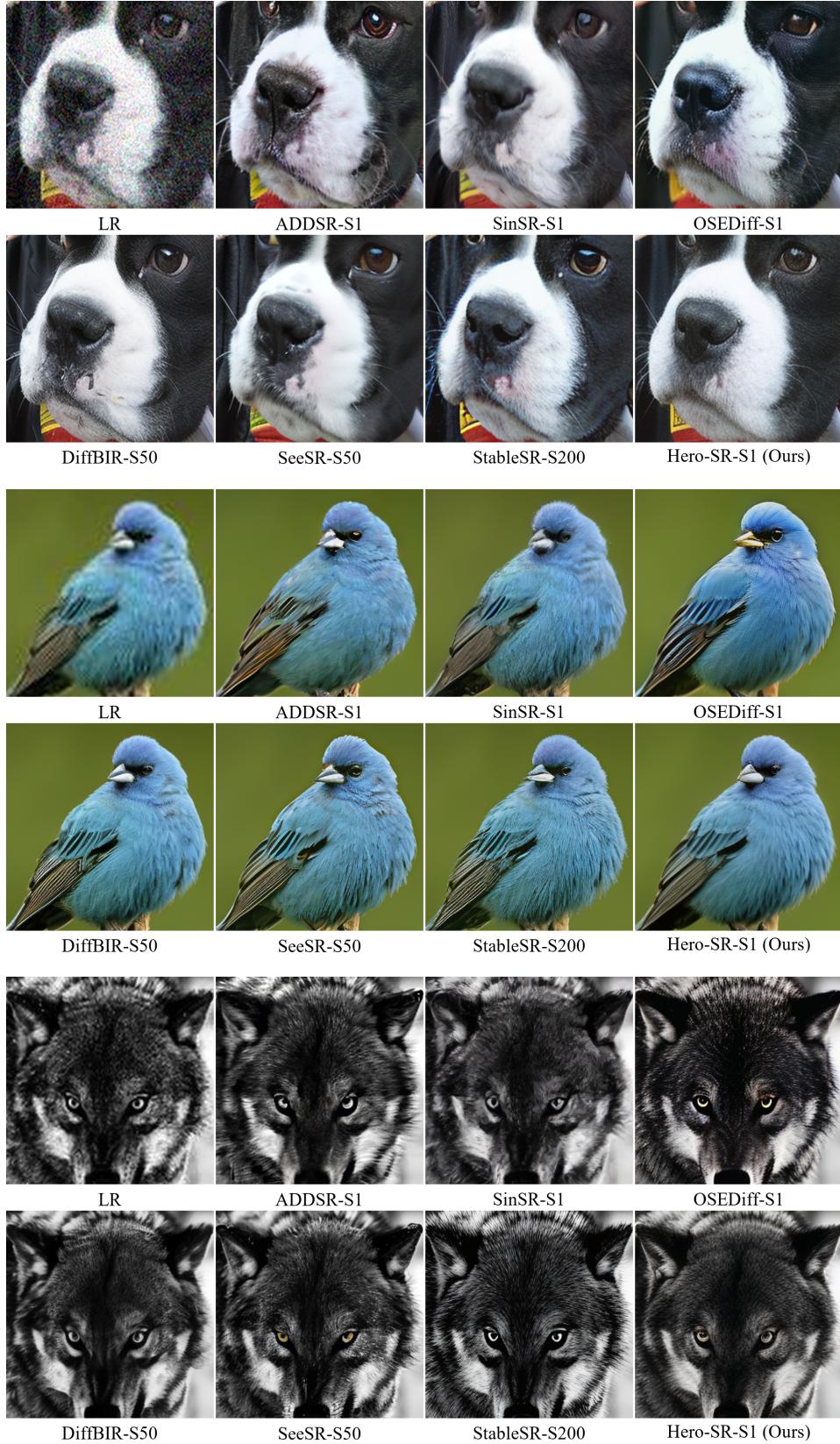


Figure 7. Qualitative comparison with one-step and multi-step methods. ‘S’ indicates the number of diffusion steps. **Zoom in for details.**

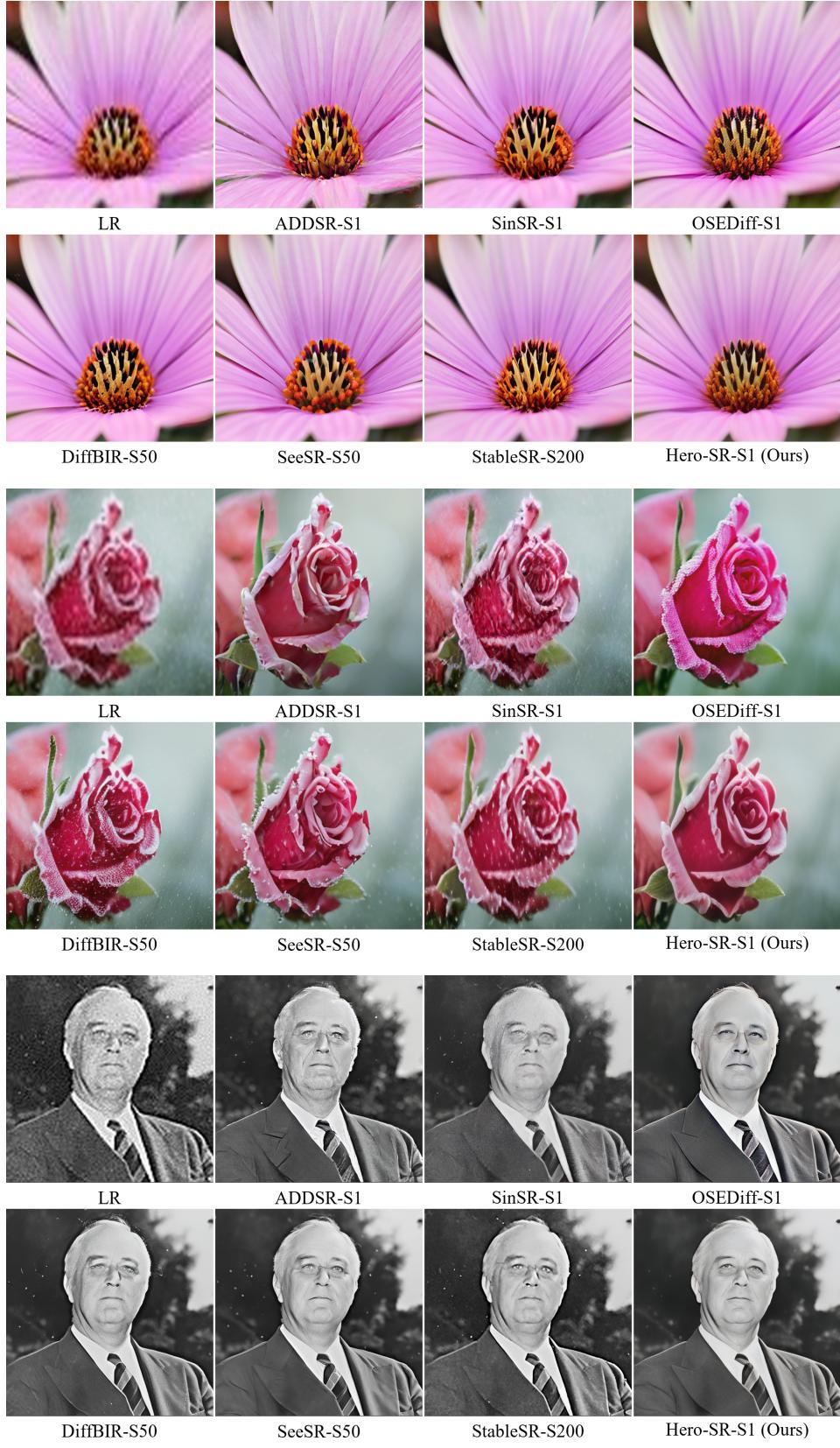


Figure 8. Qualitative comparison with one-step and multi-step methods. ‘S’ indicates the number of diffusion steps. **Zoom in for details.**



Figure 9. Qualitative comparison with one-step and multi-step methods. ‘S’ indicates the number of diffusion steps. **Zoom in for details.**