Selective Diffusion Distillation for Real-World High-Scale Image Super-Resolution

Wenli Zheng[1], Huiyuan Fu[1]*, Zekai Xu[1], Xin Wang[2], Huadong Ma[1]

[1]Beijing University of Posts and Telecommunications, China
[2]Stony Brook University, USA
{joyzheng, fhy, xzk, mhd}@bupt.edu.cn, x.wang@stonybrook.edu

Abstract

High-scale image super-resolution (SR) has become increasingly important with the rapid growth of mobile devices and high-resolution displays. However, current SR methods primarily focus on lower scales and generalize poorly to high-scale scenarios due to severe information loss and complex real-world degradations. In this paper, we propose a novel Selective Diffusion Distillation (SDD) framework for real-world high-scale SR, which distills reliable knowledge from a low-scale diffusion teacher to a high-scale student. Specifically, considering severe information loss in high-scale inputs, directly distilling from low-scale models may result in feature misalignment. To address this, we introduce a Degradation-aware Metric Learning (DML) approach to align feature distributions across different degradation levels. In addition, since the diffusion-based teacher may hallucinate artifacts in ambiguous regions, blindly imitating these unreliable outputs can degrade the student's fidelity. To tackle this, we propose a Region-aware Selective Distillation (RSD) strategy to filter out uncertain predictions and adaptively supervise only on reliable areas. To evaluate the effectiveness of our method, we introduce Real-UltraSR, a new real-world benchmark that contains diverse high-scale LR-HR pairs, including $\times 8$, $\times 10$, $\times 12$, and $\times 14$. Extensive experiments demonstrate that our SDD framework achieves state-of-the-art performance across multiple benchmarks.

Introduction

Real-world image super-resolution (Real-ISR) aims to reconstruct high-resolution (HR) images from their low-resolution (LR) counterparts under complex and unknown degradation conditions, which is a fundamental problem in the field of computer vision. This task is inherently ill-posed, as the degradation process often loses important details. Recovering these missing details requires not only strong generative capabilities but also robust generalization to diverse and unknown real-world degradations.

Generative adversarial networks (GANs) (Goodfellow et al. 2014) have played a pivotal role in modern image generation and restoration, inspiring many Real-ISR methods such as BSRGAN (Zhang et al. 2021) and Real-ESRGAN (Wang et al. 2021). Despite their success in enhancing perceptual quality, GAN-based methods often suffer from training instability, mode collapse, and limited controllability during inference. Recently, diffusion models and latent diffusion models (LDMs) (Rombach et al. 2022) have shown remarkable potential in generating high-fidelity images. Multi-step diffusion models (Wang et al. 2024; Wu et al. 2024b) have been applied to gradually denoise severely degraded inputs, but their computational cost limits real-world deployment. To address this, recent works such as OSEDiff (Wu et al. 2024a) propose the single-step diffusion architectures that balance between efficiency and reconstruction quality.

Figure 1: Visualization results of our method to achieve high-scale SR on the captured Real-UltraSR dataset.

However, existing diffusion-based methods are primarily designed for lower-scale SR (e.g., $\times 4$) and still face significant challenges when applied to high-scale scenarios (e.g., $\times 8$, $\times 10$, or $\times 12$). Firstly, compared with low-scale SR, the LR inputs of high-scale SR often suffer from more severe information loss, resulting in ambiguous textures and missing semantics. Directly training end-to-end models for high-scale SR becomes extremely difficult due to insufficient signal guidance. Secondly, real-world degradations become increasingly complex with higher scale factors. These degradations are particularly more severe in high-frequency regions such as edges and textures, making the accurate reconstruction more difficult.

To tackle these challenges, we propose a novel **Selec-**

*Corresponding author.

tive Diffusion Distillation (SDD) framework designed for real-world high-scale SR. Rather than directly training a high-scale model from scratch, SDD leverages the knowledge of a pre-trained low-scale diffusion teacher model to provide effective supervision. Specifically, considering the degradation gap between different scales, directly transferring knowledge can lead to feature misalignment and hinder student learning. To bridge this gap, we introduce a **Degradation-aware Metric Learning (DML)** scheme, which aligns latent feature distributions between low-scale and high-scale models, allowing the student to produce consistent representations across different degradation levels. Moreover, to handle the unreliability of teacher predictions in ambiguous regions, we propose a **Region-aware Selective Distillation (RSD)** strategy that estimates pixel-wise uncertainty map to identify unreliable regions of teacher predictions. The student is selectively supervised on reliable regions, thus enhancing robustness under severe degradation.

To evaluate real-world high-scale SR performance, we construct a novel benchmark, **Real-UltraSR**, which contains diverse real-world LR-HR pairs with high scale factors $\times 8$, $\times 10$, $\times 12$ and $\times 14$. Extensive experiments show that our method significantly outperforms existing state-of-the-art Real-ISR models in both quantitative metrics and visual quality, particularly in challenging high-scale SR scenarios.

In summary, our main contributions are as follows:

- We propose a novel Selective Diffusion Distillation (SDD) framework for real-world high-scale SR, which transfers reliable knowledge from a low-scale teacher to a high-scale student.
- We propose a Degradation-aware Metric Learning (DML) mechanism to align feature distributions across different degradation levels, facilitating effective cross-scale knowledge transfer.
- We introduce a Region-aware Selective Distillation (RSD) strategy that leverages pixel-wise uncertainty to supervise only reliable regions, enhancing robustness under severe degradations.
- We construct a new real-world high-scale SR dataset and conduct extensive experiments to verify the superiority of our method.

Related Work

Image Super-Resolution Classical single image super-resolution methods are mainly designed for simple degradations (eg. bicubic downsampling), achieving strong performance on synthetic datasets. However, these models often struggle in real-world SR tasks where the degradations are complex and unknown. To address these challenges, real-world SR methods simulate complex degradation processes (Zhang et al. 2021; Wang et al. 2021; Chen et al. 2022). Despite their success in perceptual quality enhancement, GAN-based models often introduce artifacts due to unstable adversarial training. Recent transformer-based SR models (Liang et al. 2021; Chen et al. 2023) demonstrate strong modeling capacity but are primarily trained under synthetic settings, limiting their robustness to unseen degradations. However, few works explore the high-scale SR, which remains an important challenge.

Diffusion-based Super-Resolution Diffusion models (Rombach et al. 2022) have recently emerged as powerful generative priors for image restoration tasks. Recently, several methods leverages generative priors to solve image SR problems (Wang et al. 2024; Wu et al. 2024b,a). StableSR (Wang et al. 2024) balances fidelity and perceptual quality by fine-tuning the time-aware encoder and employing controllable feature wrapping. SeeSR (Wu et al. 2024b) attempts to better stimulate the generative power of the SD model by extracting the semantic information in the image as a conditional guide. However, current diffusion-based SR methods primarily target low-scale SR (e.g. $\times 4$), they often perform poorly when applied into real-world high-scale SR. Our work addresses this limitation by selectively distilling reliable knowledge from low-scale diffusion teachers to guide high-scale reconstruction.

Knowledge Distillation Knowledge distillation (KD) is a widely adopted paradigm for transferring knowledge from a powerful teacher model to a lightweight student (Hinton, Vinyals, and Dean 2015). KD has been successfully applied across various computer vision tasks, including image restoration (Hui et al. 2019; Li et al. 2022), classification (Xu et al. 2020; Zhang et al. 2022; Qu et al. 2022), and object detection (Chen et al. 2017; Dai et al. 2021; Zhang and Ma 2023). However, applying KD to super-resolution remains challenging (Zhang et al. 2023a), which requires precise recovery of high-frequency details. Most KD methods treat all spatial regions equally, neglecting the varying difficulty across different areas of an image, leading to suboptimal restoration in complex areas. To address this, we propose a Region-aware Selective Distillation (RSD) strategy that adaptively distills reliable teacher knowledge, enhancing detail fidelity and robustness under severe degradations.

Methodology

Preliminary

Our proposed Selective Diffusion Distillation (SDD) framework builds upon an efficient single-step diffusion backbone (Wu et al. 2024a), which comprises a VAE encoder $\mathcal{E}$, a UNet-based denoiser ϵ_θ, a VAE decoder $\mathcal{G}$, and a semantic prompt extractor $\mathcal{C}$. All components are finetuned with lightweight adapters to ensure adaptability across scales. Given an input low-resolution image x_{LR}, we first obtain its latent representation and semantic condition:

$$z_{\text{LR}} = \mathcal{E}(x_{\text{LR}}), \quad c = \mathcal{C}(x_{\text{LR}}), \tag{1}$$

The latent is then refined via a one-step denoising process with noise level $\bar{\alpha}_T$:

$$\hat{z}_{\text{HR}} = \frac{z_{\text{LR}} - \sqrt{1-\bar{\alpha}_T} \cdot \epsilon_\theta(z_{\text{LR}}, T, c)}{\sqrt{\bar{\alpha}_T}}, \tag{2}$$

Finally, the high-resolution prediction is reconstructed through:

$$\hat{x}_{\text{HR}} = \mathcal{G}(\hat{z}_{\text{HR}}), \tag{3}$$

This single-step formulation offers a strong generative prior with high efficiency, serving as a practical backbone for our SDD framework.

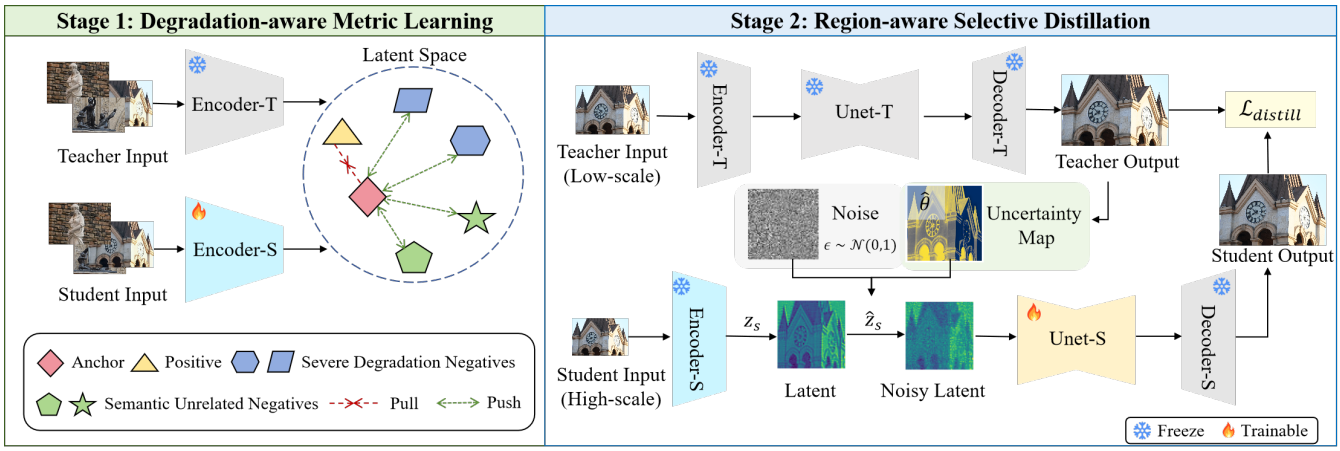

Figure 2: Overview of the proposed Selective Diffusion Distillation (SDD) framework. Given the same HR image degraded into different LR inputs, a high-scale student model learns from a pre-trained low-scale teacher through two core components: Degradation-aware Metric Learning (DML) that aligns latent features to bridge the degradation gap across different scales, and Region-aware Selective Distillation (RSD) that estimates region-wise uncertainty in the teacher's output and adaptively supervises only on reliable regions.

Overview

Our goal is to improve high-scale super-resolution by selectively transferring reliable knowledge from a pre-trained low-scale diffusion model. As illustrated in Fig. 2, we propose a Selective Diffusion Distillation framework (SDD), which enhances the robustness and generalization capability of the student model under complex real-world degradations. The training process is divided into two stages. In the first stage, we optimize the student's encoder using a Degradation-aware Metric Learning (DML) strategy to align feature distributions across different degradation levels, thereby bridging the domain gap between teacher and student inputs. In the second stage, we optimize the student's UNet denoiser with a Region-aware Selective Distillation (RSD) strategy, which estimates region-wise uncertainty in the teacher's output and adaptively supervises only on reliable regions. Throughout training, the teacher model remains frozen and serves as a reliable source of low-scale priors. By combining degradation-level alignment and region-wise supervision, our SDD framework effectively improves the performance of the student model under challenging high-scale SR scenarios.

Degradation-aware Metric Learning

In high-scale SR, the degradation of LR inputs becomes increasingly severe and complex as the scale factor increases, which leads to a significant mismatch between the feature distributions between low-scale and high-scale SR models. This degradation gap severely hinders effective knowledge transfer, as the student encounters more challenging degradation conditions than the teacher. To address this, we introduce a Degradation-aware Metric Learning (DML) strategy, which leverages metric learning to align the latent features across different degradation levels. Specifically, we aim to pull the student's latent features closer to lower-degradation features, while pushing them away from higher-degradation features.

Given a batch of N ground-truth images $\{y_i\}_{i=1}^{N}$, we generate corresponding teacher and student inputs at different degradation scales:

$$\{x_i^t, x_i^s, x_i^h\} = \mathcal{D}(y_i, \{s_t, s_s, s_h\}), \quad s_t < s_s < s_h \tag{4}$$

where $\mathcal{D}(y_i, s)$ denotes the degradation function that degrades image y_i at scale s. Here, s_t, s_s, and s_h represent mild, moderate, and severe degradations, respectively. The inputs x_i^t and x_i^h are encoded by a pretrained teacher encoder $\mathcal{E}_T$, and x_i^s is passed through the student encoder $\mathcal{E}_S$ to produce latent features:

$$\mathbf{z}_i^t = \mathcal{E}_T(x_i^t), \quad \mathbf{z}_i^s = \mathcal{E}_S(x_i^s), \quad \mathbf{z}_i^h = \mathcal{E}_T(x_i^h) \tag{5}$$

To further introduce semantic discrimination, we randomly sample a different image y_j $(j \neq i)$ from the same batch and apply the same degradation scale s_s to obtain a semantically unrelated feature:

$$x_j^s = \mathcal{D}(y_j, s_s), \quad \mathbf{z}_j^s = \mathcal{E}_S(x_j^s) \tag{6}$$

To explicitly enforce the degradation-aware alignment, we introduce a metric loss inspired by triplet-based contrastive learning (Ge 2018). Specifically, the student latent feature $\mathbf{z}_i^s$ serves as the anchor, the mildly degraded teacher feature $\mathbf{z}_i^t$ acts as the positive, and the negatives include both the severely degraded teacher feature $\mathbf{z}_i^h$ and the semantically unrelated student feature $\mathbf{z}_j^s$. The metric loss is formulated as:

$$\mathcal{L}_{\text{metric}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|\mathbf{z}_i^s - \mathbf{z}_i^t\|_2}{\|\mathbf{z}_i^s - \mathbf{z}_i^h\|_2 + \epsilon} + \lambda \cdot \frac{\|\mathbf{z}_i^s - \mathbf{z}_i^t\|_2}{\|\mathbf{z}_i^s - \mathbf{z}_j^s\|_2 + \epsilon}\right) \tag{7}$$

where ϵ is a small constant to ensure numerical stability, and λ balances the weight between degradation-based and semantic-based contrastive objectives.

By optimizing this degradation-aware metric loss, the student model is encouraged to align with lower-scale degradation representations, while remaining distinct from features arising from different semantic content. This facilitates robust knowledge transfer and enhances the student's ability to reconstruct high-quality images from severely degraded inputs.

Region-aware Selective Distillation

Super-resolution is inherently an ill-posed problem, recovering high-fidelity HR images from heavily degraded inputs is highly ambiguous, particularly in high-frequency regions such as textures and edges. These areas typically suffer from more severe degradation than flat regions, posing greater challenges for accurate reconstruction. While diffusion-based teacher models can generate visually pleasing results, their predictions in complex regions are not always reliable. Blindly imitating such uncertain outputs will lead to artifacts in the student model. To address this, we propose a Region-aware Selective Distillation (RSD) strategy that estimates uncertain regions in the teacher's output and adaptively adjusts the supervision signal to enhance the robustness of student model. Our RSD consists of three key components: Uncertainty Estimation (UE), Noise Injection (NI), and Region-aware Distillation Loss.

Uncertainty Estimation Although diffusion models achieve impressive visual quality, their predictions are not equally reliable across all regions, so it's necessary to selectively distill knowledge based on region-specific reliability. To address this, we introduce a Uncertainty Estimation (UE) mechanism that estimates the uncertainty of each region in the teacher's output. Specifically, we apply Softplus activation (Zheng et al. 2015) followed by an $\mathcal{L}_1$ normalization to compute the uncertainty map θ:

$$\theta = \psi(\text{Softplus}(\hat{y}_T)) \tag{8}$$

where ψ denotes the normalization, $\hat{y}_T$ is the teacher's SR output. As shown in Fig. 3, brighter regions in the uncertainty map correspond to areas requiring further refinement. Conversely, darker regions indicate more confident predictions suitable for direct supervision.

Noise Injection Recent works in image editing (Zhang et al. 2023b; Shi et al. 2024), image inpainting (Ju et al. 2024; Shi et al. 2024), and virtual try-on (Zhu et al. 2023; Xu et al. 2025) often employ explicit masks to guide diffusion models towards specific regions. Inspired by this, we avoid uniformly injecting Gaussian noise into all latent pixels. Instead, we selectively perturb uncertain regions of the student's latent features. Specifically, we first downsample the uncertainty map θ to obtain $\hat{\theta}$ to match the spatial resolution of latent features z_S. Then, regions with higher uncertainty are applied more noise, while regions with lower uncertainty are applied less noise. Formally, the updated latent features $\hat{z}_S$ are defined as:

$$\hat{z}_S = \hat{\theta} \cdot \epsilon + (1 - \hat{\theta}) \cdot z_S, \tag{9}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is Gaussian noise. This targeted perturbation guides the student model to reconstruct challenging regions, enhancing robustness and fidelity.

LR SR Uncertainty Map

Figure 3: The estimated uncertainty map of the teacher's SR output, brighter regions correspond to areas requiring further refinement, darker regions correspond to confident areas.

Region-aware Distillation Loss To further guide the student model with region-specific supervision, we introduce a Region-aware Distillation loss that adaptively balances soft and hard supervision based on the uncertainty map θ. Specifically, regions with low uncertainty are supervised by the teacher's output, while high-uncertainty regions are directly supervised by the ground truth. The distillation loss is formulated as:

$$\mathcal{L}_{\text{distill}} = \lambda_{\text{soft}} \cdot \|(1-\theta) \cdot (\hat{y}_S - \hat{y}_T)\|_2 + \lambda_{\text{hard}} \cdot \|\theta \cdot (\hat{y}_S - y)\|_2 \tag{10}$$

where $\hat{y}_S$ denotes the student's SR prediction, y is the ground-truth. The weighting coefficients λ_{soft} and λ_{hard} balance the contributions of soft distillation and hard supervision. In our experiments, we set $\lambda_{\text{soft}} = 1.0$ and $\lambda_{\text{hard}} = 1.5$. This loss encourages the student to rely on reliable guidance while avoiding the propagation of uncertain details.

Experiments

Experimental Settings

Training settings. Following prior works (Dong et al. 2025), we adopt LSDIR (Li et al. 2023) and the first 10K images from FFHQ (Karras, Laine, and Aila 2019) as the training data. We use the degradation pipeline Real-ESRGAN (Wang et al. 2021) to synthesize LR-HR pairs. The training patch size is 512×512.

Test Datasets. We evaluate the competing methods using synthetic and real-world test data. The synthetic dataset comprises 3000 images cropped from DIV2K (Agustsson and Timofte 2017) with size 512×512, degraded using Real-ESRGAN degradation pipeline (Wang et al. 2021). The real-world data are center-cropped from RealSR (Cai et al. 2019) and DRealSR (Wei et al. 2020) datasets with size 128×128 for LR images and 512×512 for HR images.

Additionally, to evaluate high-scale SR performance, we construct a new benchmark named **Real-UltraSR**, containing ×8, ×10, ×12 and ×14 LR-HR pairs. Specifically, we use a Nikon Z6 II camera equipped with a high-range optical-zoom lens (28-400 mm) for data collection. To ensure alignment, we adopt a two-stage image registration pipeline, including luminance normalization and SIFT (Lowe 2004) keypoint matching. After registration, all HR images are cropped to 3024×2008, and corresponding LR images are cropped according to the target scale. Our dataset includes both indoor and outdoor real-world scenes. More

Test Dataset	Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	NIQE ↓	MUSIQ ↑	CLIPIQA ↑
×8 SR (real-world)	Real-ESRGAN	22.35	0.7274	0.3614	50.85	4.76	53.25	0.6571
	BSRGAN	22.74	0.7346	0.3721	51.48	4.69	54.32	0.6780
	FeMaSR	21.05	0.6994	0.3880	53.93	4.50	51.92	0.6842
	StableSR-200s	22.03	0.7103	0.3789	46.82	4.74	55.32	0.6902
	SeeSR-50s	22.21	0.7186	0.3744	45.62	4.62	57.00	0.6984
	OSEDiff-1s	22.34	0.7237	0.3695	44.90	4.61	57.45	0.7066
	AdcSR-1s	22.51	0.7273	0.3712	44.62	4.59	58.02	0.7098
	TSD-SR-1s	22.12	0.7046	0.4042	44.42	4.56	58.36	0.7161
	Ours-1s	22.67	0.7321	0.3680	43.26	4.48	58.68	0.7216
×10 SR (real-world)	Real-ESRGAN	21.87	0.7251	0.3510	52.80	4.93	52.18	0.6384
	BSRGAN	22.35	0.7298	0.3621	54.12	4.86	53.74	0.6641
	FeMaSR	20.70	0.6890	0.3795	56.77	4.61	50.91	0.6703
	StableSR-200s	21.62	0.7014	0.3680	48.67	4.90	54.67	0.6752
	SeeSR-50s	21.88	0.7097	0.3653	47.23	4.78	56.79	0.6864
	OSEDiff-1s	21.97	0.7155	0.3596	46.30	4.75	57.20	0.6936
	AdcSR-1s	22.18	0.7182	0.3615	46.02	4.74	58.20	0.6964
	TSD-SR-1s	21.80	0.6943	0.3985	45.95	4.65	58.43	0.7012
	Ours-1s	22.32	0.7256	0.3587	42.86	4.54	59.22	0.7202
×12 SR (real-world)	Real-ESRGAN	21.72	0.7124	0.3508	53.40	4.96	52.00	0.6341
	BSRGAN	22.25	0.7235	0.3619	54.35	4.89	53.46	0.6614
	FeMaSR	20.60	0.6825	0.3783	56.70	4.67	50.45	0.6661
	StableSR-200s	21.55	0.6952	0.3674	48.89	4.91	54.20	0.6722
	SeeSR-50s	21.81	0.7044	0.3642	47.10	4.80	56.38	0.6837
	OSEDiff-1s	21.95	0.7116	0.3592	46.30	4.77	57.01	0.6905
	AdcSR-1s	22.05	0.7155	0.3612	46.05	4.75	58.00	0.6958
	TSD-SR-1s	21.66	0.6912	0.3956	45.85	4.73	58.28	0.7011
	Ours-1s	22.14	0.7199	0.3596	45.37	4.62	58.59	0.7037

Table 1: Quantitative comparison with state-of-the-art methods on the Real-UltraSR dataset. RealESRGAN (ICCV2021), BSRGAN (ICCV2021) and FeMaSR (ACMMM2022) are GAN-based method. StableSR (IJCV2024), SeeSR (CVPR2024), OSEDiff (NeurIPS2024), AdcSR and TSD-SR (CVPR2025) are diffusion-based methods. The number of diffusion inference steps is indicated by 's'. The best and second performances are marked in red and blue, respectively.

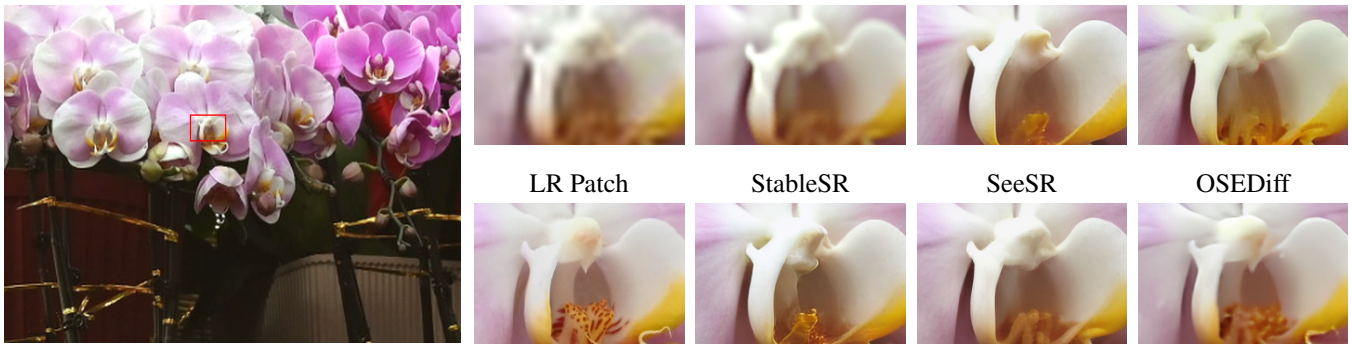

Figure 4: Visualization comparisons of different Real-ISR models on the Real-UltraSR dataset. Please zoom in for a better view.

details of Real-UltraSR are provided in the **supplementary material**.

Evaluation Metrics. To conduct a comprehensive assessment, we use both the reference-based metrics and non-reference-based metrics. Reference-based metrics, including PSNR and SSIM (Wang et al. 2004) (calculated on the Y channel in YCbCr space) for structural fidelity, LPIPS (Zhang et al. 2018) for perceptual quality, and FID (Heusel et al. 2017) for distribution comparison. Non-reference metrics such as NIQE (Mittal, Soundararajan, and Bovik 2012), MUSIQ (Ke et al. 2021), and CLIPIQA (Wang, Chan, and Loy 2023) are used to assess perceptual quality in real-world

Test Dataset	Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	NIQE ↓	MUSIQ ↑	CLIPIQA ↑
DIV2K-Val (synthetic)	Real-ESRGAN	24.29	0.6371	0.3112	37.64	4.68	61.06	0.5277
	BSRGAN	24.58	0.6269	0.3351	44.23	4.75	61.20	0.5247
	FeMaSR	23.06	0.5887	0.3126	35.87	4.74	60.83	0.5997
	StableSR-200s	23.26	0.5726	0.3113	24.44	4.76	65.92	0.6192
	SeeSR-50s	23.68	0.6043	0.3194	25.90	4.81	68.67	0.6936
	OSEDiff-1s	23.72	0.6108	0.2941	26.32	4.71	67.97	0.6683
	AdcSR-1s	23.74	0.6017	0.2853	25.52	4.36	68.00	0.6764
	TSD-SR-1s	23.02	0.5808	0.2673	29.16	4.32	71.69	0.7416
	Ours-1s	23.81	0.6083	0.2616	23.18	4.19	72.93	0.7613
RealSR (real-world)	Real-ESRGAN	25.69	0.7616	0.2727	135.18	5.83	60.18	0.4449
	BSRGAN	26.39	0.7654	0.2670	141.28	5.66	63.21	0.5001
	FeMaSR	25.07	0.7358	0.2942	141.05	5.79	58.95	0.5270
	StableSR-200s	24.70	0.7085	0.3018	128.51	5.91	65.78	0.6178
	SeeSR-50s	25.18	0.7216	0.3009	125.55	5.41	69.77	0.6612
	OSEDiff-1s	25.15	0.7341	0.2921	123.49	5.65	69.09	0.6693
	AdcSR-1s	25.47	0.7301	0.2885	118.41	5.35	69.90	0.6731
	TSD-SR-1s	24.81	0.7172	0.2743	114.45	5.12	71.19	0.7160
	Ours-1s	25.51	0.7409	0.2631	112.36	5.08	72.06	0.7314
DRealSR (real-world)	Real-ESRGAN	28.64	0.8053	0.2847	147.62	6.69	54.18	0.4422
	BSRGAN	28.75	0.8031	0.2883	155.63	6.52	57.14	0.4915
	FeMaSR	26.90	0.7572	0.3169	157.78	5.90	53.74	0.5464
	StableSR-200s	28.03	0.7536	0.3284	148.98	6.52	58.51	0.6356
	SeeSR-50s	28.17	0.7691	0.3189	147.39	6.40	64.93	0.6804
	OSEDiff-1s	27.92	0.7835	0.2968	135.30	6.49	64.65	0.6963
	AdcSR-1s	28.10	0.7726	0.3046	134.05	6.45	66.26	0.7049
	TSD-SR-1s	27.77	0.7559	0.3967	134.98	5.91	66.62	0.7344
	Ours-1s	28.33	0.7793	0.2946	126.51	5.87	67.86	0.7691

Table 2: Quantitative comparison with state-of-the-art methods on real-world x4 SR benchmarks. RealESRGAN (ICCV2021), BSRGAN (ICCV2021) and FeMaSR (ACMMM2022) are GAN-based methods. StableSR (IJCV2024), SeeSR (CVPR2024), OSEDiff (NeurIPS2024), AdcSR and TSD-SR (CVPR2025) are diffusion-based methods. The number of diffusion inference steps is indicated by 's'. The best and second performances are marked in red and blue, respectively.

scenarios.

Implementation Details. We train our SDD framework using the AdamW optimizer (Loshchilov and Hutter 2017) with a learning rate of 5e-5. The UNet denoisers of both teacher and student are initialized from the pre-trained weights of SD2.1-base. In the first training stage, the student's VAE encoder is trained using the degradation-aware metric loss $\mathcal{L}_{\text{metric}}$. In the second stage, the diffusion denoiser is optimized with the region-aware distillation loss $\mathcal{L}_{\text{distill}}$, while the VAE decoder is kept frozen to preserve its prior knowledge. All experiments are conducted on 4 NVIDIA V100 GPUs with a batch size of 16.

Comparison with state-of-the-arts

Comparison Methods. We compare our method with leading DM-based methods: StableSR (Wang et al. 2024), SeeSR (Wu et al. 2024b), OSEDiff (Wu et al. 2024a), AdcSR (Chen et al. 2025), TSD-SR (Dong et al. 2025) and GAN-based SR methods: RealESRGAN (Wang et al. 2021), BSRGAN (Zhang et al. 2021), FeMASR (Chen et al. 2022). All comparative results are obtained using officially released codes or models.

Quantitative Comparison. We first evaluate high-scale SR performance of the competing methods on the Real-UltraSR dataset. The quantitative results are shown in Tab. 1. Although GAN-based methods (Wang et al. 2021; Zhang et al. 2021) perform best in terms of PSNR and SSIM metrics, they exhibit poor performance in terms of MUSIQ and CLIPIQA metrics. Our method achieves the best results on the perceptual-oriented metrics. Note that our method clearly improves the results, especially in the challenging cases of severe degradations, which demonstrates the effectiveness of our approach. We further evaluate the proposed method on the low-scale ($\times 4$) SR benchmarks, including synthetic dataset DIV2K (Agustsson and Timofte 2017) and real-world dataset RealSR (Cai et al. 2019), DRealSR (Wei et al. 2020). As shown in Tab. 2, our method outperforms competing methods on the perceptual-oriented metrics. To provide a fair and consistent comparison in terms of model efficiency, we evaluate the model complexity and inference time for several diffusion-based Real-ISR methods. The results are summarized in Table 4.

Qualitative Comparison. The qualitative results of competing methods are shown in Fig. 4, our results exhibit clearer textures and more vivid colors than other methods.

Method	Real-UltraSR ×8 Testset				Real-UltraSR ×10 Testset			
	PSNR↑	NIQE↓	MUSIQ↑	CLIPIQA↑	PSNR↑	NIQE↓	MUSIQ↑	CLIPIQA↑
w/o DML	22.21	4.65	57.42	0.7078	21.85	4.68	57.90	0.7113
w/o RSD	22.13	4.77	57.06	0.6985	21.74	4.73	57.33	0.7026
w/o UE	22.32	4.59	57.85	0.7102	22.04	4.61	58.24	0.7150
w/o NI	22.26	4.62	57.93	0.7093	22.07	4.60	58.30	0.7147
w/o Soft Loss	22.35	4.55	58.04	0.7127	22.15	4.56	58.55	0.7171
w/o Hard Loss	22.41	4.53	58.12	0.7141	22.21	4.55	58.78	0.7190
Ours	**22.67**	**4.48**	**58.68**	**0.7216**	**22.32**	**4.54**	**59.22**	**0.7202**

Table 3: Ablation study on effects of the proposed components, including the Degradation-aware Metric Learning (DML) and Region-aware Selective Distillation (RSD) strategy. The best results are highlighted in bold.

Figure 5: Qualitative comparisons of ablation study on the Real-UltraSR dataset. Please zoom in for a better view.

Ablation Study

We conduct ablation studies to validate the contribution of each component in our SDD framework. The quantitative results on the Real-UltraSR testsets are summarized in Tab. 3 and qualitative results are shown in Fig. 5. Specifically, we perform the following ablation experiments:

Impact of Degradation-aware Metric Learning (DML). Removing DML leads to a clear performance drop, demonstrating its critical role in bridging the feature distribution gap across degradation levels. By promoting degradation-invariant representations, DML enables the student to better generalize across scales and enhances both fidelity and perceptual quality.

Impact of Region-aware Selective Distillation (RSD). Without RSD, the student receives equal supervision across all regions, including uncertain areas with hallucinated teacher outputs, leading to a clear quality drop. Removing the Uncertainty Estimation (UE) also also degrades performance, which validates the necessity of explicit uncertainty modeling. Similarly, without Noise Injection (NI), the student lacks focused guidance in uncertain regions, resulting in poor reconstruction. We further validate the effect of Region-aware Distillation Loss: soft loss (teacher-based) and hard loss (ground-truth-based). Removing the soft loss slightly reduces perceptual quality, while removing the hard loss more strongly impacts structural and semantic consistency. In Fig. 5, the full model best restores sharp text and emblem details, while removing DML, RSD, UE, or NI introduces blurring or artifacts, confirming the value of each component.

Method	**Params (M)**	**Time (s)**
StableSR-200s	1410	11.50
SeeSR-50s	2524	4.30
OSEDiff-1s	1775	0.31
AdcSR-1s	456	0.11
TSDSR-1s	2150	0.13
Ours-1s	1320	0.10

Table 4: Comparison of model complexity and inference time. The number of diffusion inference steps is indicated by 's'.

Conclusion

In this paper, we propose a novel Selective Diffusion Distillation (SDD) framework for real-world high-scale super-resolution. We introduce a Degradation-aware Metric Learning (DML) module to bridge the domain gap between low- and high-scale degradations and a Region-aware Selective Distillation (RSD) strategy to provide spatially adaptive supervision based on prediction reliability. Extensive experiments on the newly introduced Real-UltraSR benchmark demonstrate that SDD significantly outperforms existing methods in real-world high-scale scenarios.

Acknowledgments

This work is supported in part by the NSFC under No.62272059 and No.U24B20176, the Beijing Natural Science Foundation under No.JQ24020, the National Key R&D Program of China under No.2023YFF0904800, and the Beijing Nova Program under No.20230484406.

References

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

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

Chen, B.; Li, G.; Wu, R.; Zhang, X.; Chen, J.; Zhang, J.; and Zhang, L. 2025. Adversarial diffusion compression for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28208–28220.

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

Chen, G.; Choi, W.; Yu, X.; Han, T.; and Chandraker, M. 2017. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22367–22377.

Dai, X.; Jiang, Z.; Wu, Z.; Bao, Y.; Wang, Z.; Liu, S.; and Zhou, E. 2021. General instance distillation for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7842–7851.

Dong, L.; Fan, Q.; Guo, Y.; Wang, Z.; Zhang, Q.; Chen, J.; Luo, Y.; and Zou, C. 2025. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23174–23184.

Ge, W. 2018. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European conference on computer vision (ECCV)*, 269–285.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hui, Z.; Gao, X.; Yang, Y.; and Wang, X. 2019. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, 2024–2032.

Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, 150–168. Springer.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

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

Li, J.; Yang, H.; Yi, Q.; Fang, F.; Gao, G.; Zeng, T.; and Zhang, G. 2022. Multiple degradation and reconstruction network for single image denoising via knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 558–567.

Li, Y.; Zhang, K.; Liang, J.; Cao, J.; Liu, C.; Gong, R.; Zhang, Y.; Tang, H.; Liu, Y.; Demandolx, D.; et al. 2023. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1775–1787.

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

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2): 91–110.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Qu, L.; Wang, M.; Song, Z.; et al. 2022. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems*, 35: 15368–15381.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shi, Y.; Xue, C.; Liew, J. H.; Pan, J.; Yan, H.; Zhang, W.; Tan, V. Y.; and Bai, S. 2024. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8839–8849.

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

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2024. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12): 5929–5949.

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

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; and Lin, L. 2020. Component divide-and-conquer for real-world image super-resolution. In *European conference on computer vision*, 101–117. Springer.

Wu, R.; Sun, L.; Ma, Z.; and Zhang, L. 2024a. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553.

Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; and Zhang, L. 2024b. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25456–25467.

Xu, K.; Rui, L.; Li, Y.; and Gu, L. 2020. Feature normalized knowledge distillation for image classification. In *European conference on computer vision*, 664–680. Springer.

Xu, Y.; Gu, T.; Chen, W.; and Chen, A. 2025. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8996–9004.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18802–18812.

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

Zhang, L.; and Ma, K. 2023. Structured knowledge distillation for accurate and efficient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15706–15724.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Y.; Li, W.; Li, S.; Chen, H.; Tu, Z.; Wang, W.; Jing, B.; Lin, S.; and Hu, J. 2023a. Data upcycling knowledge distillation for image super-resolution. *arXiv preprint arXiv:2309.14162*.

Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D. N.; and Ren, J. 2023b. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6027–6037.

Zheng, H.; Yang, Z.; Liu, W.; Liang, J.; and Li, Y. 2015. Improving deep neural networks using softplus units. In *2015 International joint conference on neural networks (IJCNN)*, 1–4. IEEE.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4606–4615.