

Enhanced Semantic Extraction and Guidance for UGC Image Super Resolution

Yiwen Wang¹ Ying Liang¹ Yuxuan Zhang¹ Xinning Chai¹ Zhengxue Cheng^{1*}

Yingsheng Qin² Yucai Yang² Rong Xie¹ Li Song¹

¹Shanghai Jiao Tong University, China ²Transsion, China

{evonwang, forest726, 67keudyhs, chaixinning, zxcheng, xierong, song_li}@sjtu.edu.cn
{yingsheng.qin, yucai.yang}@transsion.com

Abstract

Due to the disparity between real-world degradations in user-generated content(UGC) images and synthetic degradations, traditional super-resolution methods struggle to generalize effectively, necessitating a more robust approach to model real-world distortions. In this paper, we propose a novel approach to UGC image super-resolution by integrating semantic guidance into a diffusion framework. Our method addresses the inconsistency between degradations in wild and synthetic datasets by separately simulating the degradation processes on the LSDIR dataset and combining them with the official paired training set. Furthermore, we enhance degradation removal and detail generation by incorporating a pretrained semantic extraction model (SAM2) and fine-tuning key hyperparameters for improved perceptual fidelity. Extensive experiments demonstrate the superiority of our approach against state-of-the-art methods. Additionally, the proposed model won second place in the CVPR NTIRE 2025 Short-form UGC Image Super-Resolution Challenge (Report [14, 15]), further validating its effectiveness. The code is available at <https://github.com/Moonsofang/NTIRE-2025-SRlab>.

1. Introduction

Single Image Super-Resolution (SISR) is a fundamental task in computer vision that focuses on reconstructing high-resolution (HR) images from their low-resolution (LR) counterparts. The primary goal of SISR is to recover fine-grained details and high-frequency textures that are lost during the downsampling process, ultimately enhancing the perceptual quality and fidelity of the upscaled images.

Early deep learning-based approaches predominantly relied on convolutional neural networks (CNNs) [4, 10, 11, 20, 44, 45], which leverage hierarchical feature extraction



Figure 1. Objective and subjective results of NTIRE 2025 Short-form UGC Image SR Challenge. The top six methods are included. The horizontal axis represents the objective score, which is computed as $\text{Score} = \text{PSNR} + 10 \times \text{SSIM} - 10 \times \text{LPIPS} + 0.1 \times \text{MUSIQ} + 10 \times \text{ManIQA} + 10 \times \text{CLIPQA}$. The vertical axis represents the subjective score calculated by five experts. All results above are provided by the competition organizer.

to learn LR-to-HR mappings. While these methods achieve impressive PSNR, they often produce overly smooth textures due to their reliance on pixel-wise loss functions, failing to capture high-frequency details essential for perceptual realism. By adversarial training, GAN-based methods [8, 13, 19, 32, 33, 41] synthesize visually plausible textures, significantly improving perceptual quality. However, their tendency to hallucinate visually plausible yet semantically inconsistent details remains a critical limitation for faithful super-resolution.

Unlike GANs, Diffusion models(DMs) [7] iteratively refine images through a Markov chain-based denoising process. This progressive nature allows for finer control over the trade-off between distortion and perceptual quality, making them particularly well-suited for high-fidelity image generation. Additionally, their inherent stochasticity enables the generation of diverse high-resolution (HR) sam-

*Corresponding author

ples from a single low-resolution (LR) input, a distinct advantage over traditional deterministic approaches such as CNNs and GANs, which typically produce a single fixed output. Recently, a growing number of diffusion-based super-resolution (SR) methods [3, 5, 21, 24, 30, 34, 35, 37–40] have emerged, further demonstrating the effectiveness of this framework in enhancing image quality and restoring fine details.

However, despite their promising capabilities, diffusion models still face several critical challenges when applied to real-world super-resolution tasks.

First, most existing methods rely on synthetically generated low-resolution (LR) and high-resolution training pairs, where degradation processes, such as bicubic downsampling, fail to accurately replicate the complex and heterogeneous degradations found in real-world images. This discrepancy creates a domain gap, making it difficult for models trained on synthetic data to generalize effectively to real-world scenarios. User-generated content (UGC) images, for instance, often suffer from diverse and unpredictable degradation patterns, including noise, compression artifacts, motion blur, and varying lighting conditions, further complicating the super-resolution process. As a result, models optimized for artificially degraded datasets may struggle to restore authentic details when applied to in-the-wild data.

Second, while diffusion models excel at generating plausible high-frequency details, they often struggle to maintain a delicate balance between degradation removal and semantic fidelity. In many cases, the model may introduce unnatural artifacts or fail to preserve the structural coherence of objects, leading to inconsistencies in the restored images. This issue is particularly pronounced in cases where extreme degradation has erased fine-grained textures, making it challenging for the model to reconstruct missing information in a visually meaningful way.

To address these challenges, we propose a diffusion-based super-resolution (SR) framework that integrates realistic degradation modeling, semantic-aware refinement, and perceptual optimization. Our method aims to bridge the gap between synthetic and real-world wild degradation scenarios by constructing a more representative training dataset and incorporating advanced architectural components to enhance restoration quality. Specifically, we introduce a training strategy that combines the synthetic training set provided by the competition with the LSDIR dataset [17]. To better simulate real-world conditions, we further process the LSDIR dataset by applying controlled degradation, ensuring that the model learns to handle diverse degradation patterns effectively.

Our approach leverages the strong generative prior of Stable Diffusion [27], which provides a powerful latent space for high-quality image synthesis. To further enhance reconstruction fidelity, we integrate ControlNet [26],

which enables precise spatial conditioning and improves structural detail preservation. Additionally, we incorporate a semantic-aware module to refine both structural and contextual information, ensuring that the generated high-resolution images maintain both perceptual quality and semantic coherence. Specifically, we leverage the semantic extraction capabilities of SAM2 [25] to recover fine-grained details by embedding high-level semantic information into the latent space. This integration allows our framework to strike a balance between effective degradation removal and realistic detail generation, making it more robust to diverse real-world degradations while producing visually coherent and semantically meaningful reconstructions.

Our contributions are as follows.

- To address the inconsistency between degradations in wild and synthetic datasets, we simulate both synthetic and wild degradation processes on LSDIR separately. The resulting datasets are then combined with the officially provided paired training set to form the final training dataset.
- To enhance degradation removal and semantic detail generation, we incorporate the pretrained semantic extraction model SAM, which helps refine the structural and contextual understanding of the input data.
- To better control the visual quality of the generated results, we fine-tune specific hyperparameters of the model, optimizing its performance for improved perceptual fidelity.

2. Related Works

In recent years, the advancement of deep learning technology has significantly propelled progress in Single Image Super-Resolution (SISR). Based on a variety of deep learning networks, the capability of SISR models to improve the quality of low-resolution images has been greatly enhanced. The commonly employed architectures in this domain includes Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GAN), Transformer, and Diffusion model [46].

2.1. CNN-Based SISR

CNN is a subclass of feedforward neural networks that incorporate convolutional operation. Its characteristics of local connectivity and weight sharing endow it with powerful feature-learning capabilities.

Notable examples of CNN-based single-image super-resolution (SISR) models include SRCNN [4] (the first deep learning-based SISR model), BSRN [18] (won the first place in the NTIRE 2022 Efficient SR (ER) track), and CVA-Net [43]. An important research direction for CNN-based SISR models is enhancing network depth and width to improve the model’s ability to capture fine image details.

2.2. GAN-Based SISR

Generative Adversarial Networks (GANs) consist of a generator and a discriminator. Through adversarial training, GANs effectively learn the underlying distribution of images, e.g., LDL [19] and RealSR [8], are GAN-based SISR models. GAN-based SISR models have demonstrated significant advantages in terms of perceptual quality and realistic visual effects. However, the large number of network parameters leads to unstable training and relatively long inference time. Enhancing reconstruction capabilities of GAN-based models while designing lightweight networks to stabilize the training process has become a key research direction.

2.3. Transformer-Based SISR

The Transformer architecture, originally designed for natural language processing, is a neural network framework that relies entirely on self-attention mechanisms to capture global dependencies between inputs and outputs. Compared to CNN and RNN, self-attention enables better modeling of long-range relationships while mitigating the vanishing gradient problem. Additionally, Transformers support parallel computation, significantly improving network efficiency.

A Transformer network consists of an encoder-decoder structure, with multi-head self-attention as its core component. Models such as GRL [16], HIPA [2], and ESRT [22] are based on Transformer architectures. Despite their advantages, Transformer-based SISR models still face challenges such as high computational costs and artifacts that affect perceptual quality. Addressing these issues remains an important research direction for improving the application of Transformers in image super-resolution.

2.4. Diffusion-Based SISR

In recent years, diffusion models have emerged as a powerful approach based on the principle of destruction and reconstruction, incorporating both forward and backward diffusion processes. Notable diffusion-based SISR models include IDM [5], DiffTIR [35], and SeeSR [34], etc.

Most existing diffusion-based SISR models adopt the UNet architecture from denoising diffusion probabilistic models (DDPMs) [7], while a few employ entirely novel structures, such as Transformer-integrated diffusion models. Designing a unified and foundational diffusion model architecture is a promising research direction in the SISR field.

However, the introduction of stochastic noise in the backward diffusion process often results in unstable outputs. Therefore, developing a stable and efficient diffusion-based SISR model is a forward-looking research challenge, aiming to balance reconstruction quality with computational efficiency.

3. Method

3.1. Preliminaries

Diffusion Model Diffusion models [7, 28] are a class of generative models that achieve the goal of generating target data samples from Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, 1)$. Diffusion models are based on a two-step process: a forward diffusion process and a reverse denoising process.

In the forward diffusion process, a data sample \mathbf{x}_0 is gradually perturbed by adding Gaussian noise over T timesteps:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

where β_t is a variance schedule controlling the noise level at each timestep, \mathbf{x}_t is the noised image at timestep t .

In the reverse process, a neural network $\epsilon_\theta(\mathbf{x}_t, t)$ is trained to predict the added noise, enabling the generation of new samples by denoising a random Gaussian sample \mathbf{x}_T :

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \quad (2)$$

where $\mu_\theta(\mathbf{x}_t, t)$ is typically parameterized using the predicted noise $\epsilon_\theta(\mathbf{x}_t, t)$ and the noise schedule.

In practice, denoising models adopt a denoising network $\epsilon_\theta(\mathbf{x}_t, t)$ to estimate noise ϵ . During training, the network parameters θ of the denoising network ϵ_θ are optimized by minimizing a loss function

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, t, \epsilon} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2 \quad (3)$$

Recent advancements, such as classifier-free guidance [6], have further enhanced the capability of diffusion models, making them state-of-the-art in various generative modeling tasks.

3.2. Overview

Our proposed method is built upon the diffusion framework. The overall pipeline of our approach is illustrated in Fig. 2. Given a low-quality input image I_{LR} , we first encode it into a latent representation using VAE encoder [12]. To progressively refine the latent representation, we employ a Denoising U-Net, which iteratively removes noise through a sequence of denoising steps. Additionally, we incorporate ControlNet architecture to enforce structural consistency with the low-resolution (LR) input. ControlNet enables precise conditioning, ensuring that the super-resolved output maintains fidelity to the original image while benefiting from the generative power of diffusion models.

A key challenge in real-world super-resolution tasks is handling images with varying degrees of degradation and different scaling factors, particularly when dealing with both synthetic and wild test sets. To address this, we integrate the Segment Anything Model 2 (SAM2) [25] into

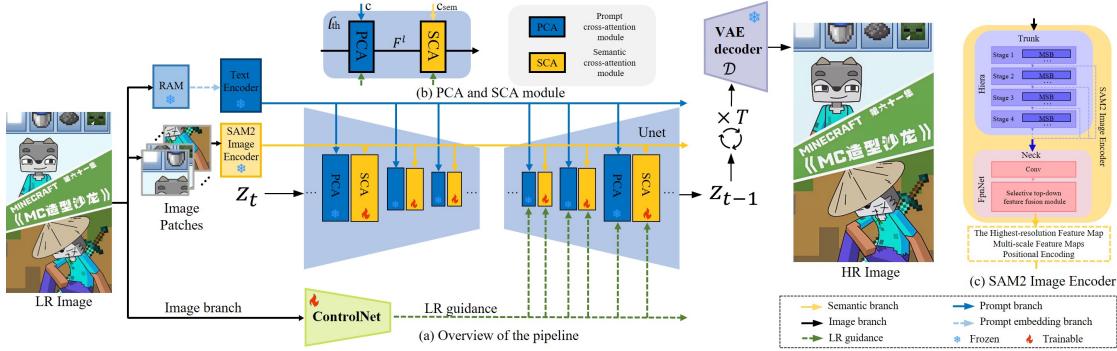


Figure 2. (a) Overview of our proposed method. Our approach builds upon the diffusion framework, incorporating a mechanism to enforce structural consistency and preserve fidelity to the original LR image. Additionally, we leverage SAM2 for semantic-guided refinement, extracting high-level semantic embeddings to enhance adaptability to diverse degradation conditions. (b) Architecture of the PCA and SCA module. (c) Architecture of the SAM2 image encoder. The encoder comprises a trunk and a neck. The trunk extracts multi-scale features from the low-resolution image through four stages with varying numbers of MSB(Multi-Scale Block) layers. The neck applies convolutions at all scales and performs top-down feature fusion on low-resolution features. The output includes refined multi-scale feature maps and corresponding positional encodings.

our framework. SAM2 is utilized to extract high-level semantic embeddings from the input images, providing additional contextual information that aids in the denoising and reconstruction process. These semantic embeddings help our model adapt to diverse image conditions, improving robustness across different datasets. The enriched latent representation, augmented with semantic information, is then iteratively refined by the Denoising U-Net over T denoising steps.

During training, we minimize a denoising objective:

$$\mathcal{L} = \mathbb{E}_{x_0, x_{LR}, t, c, c_{sem}, \epsilon} \|\epsilon - \epsilon_\theta(x_t, x_{LR}, t, c, c_{sem})\|_2^2 \quad (4)$$

where x_{LR} represents the low-resolution (LR) latent, c denotes the tag prompt, and c_{sem} is the semantic embedding. The noise estimation network ϵ_θ is responsible for predicting the noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

3.3. Semantic-Aware Module

To ensure robustness across inputs I_{LR} with different levels of degradation, as illustrated in Fig. 2(b), we introduce the Semantic-Aware Module, which extracts high-level semantic embeddings from each frame using a pre-trained SAM2 model. Specifically, as formulated in Eq. (5), the input images are processed through the frozen image semantic extractor to obtain their corresponding semantic representations c_{sem} :

$$c_{sem} = \text{Semantic Extractor}(I_{LR}) \quad (5)$$

The extracted semantic embeddings effectively capture essential structural and contextual information, even in the presence of severe degradation. These features are subsequently integrated into the denoising U-Net through a semantic attention mechanism, facilitating a more informed

restoration process. In particular, at the l -th layer of the U-Net, we compute the query vector Q from the spatial feature map F^l , while the key K and value V vectors are derived from the semantic embedding c_{sem} , as defined in Eq. (6):

$$Q = W_q(F^l), \quad K = W_k(c_{sem}), \quad V = W_v(c_{sem})$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (6)$$

By incorporating semantic priors extracted from SAM2, the Semantic-Aware Module enables the U-Net to leverage high-level contextual cues, thereby improving the reconstruction of fine details and preserving structural consistency. This integration significantly enhances the model's ability to handle complex degradations, resulting in more realistic and visually coherent super-resolved outputs.

4. Experiment

4.1. Experimental Setup

Train Dataset To enhance the model's performance in short-form user-generated content (UGC) scenarios, the training dataset is constructed by integrating the synthetic UGC training set provided by the competition with the LSDIR dataset [17]. By leveraging multiple data sources, the training set is designed to capture a diverse and realistic range of UGC image degradations.

The dataset is composed of three complementary components. First, in the LSDIR training set, each high-resolution image undergoes either downsampling or a degradation process with an equal probability of 50%, ensuring a balanced distribution of degradation patterns. Second, the synthetic UGC training dataset provides additional high-resolution

(HR) and low-resolution (LR) training pairs, where existing paired images are cropped into overlapping 512×512 sub-images to enhance sample diversity while preserving local structural details. These cropped patches retain their original HR-LR correspondence, ensuring consistency within the dataset. Finally, to further increase degradation diversity, high-resolution images from the synthetic UGC dataset are first cropped into 512×512 patches and subsequently subjected to a dedicated degradation process, simulating wild UGC distortions.

Validation Dataset and Test Dataset During the validation and testing phase, the evaluation is conducted using the official validation and test sets provided by the NTIRE 2025 Challenge on Short-form UGC Image Super-Resolution ($4\times$), ensuring a comprehensive assessment of its generalization ability in synthetic and real-world UGC scenarios.

Both the validation and test sets consist of two subsets: a synthetic dataset and a wild dataset. The synthetic dataset comprises 180 paired low-resolution (LR) and high-resolution (HR) images, where LR images are generated through $4\times$ downsampling followed by degradation, simulating distortions commonly found in real-world UGC content. Model performance on the synthetic dataset is assessed using PSNR, SSIM, and LPIPS [42]. The wild dataset consists of 190 unpaired real-world UGC images collected from short-video platforms. These images exhibit authentic distortions but lack corresponding HR references, as they undergo degradation without downsampling. We employ no-reference metrics, including MUSIQ [9], ManIQA [36], CLIPQA [31], NRQM [23], HyperIQA [29], to evaluate model’s restoration capability in this setting. Additionally, we utilize the DIV2K [1] validation dataset and apply the same degradation pipeline that was used during training to generate LR images.

Training details During training, the model is optimized using the integrated training dataset. To balance learning effectiveness with computational efficiency, training is conducted for 90,000 steps on an Nvidia RTX 3090 GPU. The Adam optimizer is utilized with a learning rate of 5×10^{-5} . To preserve the integrity of the pre-trained Stable Diffusion [27] parameters, they remain frozen throughout training. Optimization is applied exclusively to the ControlNet component and the semantic-aware module. The training strategy enables the model to adapt effectively to short-form UGC images while preserving the pretrained features of the diffusion backbone.

Testing details During testing, we analyze three key parameters: start point, guidance scale, and positive or negative prompts to assess their impact on the model’s performance across synthetic and wild datasets. The goal is to determine the optimal configuration for the final test settings.

First, we compare two initialization methods: Gaussian noise and noised low-resolution latent, assessing their im-

pact on super-resolved image quality. Next, we examine the effect of positive and negative prompts. Additional positive prompts such as clean, high-resolution, 8K, ultra-detailed, ultra-realistic are introduced to enhance image sharpness and fidelity. Conversely, negative prompts including dotted, noise, blur, low-resolution, smooth, unrealistic physics, unnatural shadows are tested to suppress artifacts and improve overall coherence. Finally, we analyze the impact of varying the guidance scale on the result. Multiple values are tested, and the resulting images are evaluated to determine the optimal balance between fidelity to the input and adherence to the learned priors of the diffusion model.

Based on the findings from these experiments, we identify the best-performing combination of parameters (as shown in Tab. 1), which is then adopted for the final test settings to ensure the optimal performance.

Datasets	Start point	gs	Text prompt
synthetic	lr	0.9	w/o prompt
wild	noise	8.5	w positive prompt

Table 1. Hyperparameter selection for testing and validation. “gs” is the abbreviation of guidance scale.

4.2. Quantitative Results

We compare our proposed method with several state-of-the-art SISR methods, including GAN-based methods BSRGAN [41], Real-ESRGAN [33] and diffusion-based methods Faithdiff [3], InvSR [40], XPSR [24], PiSA-SR [30], SeeSR [34]. The qualitative results are shown in Tab. 2.

On the synthetic validation dataset, although our model’s performance is not outstanding in any particular metric, it consistently ranks at a mid-to-high level across all metrics, and overall, it performs well among these models.

On the wild validation dataset, our method consistently achieves the best performance across all metrics, surpassing previous SOTA methods and demonstrating its robustness in handling real-world degradations.

On the DIV2K dataset, we compare our proposed method with other models using reference-based metrics. Our model ranks third in PSNR and second in SSIM. Although our model performs worse than REAL in objective metrics, it achieves the best performance in subjective evaluation.

This highlights the effectiveness of our semantic-guided diffusion framework in reconstructing fine details while preserving both structural integrity and perceptual consistency. Furthermore, these results validate the capability of our approach to bridge the gap between synthetic and real-world super-resolution scenarios, making it a promising solution for practical UGC image enhancement.

Dataset	Metrics	BSRGAN[41]	Real-ESRGAN[33]	FaithDiff[3]	InvSR[40]	XPSR[24]	PiSA-SR[30]	SeeSR[34]	Ours
synthetic	MuSIQ↑	70.5933	67.8211	72.5569	73.2262	69.8961	74.4527	73.4914	69.7667
	ManIQA↑	0.4090	0.4003	0.4344	0.4655	0.5266	0.5142	0.5170	0.4523
	CLIPQA↑	0.6191	0.5781	0.6618	0.7287	0.7449	0.7511	0.7295	0.6413
	NRQM↑	6.3076	6.3502	6.6970	6.6888	6.4588	6.6513	6.4621	6.1655
	HyperIQA↑	0.5948	0.5599	0.6007	0.6299	0.6375	0.6649	0.6593	0.6105
wild	MuSIQ↑	61.8400	63.1264	66.1299	54.9506	62.4038	70.3856	68.1723	71.0000
	ManIQA↑	0.3811	0.4794	0.3947	0.3639	0.4321	0.4980	0.4831	0.5529
	CLIPQA↑	0.5097	0.6553	0.6072	0.4897	0.6279	0.7275	0.7024	0.7633
	NRQM↑	6.2272	6.4495	6.8256	5.3240	6.4639	6.7516	6.6068	6.9242
	HyperIQA↑	0.5476	0.5901	0.5519	0.4967	0.5568	0.6391	0.6237	0.6826
DIV2K	PSNR↑	23.3098	23.3343	22.5975	21.6333	21.4079	22.9385	22.7672	23.2955
	SSIM↑	0.6237	0.6405	0.5931	0.5857	0.5526	0.6144	0.6044	0.6240

Table 2. Comparison of different SISR methods on synthetic, wild and DIV2K validation dataset. Red and blue represent the best and second score, respectively.

4.3. Qualitative Results

The images before and after the restoration process are presented in Fig. 3. As illustrated in Fig. 3, the proposed model demonstrates a strong capability in reconstructing fine details within low-resolution images. Furthermore, the restored images exhibit improved visual clarity and structural integrity, contributing to a more realistic and perceptually pleasing outcome. This suggests that the model not only enhances objective image quality but also maintains high subjective fidelity, making it suitable for applications requiring both detail preservation and natural visual appearance.

We also compared the perceived realism of the proposed method with existing state-of-the-art methods on wild and synthetic data (as illustrated in Figures Fig. 4). The images produced by our proposed method exhibit a higher degree of naturalness and realism compared to those generated by other approaches based on Generative Adversarial Networks (BSRGAN [41], Real-ESRGAN [33]) or Diffusion models (FaithDiff [3], InvSR [40], XPSR [24], PiSA-SR [30], SeeSR [34]). This demonstrates the effectiveness of our method in synthesizing visually coherent and perceptually convincing results.

4.4. Ablation Study

Text Prompt To evaluate the impact of text prompts on model performance, we conducted ablation experiments by introducing additional positive and negative prompts during inference. Results across various guidance scale settings indicate that both positive and negative prompts contribute to the final output in a comparable manner.

As shown in Tab. 3, for the wild dataset, incorporating additional positive prompts yields the highest scores, suggesting an improvement in model performance. In contrast, for the synthetic dataset, prompts have minimal effect on enhancing results.

Based on these observations, our inference strategy is as follows: for the wild dataset, we include additional positive prompts to refine outputs, whereas for the synthetic dataset, no additional prompts are introduced, as they do not contribute to further improvements.

Dataset	Text prompt	Score↑
wild	w/o prompt	20.2182
	w positive prompt	20.5054
	w negative prompt	19.7869
synthetic	w/o prompt	33.2336
	w positive prompt	33.2312
	w negative prompt	33.2336

Table 3. Impact of additional text prompts on model performance across both datasets. Positive prompt includes "clean, high-resolution, 8k, ultra-detailed, ultra-realistic". Negative prompt includes "dotted, noise, blur, lowres, smooth, unrealistic physics, unnatural shadows". For wild dataset, score is computed as: Score = $0.1 \times \text{MUSIQ} + 10 \times \text{ManIQA} + 10 \times \text{CLIPQA}$. For synthetic dataset, score is computed as: Score = PSNR + $10 \times \text{SSIM} - 10 \times \text{LPIPS}$

Guidance Scale We evaluated the final scores of the generated results across different guidance scale values, as shown in Tab. 4. For the synthetic dataset, the highest score is achieved when the guidance scale is set to 0.9, indicating that this setting optimally enhances performance in this scenario. In contrast, for the wild dataset, there is a positive correlation between the guidance scale and the final evaluation score.

To further analyze the impact of guidance scale on image perceptual quality, we examine the results for the wild test set at values of 8.5, 10, 12, and 14, and for the synthetic test set at values of 0.88, 0.9, 1 and 1.1, as illustrated in Fig. 5. For the synthetic dataset, when the guidance scale is set



Figure 3. Comparison of images from the wild and synthetic datasets before and after $1\times$ and $4\times$ super-resolution processing with our model.

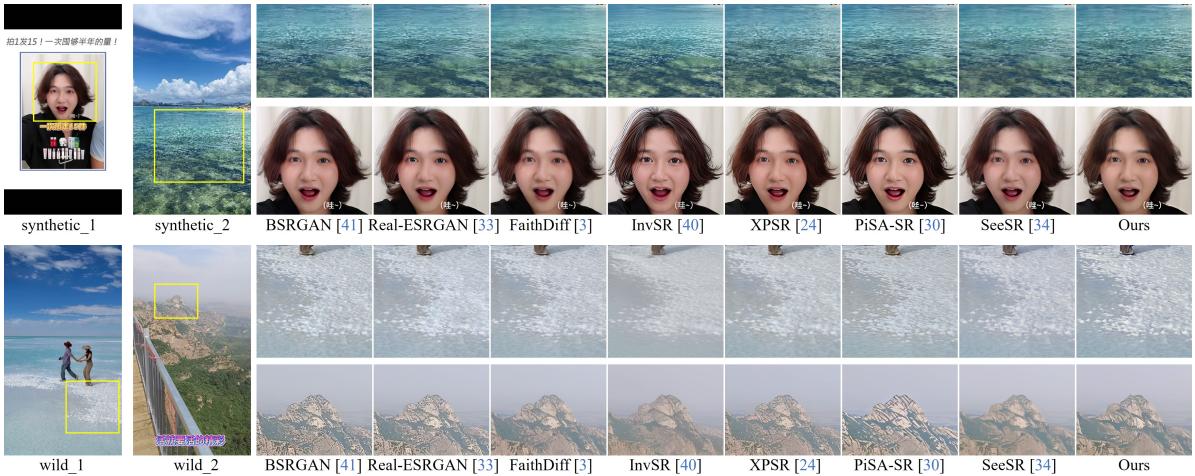


Figure 4. Comparison of images with different models on synthetic and wild validation dataset.

Dataset	gs	Score↑	Dataset	gs	Score↑
wild	8.5	20.2305	synthetic	1.1	33.2196
	10	20.3897		1.0	33.2335
	12	20.4619		0.9	33.2340
	14	20.5054		0.8	33.2338

Table 4. Impact of guidance scale on wild and synthetic dataset. "gs" is the abbreviation of guidance scale

around 0.9, the generated images exhibit minimal perceptual differences. For the wild dataset, increasing the guidance scale generally enhances the evaluation score. However, beyond a certain threshold, the generated images exhibit noticeable artifacts, resulting in a degradation of perceptual quality. To balance evaluation metrics and perceptual quality, we set the guidance scale to 8.5 for the wild test set and 0.9 for the synthetic test set.

Semantic-Aware Module To evaluate the effectiveness of the proposed semantic-aware module, we compare it with the DAPE module used in SeeSR [34]. To ensure a fair comparison, we standardize the parameter settings for both models and conduct qualitative and quantitative analyses to

assess their performance.

For the wild dataset, we evaluate performance using the MUSIQ, MANIQA, and CLIPQA metrics. As shown in Tab. 5, our semantic-aware module outperforms DAPE across all metrics, indicating that the SAM2 [25] module enhances the model's ability to process wild images compared to the DAPE module. For the synthetic dataset, due to the lack of GT images, we focus on subjective evaluations, as shown in Fig. 6. As the example shown in Fig. 6, our model demonstrates superior capability in restoring text details and natural texture compared to SeeSR. Overall, the results suggest that our proposed semantic-aware module significantly improves performance on both the wild and synthetic datasets.

Model	MUSIQ↑	MANIQA↑	CLIPQA↑	Score↑
DAPE	70.3434	0.5332	0.7345	19.7119
Ours	71.1969	0.5532	0.7579	20.2305

Table 5. Ablation study of semantic-aware model on wild dataset.

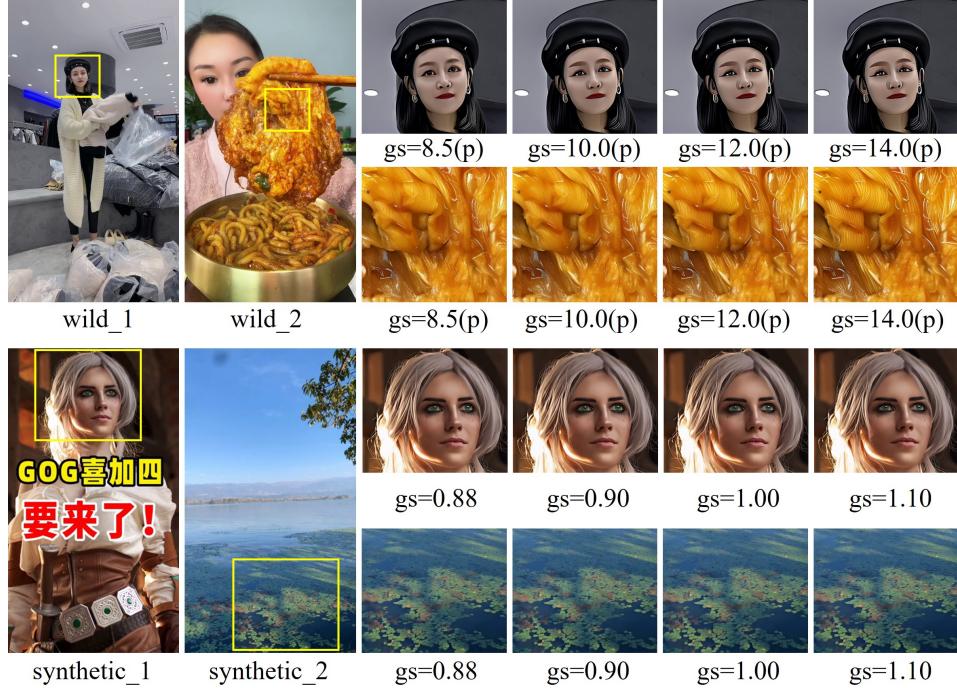


Figure 5. Image restoration results of wild and synthetic datasets under different gs values.

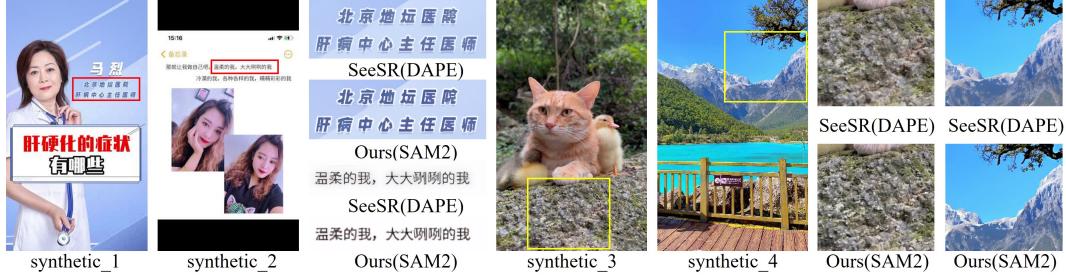


Figure 6. Comparison of the effects of the DAPE [34] and SAM2 [25] modules on synthetic dataset. The results of two models are obtained under the same parameter setting ($gs=0.9$ without extra positive prompt).

5. Conclusion

In this paper, we propose a novel approach to UGC image super-resolution by incorporating semantic guidance into a diffusion-based framework. By leveraging semantic embeddings, our method enriches the model with high-level contextual information, enabling more accurate reconstruction of fine details while mitigating artifacts. To address the disparity between synthetic and real-world degradations, we separately simulate both types on the LSDIR dataset and integrate them with the official training set, resulting in a more diverse and representative dataset. Additionally, we utilize the pretrained SAM2 model to extract refined structural and semantic features, further enhancing the quality of super-resolved images. Through meticulous selection of key hyperparameters, we optimize the model for perceptual

fidelity, ensuring visually coherent and high-quality results. Our approach effectively bridges the gap between synthetic and real-world super-resolution, contributing to more robust, perceptually faithful, and practically applicable image enhancement.

Despite the promising results, our method still presents several limitations. First, it exhibits notable difficulties in reconstructing text regions under conditions of severe degradation or occlusion. Secondly, in regions with substantial corruption or low resolution, our method may inadvertently introduce excessive high-frequency components, potentially leading to a degradation in the subjective perceptual quality. While the current approach has certain constraints, it provides a strong foundation for future extensions in this domain.

Acknowledgements

This work was partly supported by Science and Technology Commission of Shanghai Municipality (No. 24511106200), the Shanghai Key Laboratory of Digital Media Processing and Transmission under Grant 22DZ2229005, 111 project BP0719010.

References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5
- [2] Qing Cai, Yiming Qian, Jinxing Li, Jun Lyu, Yee-Hong Yang, Feng Wu, and David Zhang. Hipa: Hierarchical patch transformer for single image super resolution. *IEEE Transactions on Image Processing*, 32:3226–3237, 2023. 3
- [3] Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. *arXiv preprint arXiv:2411.18824*, 2024. 2, 5, 6
- [4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2
- [5] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023. 2, 3
- [6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3
- [8] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 466–467, 2020. 1, 3
- [9] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 5
- [10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 1
- [11] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 1
- [12] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 3
- [13] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1
- [14] Xin Li, Xijun Wang, Bingchen Li, Kun Yuan, Yizhen Shao, Suhang Yao, Ming Sun, Chao Zhou, Radu Timofte, and Zhibo Chen. NTIRE 2025 challenge on short-form ugc video quality assessment and enhancement: Kwair dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 1
- [15] Xin Li, Kun Yuan, Bingchen Li, Fengbin Guan, Yizhen Shao, Zihao Yu, Xijun Wang, Yiting Lu, Wei Luo, Suhang Yao, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2025 challenge on short-form ugc video quality assessment and enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 1
- [16] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demedolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023. 3
- [17] Yawei Li, Kai Zhang, Jingyun Liang, Jiezheng Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demedolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023. 2, 4
- [18] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–843, 2022. 2
- [19] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5657–5666, 2022. 1, 3
- [20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1
- [21] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, pages 430–448. Springer, 2024. 2
- [22] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022. 3
- [23] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for

- single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 5
- [24] Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *European Conference on Computer Vision*, pages 285–303. Springer, 2024. 2, 5, 6
- [25] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädl, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3, 7, 8
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 5
- [28] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 3
- [29] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020. 5
- [30] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. *arXiv preprint arXiv:2412.03017*, 2024. 2, 5, 6
- [31] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 5
- [32] 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 the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1
- [33] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 1, 5, 6
- [34] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024. 2, 3, 5, 6, 7, 8
- [35] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13095–13105, 2023. 2, 3
- [36] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniq: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022. 5
- [37] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 2
- [38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024.
- [39] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36:13294–13307, 2023.
- [40] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. *arXiv preprint arXiv:2412.09013*, 2024. 2, 5, 6
- [41] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4791–4800, 2021. 1, 5, 6
- [42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5
- [43] Weidong Zhang, Wenyi Zhao, Jia Li, Peixian Zhuang, Haihan Sun, Yibo Xu, and Chongyang Li. Cvanet: Cascaded visual attention network for single image super-resolution. *Neural Networks*, 170:622–634, 2024. 2
- [44] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 1
- [45] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 1
- [46] Zhicun Zhang, Yu Han, Linlin Zhu, Xiaoqi Xi, Lei Li, Mengnan Liu, Siyu Tan, and Bin Yan. Network architecture for single image super-resolution: A comprehensive review and comparison. *IET Image Processing*, 18(9):2215–2243, 2024. 2