# Semantic-Guided Diffusion Model for Single-Step Image Super-Resolution

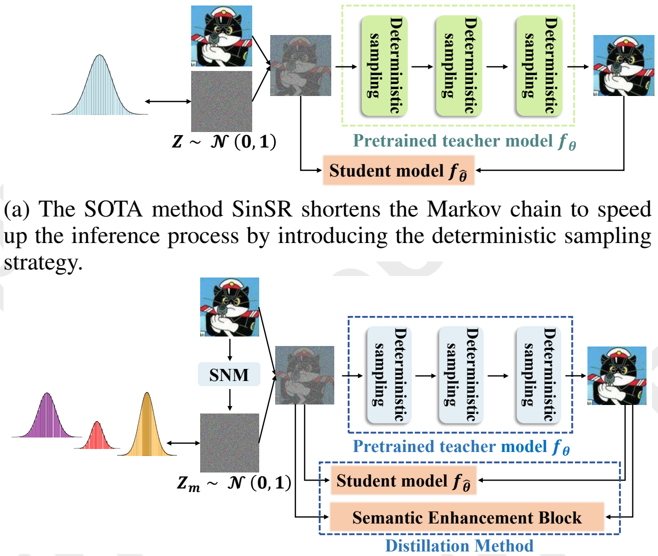
**Zihang Liu**[1], **Zhenyu Zhang**[2], **Hao Tang**[3*]
[1]Beijing Institute of Technology
[2]Nanjing University
[3]School of Computer Science, Peking University
liuzihang@bit.edu.cn, zhangjesse@foxmail.com, haotang@pku.edu.cn

## Abstract

Diffusion-based image super-resolution (SR) methods have demonstrated remarkable performance. Recent advancements have introduced deterministic sampling processes that reduce inference from 15 iterative steps to a single step, thereby significantly improving the inference speed of existing diffusion models. However, their efficiency remains limited when handling complex semantic regions due to the single-step inference. To address this limitation, we propose SAMSR, a semantic-guided diffusion framework that incorporates semantic segmentation masks into the sampling process. Specifically, we introduce the SAM-Noise Module, which refines Gaussian noise using segmentation masks to preserve spatial and semantic features. Furthermore, we develop a pixel-wise sampling strategy that dynamically adjusts the residual transfer rate and noise strength based on pixel-level semantic weights, prioritizing semantically rich regions during the diffusion process. To enhance model training, we also propose a semantic consistency loss, which aligns pixel-wise semantic weights between predictions and ground truth. Extensive experiments on both real-world and synthetic datasets demonstrate that SAMSR significantly improves perceptual quality and detail recovery, particularly in semantically complex images.

## 1 Introduction

Image super-resolution (SR) is a fundamental yet challenging problem in low-level vision, aiming to reconstruct a high-resolution (HR) image from a given low-resolution (LR) input [Wang *et al.*, 2020]. The task is inherently ill-posed due to the complexity and unknown nature of degradation models in real-world scenarios. Recently, diffusion models, as an emerging generative paradigm, have achieved unprecedented success in image generation and demonstrated remarkable potential in various low-level vision tasks, including image editing, inpainting, and colorization.


Deterministic sampling
Deterministic sampling
Deterministic sampling
$Z \sim \mathcal{N}(0,1)$
Pretrained teacher model $f_\theta$
Student model $f_{\hat{\theta}}$

(a) The SOTA method SinSR shortens the Markov chain to speed up the inference process by introducing the deterministic sampling strategy.

SNM
Deterministic sampling
Deterministic sampling
Deterministic sampling
$Z_m \sim \mathcal{N}(0,1)$
Pretrained teacher model $f_\theta$
Student model $f_{\hat{\theta}}$
Semantic Enhancement Block
Distillation Method

(b) A simplified pipeline of the proposed method SAMSR. It refines Gaussian noise and distillation method using segmentation masks to preserve spatial and semantic features.

Figure 1: A comparison between the most recent SOTA one-step SR model and our SAMSR model. Different from recent works with simple noise distribution, the proposed method incorporates semantic segmentation information into the noise distribution and gause diffusion process.

Currently, strategies for employing diffusion models in image SR can be broadly categorized into two approaches: (i) inserting the LR image as input to the denoiser and retraining the model from scratch, [Rombach *et al.*, 2022; Saharia *et al.*, 2022] and (ii) utilizing an unconditional pre-trained diffusion model as a prior and modifying its reverse path to generate the desired HR image [Choi *et al.*, 2021; Chung *et al.*, 2022; Wang *et al.*, 2021]. However, both strategies face significant computational efficiency challenges. Conventional methods typically initiate from pure Gaussian noise, failing to leverage the prior knowledge embedded in the LR image, consequently requiring a substantial number of inference steps to achieve satisfactory performance and severely constraining the practical application of diffusion-based SR techniques.

[*] Hao Tang is the corresponding author. This work was done while Zihang Liu was visiting Peking University. Our code is released at https://github.com/Liu-Zihang/SAMSR.

Although various acceleration techniques have been proposed to accelerate diffusion model sampling [Lu *et al.*, 2022; Lugmayr *et al.*, 2020; Song *et al.*, 2020], these methods often compromise performance in low-level vision domains that demand high fidelity. Recent innovative research has begun to reformulate the diffusion process in image restoration tasks, attempting to model the initial step as a combination of LR images and random noise [Yue *et al.*, 2024]. However, the inference speed remains limited. Some subsequent works have explored deterministic sampling strategies for image SR, learning bidirectional deterministic mappings between noise and HR image generation to improve inference speed [Wang *et al.*, 2024], which is shown in Fig. 1a. However, these models frequently suffer from limited authenticity and reduced capability in processing complex semantic images due to constrained inference steps.

To address these challenges, we propose a semantic segmentation-based pixel-wise sampling framework, SAMSR as shown in Fig. 1b, a semantic segmentation guided framework to address the limitations of deterministic sampling in diffusion-based image SR. Existing methods often apply uniform noise addition and global parameters, making it challenging to recover fine details in semantically complex regions. To overcome this, we introduce the SAM-Noise Module, which leverages segmentation masks generated by the Segment Anything Model (SAM) to perform spatially adaptive noise sampling, preserving both semantic and spatial features. Additionally, we propose a semantic-guided forward process that dynamically adjusts the residual transfer rate and noise strength at the pixel level based on semantic weights, enabling prioritized recovery of semantically rich regions. To enhance training, a semantic consistency loss is introduced to align the pixel-wise semantic weights between the prediction and the ground truth. These innovations allow SAMSR to effectively utilize semantic information, achieving superior performance in both real-world and synthetic datasets, particularly in recovering fine details and textures.

Our main contributions are summarized as follows: (i) We introduce, for the first time, a segmentation-mask-based random noise sampling method. This approach performs single-distribution noise sampling separately within each masked region and normalizes and combines them, allowing the Gaussian noise to retain both the spatial and semantic features of the original image. (ii) We leverage the segmentation masks obtained from SAM to derive pixel-level sampling hyperparameters, differentiating the noise addition speed for pixels with varying levels of semantic richness. This ensures that semantically rich regions are distinctly recovered within a single sampling step. (iii) We propose a novel consistency semantic loss that utilizes ground-truth images during training to enhance the model's understanding and application of region segmentation masks, leading to improved performance.

## 2 Related Work

**Advances in Super-Resolution Techniques.** Super-resolution (SR) has undergone significant evolution, transitioning from early methods based on handcrafted priors to deep learning-based approaches. Early SR algorithms leveraged priors such as non-local similarity [Sun *et al.*, 2008], sparse coding [Yang *et al.*, 2010; Cai *et al.*, 2019a], and low-rankness [Milanfar, 2012; Cai *et al.*, 2019a]. These handcrafted approaches were effective for basic degradations but lacked flexibility and generalization for complex real-world scenarios [Zhang *et al.*, 2021].

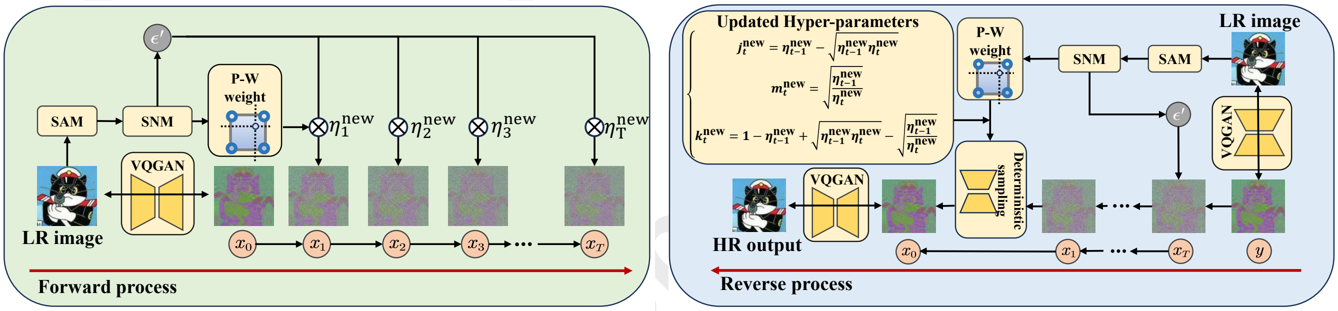
Deep learning revolutionized SR with the introduction of convolutional neural networks (CNNs) in SRCNN [Dong *et al.*, 2014], which marked the beginning of a series of innovations. Residual learning [Zhang *et al.*, 2018b], attention mechanisms and transformers further improved SR performance in terms of fidelity and perceptual quality [Cao *et al.*, 2023]. Generative adversarial networks (GANs) like ESRGAN [Wang *et al.*, 2018] pushed SR towards generating more realistic textures but often introduced perceptual artifacts and training instability [Ledig *et al.*, 2017].

Recently, diffusion models, initially developed for generative tasks [Saharia *et al.*, 2022], have emerged as promising tools for SR. Unlike GANs, diffusion models iteratively refine Gaussian noise into structured HR images, offering superior theoretical guarantees and perceptual quality. Methods like SR3 [Saharia *et al.*, 2022] and SRDiff [Song *et al.*, 2020] adapt diffusion models for SR by modifying their reverse processes or combining LR images with noise. However, these methods are often computationally expensive, requiring hundreds or thousands of iterative steps [Ho *et al.*, 2020; Lu *et al.*, 2022; Lugmayr *et al.*, 2020].

**Acceleration of Diffusion Models for SR.** The inefficiency of diffusion models has motivated extensive research into acceleration techniques. DDIM introduced deterministic sampling paths, which reduced inference steps but often compromised image fidelity in SR tasks [Ho *et al.*, 2020; Lu *et al.*, 2022]. Progressive sampling methods, like those used in Latent Diffusion Models (LDM), offered more efficient sampling but still required tens of steps to achieve satisfactory results [Rombach *et al.*, 2022].

To address these challenges, ResShift [Yue *et al.*, 2024] proposed embedding LR information directly into the Markov chain, significantly reducing sampling steps while preserving fidelity. Similarly, SinSR [Wang *et al.*, 2024] distilled the mapping between Gaussian noise and HR images into a lightweight student network, achieving single-step SR with up to a tenfold speedup. However, these methods face limitations in semantically complex regions, as they rely on uniform noise priors rather than adaptive spatial information [Yue *et al.*, 2024; Wang *et al.*, 2024]. Recent advancements such as HoliSDiP [Tsao *et al.*, 2024] integrate semantic segmentation with diffusion frameworks, providing global and localized semantic information for better spatial fidelity. This approach demonstrates the potential to further reduce sampling complexity while enhancing image quality by leveraging holistic semantic priors [Cao *et al.*, 2023; Tsao *et al.*, 2024].

**Applications of SAM in Various Domains.** Segment Anything Model (SAM) [Kirillov *et al.*, 2023] and its successor SAM2 [Ravi *et al.*, 2024] have been adapted to diverse fields, including medical imaging, video analysis, and super-resolution. Below, we summarize its applications in these areas and their relevance to this



(a) Forward process

(b) Reverse process

Figure 2: The overall framework of the proposed SAMSR method. The SAM-Noise Module (SNM) generates semantically-guided noise maps $\epsilon'$ for the forward process while computing pixel-wise semantic weights. These weights are utilized to adaptively adjust the residual transfer rate and noise strength in both forward and reverse processes, enabling fine-grained control over semantic region reconstruction.

work. (i) SAM in Medical Image Segmentation: SAM has been widely adopted in medical imaging. Models like Med-SAM-Adapter and SAMed leverage parameter-efficient fine-tuning strategies, such as adapters and low-rank adaptations, to tailor SAM for domain-specific applications [Wu *et al.*, 2023; Zhang and Liu, 2023; Leng *et al.*, 2024]. For instance, SAM-based segmentation has been explored in retinal vessel segmentation [Zhang *et al.*, 2024] and polyp segmentation [Xiong *et al.*, 2024]. Additionally, MedSAM-2 [Zhu *et al.*, 2024] demonstrates how memory mechanisms can adapt SAM for 3D segmentation tasks, allowing it to handle unordered medical image slices. These works emphasize SAM's ability to address challenges like data scarcity and fine-grained segmentation in medical domains [Shen *et al.*, 2024; Yao *et al.*, 2023]. (ii) SAM in Video Understanding and Analysis: In video segmentation and tracking, SAM has been adapted to dynamic contexts by incorporating temporal modeling and memory management. For example, SAMURAI introduces motion-aware memory mechanisms to enhance object tracking under occlusion and rapid motion [Yang *et al.*, 2024]. Similarly, SAM2Long uses tree-based memory architectures to improve segmentation consistency across long video sequences [Ding *et al.*, 2024]. In surgical video segmentation, Surgical SAM2 achieves real-time segmentation by employing efficient frame pruning, reducing computational demands while maintaining accuracy [Liu *et al.*, 2024b]. These adaptations enable SAM to excel in spatiotemporal tasks requiring consistent object identity across frames. (iii) SAM in Image and Video SR: SAM's semantic capabilities have recently been extended to super-resolution tasks. SAM Boost utilizes semantic priors to improve alignment and fusion in video super-resolution, enabling better handling of large motions and occlusions [Lu *et al.*, 2023; Liu *et al.*, 2024a]. HoliSDiP combines SAM-derived segmentation maps with diffusion models, providing spatial guidance for improved image super-resolution in real-world scenarios [Tsao *et al.*, 2024]. By integrating SAM's zero-shot segmentation and spatial adaptability, these works demonstrate the potential of SAM in enhancing detail reconstruction and semantic consistency in SR tasks.

# 3 Methodology

## 3.1 Overview

The SinSR model and the ResShift model differ primarily in their ability to reduce the number of inference steps from 15 to a single step through a deterministic sampling strategy. In the original SinSR, the forward diffusion process begins by combining a low-resolution (LR) image $y$ with Gaussian noise $\epsilon$, scaled by a residual transfer rate $\eta_t$ and noise strength $\kappa$. This is formulated as:

$$q(x_t|x_0, y) = N(x_t; x_0 + \eta_t(y - x_0), \kappa^2\eta_t I), \quad (1)$$

while the reverse process is represented by a deterministic mapping:

$$x_{t-1} = k_t\hat{x}_0 + m_t x_t + j_t y, \quad (2)$$

where $k_t$, $m_t$, and $j_t$ are coefficients derived from $\eta_t$. Although SinSR achieves single-step sampling efficiency, its reliance on global Gaussian noise $\epsilon$ and uniform diffusion parameters $\eta_t$ and $\kappa$ limits its flexibility in handling semantically complex regions.

To address this, we propose the **SAMSR** framework, which introduces semantic guidance via the SAM-Noise Module and SAM-based Forward Process. The SAM-Noise Module refines the global Gaussian noise $\epsilon$ into a spatially adaptive noise map $\epsilon'$, as described in Sec. 3.2. The SAM-based Forward Process dynamically adjusts the residual transfer rate $\eta_t$ and noise strength $\kappa$ based on semantic weights derived from SAM masks, as detailed in Sec. 3.3. Furthermore, to enhance the model's understanding of semantic information during training, we introduce a semantic consistency loss that aligns semantic features between predictions and ground truth, which will be elaborated in Sec. 3.4.

The overall framework of SAMSR is illustrated in Fig. 2. As shown in Fig. 2(a), the forward process combines the SAM-Noise Module and pixel-wise weight computation to generate semantically-guided noise. In the reverse process (Fig. 2(b)), these semantic cues are utilized to dynamically adjust the sampling hyperparameters, enabling region-aware image reconstruction. This semantic-guided framework allows SAMSR to better preserve details in semantically rich regions while maintaining global consistency.

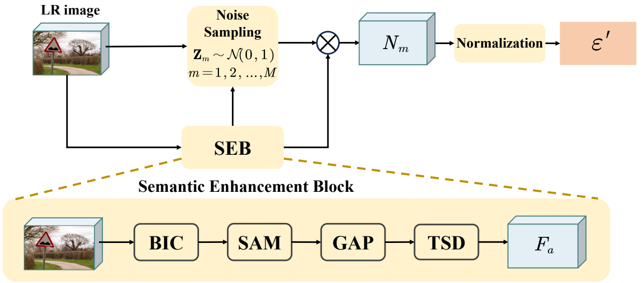


Figure 3: Architecture of the SAM-Noise Module. The module consists of two main components: (1) A Semantic Enhancement Block integrating bicubic interpolation (BIC), Segment Anything Model (SAM), global average pooling (GAP), and thresholding (TSD) operations for mask generation; (2) A noise sampling and normalization pipeline that leverages semantic information to produce spatially-adaptive noise distributions. This design enables semantically-guided noise generation while preserving structural consistency.

### 3.2 SAM-Noise Module

The SAM-Noise Module aims to integrate the spatial and semantic features of the original image into Gaussian noise, thereby enhancing the diffusion model's ability to handle images with complex semantic information. As shown in Fig. 3, to improve the accuracy of semantic segmentation, the given LR image is first passed through a bicubic interpolation process as input to the SAM. The resulting mask information is then processed through global average pooling and thresholding operations to obtain a binary tensor mask of the original LR image. The specific computation formula is as follows:

$$\begin{aligned} F_p &= BIC(LR), \quad F_p \in \mathbb{R}^{3\times 4H\times 4W}, \\ F_d &= SAM(F_p), \quad F_d \in \mathbb{R}^{M\times 4H\times 4W}, \\ F_u &= GAP(F_d), \quad F_u \in \mathbb{R}^{M\times H\times W}, \\ F_a &= TSD(F_u), \quad F_a \in \mathbb{R}^{M\times H\times W}. \end{aligned} \tag{3}$$

where $BIC$ is the bicubic interpolation, the $SAM$ is the segment anything model, and the $GAP$ is the global average pooling. To obtain the final binary mask, a thresholding operation is applied to $F_u$, where values greater than a predefined threshold $T$ are set to 1, and the rest are set to 0:

$$F_a(i,j) = \begin{cases} 1, & \text{if } F_u(i,j) > T, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

We set the threshold $T$ to $0.5$. To generate refined noise for the forward diffusion process, we incorporate the binary mask $F_a$ into the noise sampling procedure. Specifically, we first sample $M$ independent noise maps $\mathbf{Z}_m \in \mathbb{R}^{3\times H\times W}$ from a standard normal distribution $\mathcal{N}(0,1)$, where $m = 1, 2, \dots, M$. These noise maps are then multiplied by the binary mask in the element $F_a$, ensuring that the noise is applied only within the regions covered by the mask. Mathematically, the masked noise $\mathbf{N}_m$ is defined as:

$$\mathbf{N}_m = F_a \odot \mathbf{Z}_m, \quad \mathbf{Z}_m \sim \mathcal{N}(0,1),\ m = 1, 2, \dots, M, \tag{5}$$

where $\odot$ denotes the element-wise multiplication. Next, the $M$ masked noise maps are summed to produce a combined noise map $\mathbf{N}_{\text{sum}}$:

$$\mathbf{N}_{\text{sum}} = \sum_{m=1}^{M} \mathbf{N}_m. \tag{6}$$

To ensure the noise map is normalized for the diffusion process, we standardize $\mathbf{N}_{\text{sum}}$ to have zero mean and unit variance. The final noise map $\epsilon'$ is computed as follows:

$$\epsilon' = \frac{\mathbf{N}_{\text{sum}} - \mu_{\mathbf{N}_{\text{sum}}}}{\sigma_{\mathbf{N}_{\text{sum}}}}, \tag{7}$$

where $\mu_{\mathbf{N}_{\text{sum}}}$ and $\sigma_{\mathbf{N}_{\text{sum}}}$ represent the mean and standard deviation of $\mathbf{N}_{\text{sum}}$. The resulting noise $\epsilon'$ is used as the input to the forward diffusion process, ensuring that noise is spatially restricted to the mask-covered regions while maintaining a normalized distribution.

### 3.3 SAM-based Forward and Reverse Process

To refine the residual transfer rate and noise strength based on semantic regions, we introduce a pixel-wise weight matrix $W(x,y)$, which is derived from the binary masks $F_a$. Specifically, for each pixel location $(x,y)$, $W(x,y)$ is defined as the normalized coverage across all $M$ masks, where normalization is performed using the maximum pixel coverage:

$$W(x,y) = \frac{\sum_{m=1}^{M} F_a^m(x,y)}{\max_{(x',y')} \sum_{m=1}^{M} F_a^m(x',y')}, \quad W(x,y) \in [0,1], \tag{8}$$

where $F_a^m(x,y) \in \{0,1\}$ represents the binary value of the $m$-th mask at pixel location $(x,y)$, and $M$ denotes the total number of masks. The denominator represents the maximum coverage among all pixels in the image.

Using the weight matrix $W(x,y)$, the residual transfer rate $\eta_t$ and noise strength $\kappa$ are adjusted as follows:

$$\begin{aligned} \eta_t^{\text{new}}(x,y) &= \eta_t \cdot (1 + m \cdot W(x,y)), \\ \kappa^{\text{new}}(x,y) &= \kappa \cdot (1 - m \cdot W(x,y)), \end{aligned} \tag{9}$$

where $m$ is the hyper-parameter that controls the noise addition speed and intensity for pixels with different levels of semantic richness during the forward diffusion process.

Utilizing the Pixel-wise Weight Matrix, we update the forward and reverse process to introduce semantic adaptiveness into the deterministic sampling framework. Specifically, the adjustments are applied to the residual transfer rate $\eta_t$ and noise strength $\kappa$, enabling pixel-wise control based on semantic guidance.

Therefore, using the noise map $\epsilon'$, we can update the forward process of the diffusion model starting from an initial state from the LR image $y$ as below:

$$x_T = y + \kappa^{\text{new}} \sqrt{\eta_t^{\text{new}}}\epsilon'. \tag{10}$$

The updated reverse process at time step $t$ is given as:

$$x_{t-1}(x,y) = k_t^{\text{new}} \hat{x}_0(x,y) + m_t^{\text{new}} x_t(x,y) + j_t^{\text{new}} y(x,y), \tag{11}$$

**Algorithm 1** Training the Pixel-wise Sampling Framework

```
Require: Pre-trained teacher diffusion model $f_\theta$
Require: Paired training set $(X, Y)$
Init $f_{\hat{\theta}}$ from the pre-trained model, i.e., $\hat{\theta} \leftarrow \theta$
while not converged do
    Sample $x_0, y \sim (X, Y)$
    Compute $W_y(x, y)$ using Equation 8
    Compute $\kappa^{new}, \eta_T^{\text{new}}$ using Equation 9
    Compute $k_t^{\text{new}}, m_t^{\text{new}}, j_t^{\text{new}}$ using Equation 12
    Sample $\epsilon \sim \mathcal{N}(0, (\kappa^{new})^2 \, \eta_T^{\text{new}} \mathbf{I})$
    $x_T = y + \epsilon$
    for $t = T, T-1, \ldots, 1$ do
        if $t = 1$ then
            $\hat{x}_0 = f_\theta(x_1, y, 1)$
        else
            $x_{t-1} = k_t^{\text{new}} f_\theta(x_t, y, t) + m_t^{\text{new}} x_t + j_t^{\text{new}} y$
        end if
    end for
    $\mathcal{L}_{\text{distill}} = L_{\text{MSE}}(f_{\hat{\theta}}(x_T, y, T), \hat{x}_0)$
    $\mathcal{L}_{\text{inverse}} = L_{\text{MSE}}(f_{\hat{\theta}}(\hat{x}_0, y, 0), x_T)$
    $\hat{x}_T = f_{\hat{\theta}}(x_0, y, 0)$
    $\mathcal{L}_{\text{gt}} = L_{\text{MSE}}(f_{\hat{\theta}}(\text{detach}(\hat{x}_T), y, T), x_0)$
    Compute $W_{\hat{x}_0}(x, y), W_{x_0}(x, y)$ using Equation 8
    $\mathcal{L}_{\text{SC}} = L_{\text{MSE}}(W_{\hat{x}_0}(x, y), W_{x_0}(x, y))$
    $\mathcal{L} = \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{inverse}} + \mathcal{L}_{\text{gt}} + \lambda \mathcal{L}_{\text{SC}}$
    Perform a gradient descent step on $\nabla_{\hat{\theta}} \mathcal{L}$
end while
return The student model $f_{\hat{\theta}}$
```

where $k_t^{\text{new}}$, $m_t^{\text{new}}$, and $j_t^{\text{new}}$ are the updated coefficients derived from the pixel-wise adjusted residual transfer rate $\eta_t^{\text{new}}$. These parameters are defined as:

$$\begin{aligned} k_t^{\text{new}} &= 1 - \eta_{t-1}^{\text{new}} + \sqrt{\eta_{t-1}^{\text{new}} \eta_t^{\text{new}}} - \sqrt{\frac{\eta_{t-1}^{\text{new}}}{\eta_t^{\text{new}}}}, \\ m_t^{\text{new}} &= \sqrt{\frac{\eta_{t-1}^{\text{new}}}{\eta_t^{\text{new}}}}, \\ j_t^{\text{new}} &= \eta_{t-1}^{\text{new}} - \sqrt{\eta_{t-1}^{\text{new}} \eta_t^{\text{new}}}, \end{aligned} \tag{12}$$

where $\eta_t^{\text{new}}$ represents the dynamically adjusted residual transfer rate at each pixel location $(x, y)$, which is computed in Eq. (9).

### 3.4 Semantic Consistency Loss

To further incorporate semantic guidance into the training process, we introduce a Semantic Consistency Loss $L_{\text{SC}}$, which aligns the semantic weights between the predicted output $\hat{x}_0$ and the ground truth $x_0$. Using the Pixel-wise Weight Matrix $W(x, y)$ defined in Sec. 3.3, we compute the normalized semantic weights $W_{\hat{x}_0}(x, y)$ for the predicted image and $W_{x_0}(x, y)$ for the ground truth. The loss is formulated as:

$$\mathcal{L}_{\text{SC}} = L_{\text{MSE}} \left( W_{\hat{x}_0}(x, y), W_{x_0}(x, y) \right). \tag{13}$$

This additional loss is integrated into the original training objective, which consists of the distillation loss $\mathcal{L}_{\text{distill}}$, reverse loss $\mathcal{L}_{\text{reverse}}$, and ground truth loss $\mathcal{L}_{\text{gt}}$. The updated training objective is defined as:

$$\hat{\theta} = \arg\min_{\hat{\theta}} \mathbb{E}_{y, x_0, x_T} \left[ \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{reverse}} + \mathcal{L}_{\text{gt}} + \lambda \mathcal{L}_{\text{SC}} \right], \tag{14}$$

where $\lambda$ is a hyper-parameter that controls the contribution of the semantic consistency loss.

By explicitly enforcing alignment between the semantic weights of the prediction and ground truth, the proposed $L_{\text{SC}}$ improves the model's ability to utilize semantic segmentation masks effectively, leading to better semantic understanding during training. The overall of the proposed method is summarized in Algorithm 1.

## 4 Experiments

### 4.1 Experimental Setup

**Compared Methods.** We compare our method with several representative SR models, including RealSR-JPEG [Ji *et al.*, 2020], ESRGAN [Wang *et al.*, 2018], BSRGAN [Zhang *et al.*, 2021], SwinIR [Liang *et al.*, 2021], RealESRGAN [Wang *et al.*, 2021], DASR [Liang *et al.*, 2022], LDM [Rombach *et al.*, 2022], ResShift [Yue *et al.*, 2024] and SinSR [Wang *et al.*, 2024].

**Metrics.** To evaluate the fidelity of our method on synthetic datasets with reference images, we used PSNR, SSIM, and LPIPS [Zhang *et al.*, 2018a]. Additionally, two recent non-reference metrics were employed to assess the realism of the generated images: CLIPIQA [Wang *et al.*, 2023], which leverages a pretrained CLIP [Radford *et al.*, 2021] model on a large-scale dataset, and MUSIQ [Ke *et al.*, 2021].

**Training Details.** To ensure a fair comparison, we adopted the same experimental configuration and backbone architecture as described in prior work. However, our approach introduces modifications to the forward diffusion process and corresponding loss function. These adjustments enable a significant reduction in the number of training iterations compared to existing models. Specifically, our model achieves convergence in only 10,000-15,000 iterations. We attribute this improvement to the integration of a semantic consistency loss, which accelerates the convergence of the student model and further optimizes the training efficiency.

### 4.2 Experimental Results

**Evaluation on Real-world Datasets.** We comprehensively evaluate SAMSR on both real-world and synthetic datasets to demonstrate its robustness and effectiveness across diverse scenarios. For real-world evaluation, we utilize RealSR [Cai *et al.*, 2019b] and RealSet65 [Yue *et al.*, 2024]. Both datasets exhibit diverse degradation patterns and lack ground truth. SAMSR is compared against SOTA SR methods, using non-reference metrics CLIPIQA and MUSIQ [Wang *et al.*, 2023; Ke *et al.*, 2021]. As shown in Table 1, SAMSR achieves superior performance in both metrics, benefiting from its semantic-guided noise sampling and region-aware diffusion dynamics.

**Evaluation on Synthetic Datasets.** For synthetic datasets, we follow the standard pipeline to create LR inputs from 3000 HR images randomly selected from ImageNet [Wang

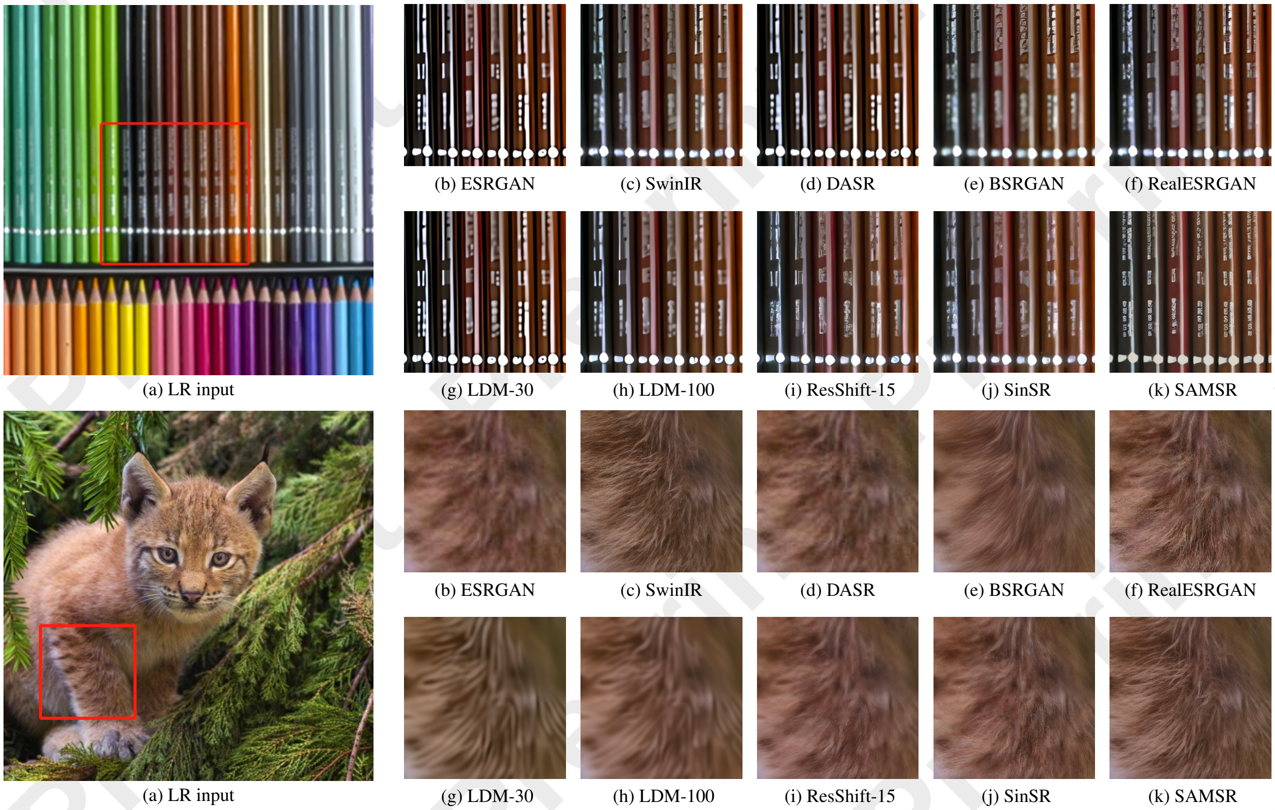


Figure 4: Qualitative comparisons on real-world examples. Please zoom in for a better view.

| Methods | RealSR | | RealSet65 | |
|---|---|---|---|---|
| | CLIPIQA↑ | MUSIQ↑ | CLIPIQA↑ | MUSIQ↑ |
| ESRGAN [Wang *et al.*, 2018] | 0.2362 | 29.048 | 0.3739 | 42.369 |
| RealSR-JPEG [Ji *et al.*, 2020] | 0.3615 | 36.076 | 0.5282 | 50.539 |
| BSRGAN [Zhang *et al.*, 2021] | 0.5439 | <u>63.586</u> | 0.6163 | **65.582** |
| SwinIR [Liang *et al.*, 2021] | 0.4654 | 59.636 | 0.5782 | 63.822 |
| RealESRGAN [Wang *et al.*, 2021] | 0.4898 | 59.678 | 0.5995 | 63.220 |
| DASR [Liang *et al.*, 2022] | 0.3629 | 45.825 | 0.4965 | 55.708 |
| LDM-15 [Rombach *et al.*, 2022] | 0.3836 | 49.317 | 0.4274 | 47.488 |
| ResShift-15 [Yue *et al.*, 2024] | 0.5958 | 59.873 | 0.6537 | 61.330 |
| SinSR-1 [Wang *et al.*, 2024] | <u>0.6887</u> | 61.582 | <u>0.7150</u> | 62.169 |
| SAMSR (Ours) | **0.7179** | **63.696** | **0.7324** | <u>65.089</u> |

Table 1: Quantitative results on two real-world datasets. The best and second best results are highlighted in **bold** and <u>underline</u>, respectively.

*et al.*, 2024]. Evaluation metrics include fidelity measures (PSNR, SSIM, LPIPS) [Zhang *et al.*, 2018a] and perceptual quality metrics (CLIPIQA, MUSIQ) [Wang *et al.*, 2023; Ke *et al.*, 2021]. Results in Table 2 indicate that SAMSR attains comparable fidelity metrics to existing diffusion-based models while significantly improving perceptual quality. The introduction of semantic masks enables pixel-wise adjustment of residual transfer rates and noise strengths, enhancing detail preservation in semantically rich regions and allowing SAMSR to outperform SinSR in balancing detail recovery and perceptual realism. These comprehensive evaluations demonstrate SAMSR's effectiveness and versatility in both real-world and controlled synthetic environments.

### 4.3 Model Analysis

**Hyper-parameter $m$.** The hyper-parameter $m$ controls the noise addition speed and intensity for pixels with different levels of semantic richness during the forward diffusion process. Table 3 summarizes the performance of SAMSR on the Realset65 and RealSR dataset under different values of $m$. We observe that both excessively large and small values of $m$ degrade the model's authenticity. Experiments show that when $m$ is within the range of [1/5, 1/8], our method achieves outstanding results on both the Realset65 and RealSR datasets. Therefore, in this paper, we set $m$ to 1/5.

**Hyper-parameters $k$ and $\eta_t$.** $\eta_t$ is the residual transition ratio defined during the diffusion process, which controls the

| Method | PSNR↑ | SSIM↑ | LPIPS↓ | CLIPIQA↑ | MUSIQ↑ |
|---|---|---|---|---|---|
| ESRGAN [Wang *et al.*, 2018] | 20.67 | 0.448 | 0.485 | 0.451 | 43.615 |
| RealSR-JPEG [Ji *et al.*, 2020] | 23.11 | 0.591 | 0.326 | 0.537 | 46.981 |
| BSRGAN [Zhang *et al.*, 2021] | 24.42 | 0.659 | 0.259 | 0.581 | **54.697** |
| SwinIR [Liang *et al.*, 2021] | 23.99 | 0.667 | 0.238 | 0.564 | 53.790 |
| RealESRGAN [Wang *et al.*, 2021] | 24.04 | 0.665 | 0.254 | 0.523 | 52.538 |
| DASR [Liang *et al.*, 2022] | 24.75 | **0.675** | 0.250 | 0.536 | 48.337 |
| LDM-30 [Rombach *et al.*, 2022] | 24.49 | 0.651 | 0.248 | 0.572 | 50.895 |
| LDM-15 [Rombach *et al.*, 2022] | 24.89 | 0.670 | 0.269 | 0.512 | 46.419 |
| ResShift-15 [Yue *et al.*, 2024] | **24.90** | 0.673 | 0.228 | 0.603 | 53.897 |
| SinSR-1 [Wang *et al.*, 2024] | 24.56 | 0.657 | 0.221 | 0.611 | 53.357 |
| SAMSR (Ours) | 24.74 | 0.666 | **0.217** | **0.619** | 54.146 |

Table 2: Quantitative results on *ImageNet-Test*. The best and second best results are highlighted in **bold** and underline, respectively.

| Hyper-parameters | | | RealSR | | RealSet65 | |
|---|---|---|---|---|---|---|
| $m$ | $p$ | $\kappa$ | CLIPIQA↑ | MUSIQ↑ | CLIPIQA↑ | MUSIQ↑ |
| 1/2 | 0.3 | 2.0 | 0.6992 | 60.483 | 0.7193 | 62.451 |
| 1/4 | 0.3 | 2.0 | 0.7019 | 61.642 | 0.7221 | 62.633 |
| **1/5** | **0.3** | **2.0** | **0.7179** | **63.696** | **0.7324** | **65.089** |
| 1/6 | 0.3 | 2.0 | 0.7092 | 62.734 | 0.7251 | 62.914 |
| 1/8 | 0.3 | 2.0 | 0.7119 | 62.385 | 0.7291 | 64.218 |
| 1/10 | 0.3 | 2.0 | 0.7069 | 61.982 | 0.7216 | 63.814 |
| 1/20 | 0.3 | 2.0 | 0.6953 | 61.492 | 0.7194 | 62.843 |

Table 3: Quantitative results of models under different Hyper-parameters ($m$, $p$, $\kappa$).

| Method | CLIPIQA↑ | MUSIQ↑ |
|---|---|---|
| **SAMSR**($\eta_t^{\mathbf{new}}, \kappa$) | 0.7208 | 63.613 |
| **SAMSR**($\eta_t, \kappa^{\mathbf{new}}$) | 0.7194 | 63.241 |
| **SAMSR**($\eta_t^{\mathbf{new}}, \kappa^{\mathbf{new}}$) | **0.7324** | **65.089** |

Table 4: A comparison of the SAMSR method with different hyper-parameters (evaluation on RealSet65 datasets).

| | Num. of Iters | Train. Time | CLIPIQA↑ | MUSIQ↑ |
|---|---|---|---|---|
| ResShift | 500k | 7.64days | 0.6537 | 61.330 |
| SinSR | 30k | 2.57days | 0.7150 | 62.169 |
| **SAMSR** | **10-15k** | **1.89days** | **0.7324** | **65.089** |

Table 5: A comparison of the training time cost and results on NVIDIA RTX4090.

gradual transition speed from the HR image to the LR image in the Markov chain. $\kappa$ is the control parameter for noise intensity during the diffusion process, affecting the strength of noise introduced at each diffusion step. In this paper, we modify $\eta_t$ and $\kappa$ using the hyper-parameter $m$, incorporating dense semantic guidance. This ensures that semantically significant regions are prioritized for finer recovery during a single step of reverse diffusion, while background regions maintain higher global consistency. Table 4 further explores the effects of individually modifying $\eta_t$ and $\kappa$.

**Effective of Pixel-wise Sampling.** Previous research has shown that learning the deterministic mapping between $x_t$ and $x_0$ is hindered by the non-causal nature of the generative process. However, our experiments demonstrate that modifying the local noise intensity across different masked regions of an image effectively mitigates this issue. This allows the student network in the knowledge distillation process to better solve the ODE process in a single step, while maintaining the same model size.

**Effective of Consistency Semantic Loss.** Our consistency semantic loss aims to improve the model's ability to understand and apply semantic information. Specifically, we establish a loss function based on the semantic weight differences between the ground-truth image and the predicted image $x_0$, enhancing the model's ability to recover semantic details. From Table 5, we observe that the consistency semantic loss not only significantly accelerates the model's convergence speed but also effectively improves its performance, demonstrating its effectiveness during the training process.

## 5 Conclusion

In this paper, we propose a semantic segmentation-guided diffusion model named SAMSR. Specifically, we introduce a pixel-wise sampling framework based on semantic segmentation, where noise is sampled within masked regions to retain the spatial and semantic characteristics of the original image. Additionally, we leverage segmentation masks to derive pixel-level sampling hyperparameters, enabling differentiated noise schedules for pixels with varying semantic richness. This ensures that semantically rich regions achieve significant recovery within a single sampling step. Furthermore, we propose a semantic consistency loss to accelerate the convergence of the model. Experimental results show that our approach achieves significant performance improvements, particularly for SR tasks on semantically complex images.

## References

[Cai *et al.*, 2019a] 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 *ICCV*, October 2019.

[Cai *et al.*, 2019b] 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 *ICCV*, pages 3086–3095, 2019.

[Cao *et al.*, 2023] Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Van Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In *CVPR*, pages 1796–1807, 2023.

[Choi *et al.*, 2021] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[Chung *et al.*, 2022] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, pages 12413–12422, 2022.

[Ding *et al.*, 2024] Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. *arXiv preprint arXiv:2410.16268*, 2024.

[Dong *et al.*, 2014] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Ji *et al.*, 2020] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPRW*, pages 466–467, 2020.

[Ke *et al.*, 2021] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023.

[Ledig *et al.*, 2017] 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 *CVPR*, pages 4681–4690, 2017.

[Leng *et al.*, 2024] Tianang Leng, Yiming Zhang, Kun Han, and Xiaohui Xie. Self-sampling meta sam: enhancing few-shot medical image segmentation with meta-learning. In *WACV*, pages 7925–7935, 2024.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021.

[Liang *et al.*, 2022] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, pages 574–591. Springer, 2022.

[Liu *et al.*, 2024a] Chuanhong Liu, Caili Guo, Yang Yang, Wanli Ni, Yanquan Zhou, Lei Li, and Tony QS Quek. Explainable semantic communication for text tasks. *IEEE Internet of Things Journal*, 2024.

[Liu *et al.*, 2024b] Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. *arXiv preprint arXiv:2408.07931*, 2024.

[Lu *et al.*, 2022] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[Lu *et al.*, 2023] Zhihe Lu, Zeyu Xiao, Jiawang Bai, Zhiwei Xiong, and Xinchao Wang. Can sam boost video super-resolution? *arXiv preprint arXiv:2305.06524*, 2023.

[Lugmayr *et al.*, 2020] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *ECCV*, pages 715–732. Springer, 2020.

[Milanfar, 2012] Peyman Milanfar. A tour of modern image filtering: New insights and methods, both practical and theoretical. *IEEE signal processing magazine*, 30(1):106–128, 2012.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

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

[Shen *et al.*, 2024] Chuyun Shen, Wenhao Li, Yuhang Shi, and Xiangfeng Wang. Interactive 3d medical image segmentation with sam 2. *arXiv preprint arXiv:2408.02635*, 2024.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Sun *et al.*, 2008] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *CVPR*, pages 1–8. IEEE, 2008.

[Tsao *et al.*, 2024] Li-Yuan Tsao, Hao-Wei Chen, Hao-Wei Chung, Deqing Sun, Chun-Yi Lee, Kelvin CK Chan, and Ming-Hsuan Yang. Holisdip: Image super-resolution via holistic semantics and diffusion prior. *arXiv preprint arXiv:2411.18662*, 2024.

[Wang *et al.*, 2018] 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 *ECCVW*, pages 0–0, 2018.

[Wang *et al.*, 2020] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020.

[Wang *et al.*, 2021] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021.

[Wang *et al.*, 2023] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, volume 37, pages 2555–2563, 2023.

[Wang *et al.*, 2024] 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 *CVPR*, pages 25796–25805, 2024.

[Wu *et al.*, 2023] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[Xiong *et al.*, 2024] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*, 2024.

[Yang *et al.*, 2010] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[Yang *et al.*, 2024] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024.

[Yao *et al.*, 2023] Xing Yao, Han Liu, Dewei Hu, Daiwei Lu, Ange Lou, Hao Li, Ruining Deng, Gabriel Arenas, Baris Oguz, Nadav Schwartz, et al. False negative/positive control for sam on noisy medical images. *arXiv preprint arXiv:2308.10382*, 2023.

[Yue *et al.*, 2024] 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, 2024.

[Zhang and Liu, 2023] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[Zhang *et al.*, 2018a] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[Zhang *et al.*, 2018b] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018.

[Zhang *et al.*, 2021] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021.

[Zhang *et al.*, 2024] Yifan Zhang, Zhuangzhuang Chen, and Xuan Yang. Light-m: An efficient lightweight medical image segmentation framework for resource-constrained iomt. *Computers in Biology and Medicine*, 170:108088, 2024.

[Zhu *et al.*, 2024] Jiayuan Zhu, Yunli Qi, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.