

SCEESR: Semantic-Control Edge Enhancement for Diffusion-Based Super-Resolution

Yun Kai Zhuang

zhuangyk2023@shanghaitech.edu.cn

ShanghaiTech University

Research Article

Keywords: Super-Resolution, Diffusion Model, ControlNet, Edge Detector

Posted Date: August 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-7913819/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SCEESR: Semantic-Control Edge Enhancement for Diffusion-Based Super-Resolution

Yun Kai Zhuang ^{1*}

^{1*}School of Information Science and Technology, ShanghaiTech University, 393 Middle Huaxia Road, Shanghai, 201210, China.

Corresponding author(s). E-mail(s): zhuangyk2023@shanghaitech.edu.cn;

Abstract

Real-world image super-resolution (Real-ISR) must handle complex degradations and inherent reconstruction ambiguities. While generative models have improved perceptual quality, a key trade-off remains with computational cost. One-step diffusion models offer speed but often produce structural inaccuracies due to distillation artifacts. To address this, we propose a novel SR framework that enhances a one-step diffusion model using a ControlNet mechanism for semantic edge guidance. This integrates edge information to provide dynamic structural control during single-pass inference. We also introduce a hybrid loss combining L2, LPIPS, and an edge-aware AME loss to optimize for pixel accuracy, perceptual quality, and geometric precision. Experiments show our method effectively improves structural integrity and realism while maintaining the efficiency of one-step generation, achieving a superior balance between output quality and inference speed.

Keywords: Super-Resolution, Diffusion Model, ControlNet, Edge Detector

1 Introduction

Real-world image super-resolution (Real-ISR) is a challenging and rapidly evolving research area aimed at reconstructing high-quality (HQ) images from their low-quality (LQ) counterparts. These LQ inputs may be captured directly in real-world conditions or synthesized using simulated degradation models that often incorporate complex and unpredictable distortions such as noise, blur, and non-ideal downsampling. A fundamental difficulty in real-world SR stems from the inherent ambiguity in the degradation process and the irreversible loss of high-frequency information, making it difficult to recover realistic texture details.

Initially, Convolutional Neural Networks (CNNs) with substantial depth were established as the backbone for learning the complex LQ-to-HQ mapping, primarily optimized with L1/L2 losses. To escape the perceptual plateau of these deterministic models, Generative Adversarial Networks (GANs) introduced a paradigm shift. The core principle involves a generator that produces the SR image and a discriminator that distinguishes it from real HQ images, which directly encourages the generation of highly realistic and

textured outputs. More recently, Diffusion Models (DMs) have emerged as a powerful alternative. Their principle is based on a two-stage process: a forward pass that progressively adds noise to a ground-truth image until it becomes pure noise, and a reverse pass where a neural network is trained to iteratively denoise a random vector, conditioned on the LQ image, to reconstruct the HQ output. This iterative refinement process allows DMs to capture a rich distribution of image details, often yielding superior diversity and fidelity in generated textures compared to GANs.

To address the significant computational bottleneck of iterative sampling in diffusion models, researchers have developed one-step generation. The core principle involves distilling the complex multi-step denoising process of a pre-trained diffusion model into a single, much larger neural network pass. Despite these advancements, one-step diffusion models still exhibit certain limitations. A primary shortcoming is their tendency to occasionally produce outputs with compromised structural integrity or subtle artifacts, as the single-pass generation may fail to fully capture complex, high-frequency details and precise edge information that are critical for perceptual

realism. This is partly because the distillation process, which compresses the multi-step denoising pipeline into one step, can lead to a loss of granular control over structural elements.

To overcome these limitations, our work introduces a key innovation by integrating a ControlNet mechanism[1] into a one-step diffusion model. This enhancement allows the model to be dynamically guided by semantic input conditions, specifically by adjusting the generative weights based on different types of edge maps, such as Canny[2] and HED (Holistically-Nested Edge Detection)[3]. The core idea is to empower the model to emphasize critical edge information during the single-pass synthesis process. The ControlNet acts as a conditional adapter, processing the input edge maps and injecting this structural guidance into the diffusion model’s sampling process. This ensures that the generated high-resolution images not only possess realistic textures but also maintain accurate and sharp structural outlines.

Furthermore, to explicitly reinforce the learning of edge details, we augment the loss function by incorporating an edge-aware AME loss term. This term specifically penalizes discrepancies between the edge maps of the super-resolved output and the ground-truth high-resolution image. By combining this AME loss with traditional L2 and LPIPS[4] losses, our hybrid loss function simultaneously optimizes for pixel-level accuracy, overall structural similarity, and geometric precision along edges. As a result, our method achieves a superior balance, enhancing the perceptual quality and structural fidelity of the generated images while maintaining the computational efficiency of a one-step generation framework.

2 Related Work

The pursuit of high-fidelity image super-resolution has led to significant methodological evolution, primarily driven by advancements in deep neural networks. This section reviews the key developments in loss functions, network architectures, and models that form the foundation for our work.

2.1 Evolution of Loss Functions for Perceptual Quality

Early deep learning-based SR models predominantly relied on pixel-wise regression losses, such as L1 or L2 (MSE) loss. While these losses are effective at minimizing the average pixel error and achieving high Peak Signal-to-Noise Ratio (PSNR), they inherently lead to perceptually unsatisfactory, over-smoothed reconstructions with a lack of high-frequency texture. To address this, the research community shifted towards

perception-driven metrics. The Structural Similarity Index (SSIM) loss was introduced to better align with human vision by measuring local patterns of luminance and contrast. A more significant leap was made with the advent of perceptual losses, most notably the Learned Perceptual Image Patch Similarity (LPIPS)[4] metric, which leverages features from pre-trained deep networks to measure semantic similarity in a high-dimensional feature space, encouraging outputs that are perceptually convincing even if pixel-level accuracy is slightly compromised.

2.2 From CNNs to Diffusion Models

The architecture of SR models has evolved in tandem with their training objectives. Pioneering works employed deep Convolutional Neural Networks (CNNs)[5–8], which have difficulties in perceptual losses.

A paradigm shift in super-resolution occurred with the introduction of Generative Adversarial Networks (GANs). Pioneering works such as SRGAN[9] demonstrated that adversarial training could yield results with highly realistic textures and significantly improved perceptual quality. Subsequent models like BSRGAN[10] and RealESRGAN[11] further advanced real-world image super-resolution (Real-ISR) by handling more complex and diverse degradations. FeMaSR[12] recovers realistic high-resolution images by matching distorted low-resolution features to pretrained HR priors in a compact feature space, avoiding GAN instability without needing explicit references. EFRG-SRGAN[13] introduces an Enhanced Feature Refinement Generator with a Multi-Information Attention mechanism, improving the realism and detail of textures in super-resolved images. RGSRGAN[14] introduces a recursive gated convolution network with global filtering and a novel artifact discriminative loss, effectively restoring high-frequency details while reducing computational costs in image super-resolution. Despite these advancements, GAN-based methods face inherent limitations. The adversarial training process is often unstable, and the discriminator’s ability to assess the quality of diverse natural image content remains limited.

More recently, Diffusion Models (DMs) have emerged as a state-of-the-art generative framework. DMs excel at capturing a rich distribution of natural image details, often yielding outputs with superior diversity and fidelity compared to GANs.

2.3 From Multi-Steps Diffusion Models to One-Step Diffusion Models

Recent studies have increasingly explored multi-step diffusion models for image super-resolution

(SR). For instance, ResShift[15] trains a diffusion model from scratch on paired low-quality and high-quality (LQ-HQ) image data. StableSR[16] introduces a trainable encoder and uses the LQ image as a conditioning input to guide a pre-trained Stable Diffusion model. PASD[17] incorporates an encoder for degradation removal and further proposes a pixel-aware cross-attention module to integrate both low-level and high-level image features into the diffusion process. SeeSR[18] improves semantic robustness by leveraging degradation-aware tag-style prompts to steer the generative process. DDiffSR[19] integrates a Transformer for compact SR priors and a residual denoising diffusion model, improving both quantitative metrics and perceptual quality. SDGDiff[20] uses a diffusion model in the spectral domain with a wavelet-based loss to enhance high-frequency details and color reproduction for superior visual perception. However, the multi-step sampling procedure in such models leads to increased computational cost and a higher risk of generating unrealistic or unfaithful image content.

To mitigate these issues, researchers have begun developing diffusion-based SR methods with reduced sampling steps. SinSR[21] applies consistency-preserving distillation to accelerate the diffusion process originally introduced in ResShift[15]. Meanwhile, OSEDiff[22] adopts the LQ image as the direct input—bypassing random noise sampling—and utilizes variational score distillation (VSD)[23, 24] loss to distill the generative capability of a multi-step diffusion model into a one-step inference framework, thereby providing an efficient DM-based SR solution.

2.4 The Role of Structural Guidance

Incorporating explicit structural guidance is crucial for enhancing the geometric fidelity of generative models. This has led to the use of various edge detection techniques to guide the image generation process. The Sobel operator, for instance, is a classic method that approximates the image gradient to highlight regions of high spatial frequency. The Laplacian of Gaussian (LoG) filter, another foundational technique, identifies edges by locating zero-crossings after applying a Laplacian filter to a Gaussian-smoothed image. More advanced methods like the Canny detector[2] provide cleaner, well-localized edges by employing non-maximum suppression and hysteresis thresholding. Similarly, Holistically-nested Edge Detection (HED)[3] is a deep learning-based approach that produces semantically rich edge maps by leveraging multi-scale features.

The advent of ControlNet[1] has provided a powerful framework for adding spatial conditioning to large-scale text-to-image multi-steps diffusion models. A prominent application[25] is the use of Canny edge detection to steer the generative process, ensuring that the output adheres to specific structural outlines.

Our work integrates a ControlNet-like guidance mechanism, conditioned on such edge information, into a one-step diffusion model. This adapter provides granular structural control, directly addressing the inaccuracies of one-step generation. Furthermore, we augment the training with an edge-aware loss to explicitly reinforce geometric detail learning, achieving a superior balance between perceptual quality and structural fidelity.

3 Methodology

Our proposed framework is designed to enhance the structural fidelity of one-step diffusion models for real-world image super-resolution. The overall architecture, illustrated in the Figure 1, integrates a gated ControlNet mechanism and a hybrid loss to guide the generation process. The following subsections detail its core components.

3.1 One-Step Diffusion Model for Super-Resolution

As discussed, existing diffusion-based Real-ISR methods typically require multiple timesteps to estimate the HQ image from random noise initialization with LQ images as conditional inputs. These approaches are computationally expensive and inherently introduce stochasticity. We propose a super-resolution network based on one-step diffusion for Real-ISR.

Our generator G_θ consists of three key components: a frozen encoder E_θ , a fine-tuned diffusion network ϵ_θ , and a frozen decoder D_θ . Let E_ϕ , ϵ_ϕ and D_ϕ denote the VAE encoder, latent diffusion network, and VAE decoder of a pretrained Stable Diffusion (SD) model, where ϕ represents the pretrained parameters. Following recent success of LoRA[26–28] in fine-tuning SD for downstream tasks, we employ LoRA to adapt the pretrained SD model for Real-ISR.

We maintain SD’s original generation capacity by introducing trainable LoRA layers to the diffusion network ϵ_ϕ , transforming them into ϵ_θ through fine-tuning with our training data. The encoder parameters and decoder parameters remain fixed ($D_\theta = D_\phi$) to ensure consistency between the diffusion network output.

Recall that the diffusion process transforms input latent z through $z_t = \alpha_t z + \beta_t \epsilon$, where α_t, β_t are timestep-dependent scalars for $t \in \{1, \dots, T\}$.

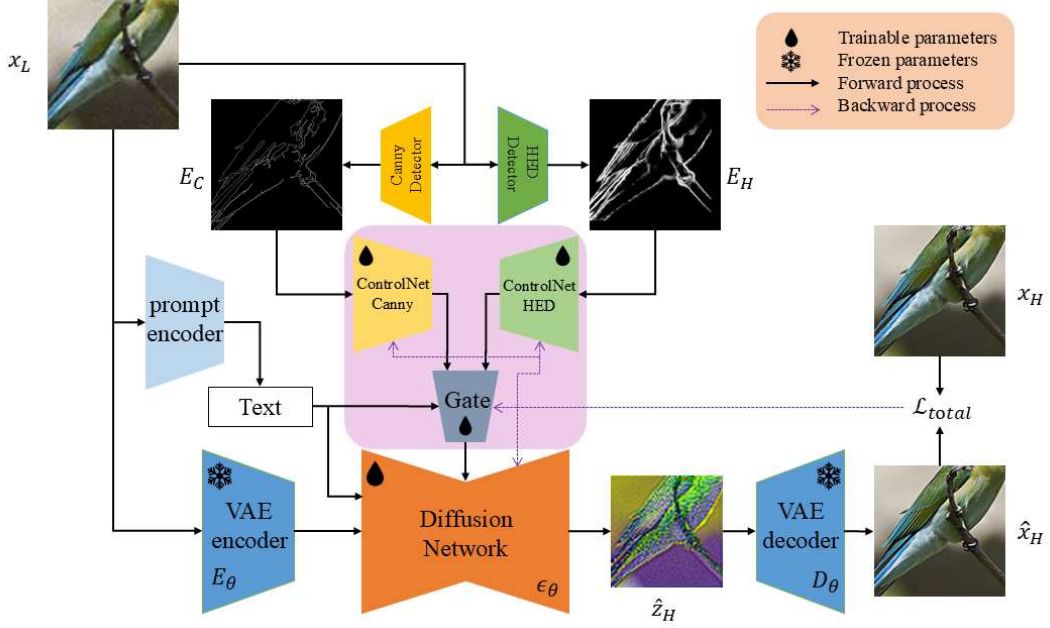


Fig. 1 The training process of SCEESR. The LQ image is passed through a frozen encoder E_θ , a LoRA finetuned diffusion network ϵ_θ and a frozen decoder D_θ . Text prompts, generated by a prompt encoder, are fed into both the diffusion network ϵ_θ and a semantic control gate. Simultaneously, Canny edge maps and HED edge maps are extracted from the LQ image using a Canny detector and an HED detector, respectively. The outputs from the two ControlNets—corresponding to the edge maps—are fused in the semantic control gate under the guidance of the text prompts. The fused control information is then injected into each layer of the diffusion network ϵ_θ . During inference, only the diffusion network ϵ_θ , the two ControlNets, and the parameters of the semantic control gate are utilized.

Given a neural network that predicts noise in z_t as $\hat{\epsilon}$, the denoised latent can be obtained as:

$$\hat{z}_0 = \frac{z_t - \beta_t \hat{\epsilon}}{\alpha_t} \quad (1)$$

which should yield cleaner and more photorealistic results than z_t . The original SD performs text-conditioned generation by extracting text embeddings c_y from text description y , enabling noise prediction as $\hat{\epsilon} = \epsilon_\theta(z_t; t, c_y)$.

We adapt this text-to-image denoising process for Real-ISR by formulating the LQ-to-HQ latent transformation F_θ as a text-conditioned image-to-image denoising process:

$$\hat{z}_H = \frac{z_L - \beta_T \epsilon_\theta(z_L; T, c_y)}{\alpha_T} \quad (2)$$

where we perform single-step denoising on the LQ latent z_L at the final timestep T without adding noise. The text embeddings $c_y = \mathcal{Y}(x_L)$ are extracted from the LQ input x_L using a prompt extractor[29]. The complete LQ-to-HQ image synthesis is:

$$\hat{x}_H = D_\theta(F_\theta(E_\theta(x_L); \mathcal{Y}(x_L))) \quad (3)$$



Fig. 2 The figure presents a comparative visualization of edge detection results from the same input image (left). The middle image shows the output of the Canny detector[2]. The right image displays results from the HED detector[3].

3.2 Structural Guidance with Canny and HED Edge Detectors

To compensate for the structural limitations of one-step generation, we employ explicit edge guidance. We extract edge maps from the upsampled LR image using two complementary detectors: Canny[2] and Holistically-Nested Edge Detection (HED)[3].

Canny Detector: The Canny operator is a classic multi-stage algorithm involving Gaussian filtering, gradient calculation, non-maximum suppression, and hysteresis thresholding. It is renowned for its strong response to prominent, sharp edges and is less sensitive to texture noise. This makes it highly effective for capturing clear structural boundaries and object contours.

HED Detector: HED is a deep learning-based approach that performs image-to-image prediction by leveraging holistically-nested networks with

deep supervision. HED excels at capturing richer edge information, including semantically meaningful soft edges and fine-grained textures that the Canny detector might omit. However, it may also be more susceptible to producing thicker or noisier edge lines.

Performance Comparison and Rationale: The choice of these two detectors is strategic. As shown in the Figure 2, the Canny detector provides precise geometric guidance for strong, unambiguous edges, ensuring structural integrity. In contrast, the HED detector provides comprehensive semantic guidance, capturing finer details and object-level boundaries that contribute to perceptual richness. By leveraging both, our model can access a more complete and multi-faceted representation of the image’s structural information.

3.3 Gated ControlNet with MLP-based Fusion

To dynamically and adaptively fuse the complementary edge features from the Canny and HED detectors, we introduce a gated ControlNet mechanism driven by a Multi-Layer Perceptron (MLP). Instead of simply concatenating the edge maps, this design allows the model to learn a content-aware fusion strategy.

The process begins by extracting structural conditions from the low-resolution input. The Canny edge map E_C and the HED edge map E_H are first processed by their respective ControlNet adapters. These adapters, which are trainable copies of specific encoder blocks from the generator G_θ , encode the edge maps into high-dimensional feature representations, denoted as $F_C = \text{ControlNet}_C(E_C)$ and $F_H = \text{ControlNet}_H(E_H)$.

The core innovation is a gating function implemented by a compact MLP, which acts as a semantic router. This MLP takes as input a globally pooled semantic feature vector z_{sem} extracted from the prompt extractor. This vector z_{sem} encapsulates the global semantic context of the input image. The MLP g_ϕ , parameterized by weights ϕ , then maps this semantic vector to a set of fusion weights:

$$[\beta_C, \beta_H] = \text{Softmax}(g_\phi(z_{sem})) \quad (4)$$

Here, β_C and β_H are scalar weights that control the global contribution of the Canny-guided and HED-guided features, respectively. The Softmax function ensures $\beta_C + \beta_H = 1$.

The final, fused edge guidance feature F_{edge} that is injected into the Unet is computed as a weighted sum of the two ControlNet outputs:

$$F_{edge} = \beta_C \cdot F_C + \beta_H \cdot F_H \quad (5)$$

This gating mechanism enables a dynamic and semantically-informed interaction between the two structural guidance streams. Based on the input image’s content, the MLP can learn to assign a higher weight β_C to the Canny stream for precise geometric boundaries or a higher weight β_H to the HED stream for richer semantic edges and textures. This adaptive fusion ensures that the most relevant structural guidance is emphasized during the single-pass synthesis, leading to outputs with enhanced structural integrity.

3.4 Hybrid Loss Function

To train the entire model end-to-end and explicitly enforce geometric accuracy, we employ a hybrid loss function $\mathcal{L}_{total}$ that combines pixel-level, perceptual, and edge-aware objectives:

$$\mathcal{L}_{total} = \lambda_{L2}\mathcal{L}_{L2} + \lambda_{lpips}\mathcal{L}_{lpips} + \lambda_{AME}\mathcal{L}_{AME} \quad (6)$$

Pixel Reconstruction Loss ($\mathcal{L}_2$): We use the L2 loss to ensure basic pixel-level fidelity between the generated image $\hat{x}_H$ and the ground-truth high-resolution image x_H . This term provides a stable training signal for the generator.

Perceptual Loss ($\mathcal{L}_{lpips}$): To improve the perceptual quality, we employ the LPIPS loss[4], which measures the similarity in the feature space of a pre-trained network. This loss encourages the output to be semantically similar to the target.

Adaptive Multi-Detector Edge Loss ($\mathcal{L}_{AME}$): is based on the Entropy Weight Method. This loss function objectively evaluates the amount of information contributed by different edge detectors for each training batch and dynamically allocates weights, thereby achieving a more scientific and adaptive edge constraint. Our approach utilizes a set of four edge detectors to form a comprehensive edge-sensing system: the Sobel operator, which provides dense gradient magnitude information and is sensitive to subtle intensity changes; the Laplacian of Gaussian (LoG) operator, which offers better capture of edge scale characteristics; the Canny detector[2], which produces fine, well-connected structural edges; and the Holistically-Nested Edge Detection (HED) detector[3], which extracts semantic-level edge information. For each detector d , we calculate its L1 loss and Structural Similarity (SSIM) loss components, defined as

$$\mathcal{L}_{L1}^d = \|E_d(x_H) - E_d(\hat{x}_H)\|_1 \quad (7)$$

and

$$\mathcal{L}_{SSIM}^d = 1 - SSIM(E_d(x_H), E_d(\hat{x}_H)) \quad (8)$$

, respectively.

The weighting process based on the Entropy Weight Method involves four key steps. First, for

a batch of N samples, a loss matrix $X \in R^{N \times 8}$ is constructed. Each column of this matrix corresponds to the values of a specific loss item across all samples in the batch. Second, each column of the loss matrix is normalized to the range $[0, 1]$ using min-max standardization. Third, the information entropy for the j -th loss item is calculated as $e_j = -\frac{1}{\ln(N)} \sum_{i=1}^N p_{ij} \ln(p_{ij})$, where $p_{ij} = \frac{r_{ij}}{\sum_{i=1}^N r_{ij}}$ is the proportion value. Finally, the weight for each loss component is determined as

$$\omega_j = \frac{1 - e_j}{\sum_{k=1}^M (1 - e_k)} \quad (9)$$

The final form of the proposed Adaptive Multi-Detector Edge Loss (AME) is defined based on the adaptively obtained weights from the Entropy Weight Method:

$$\mathcal{L}_{AME} = \sum_{d \in D} (\omega_{L1}^d \cdot \mathcal{L}_{L1}^d + \omega_{SSIM}^d \cdot \mathcal{L}_{SSIM}^d) \quad (10)$$

, where ω_{L1}^d and ω_{SSIM}^d are the weights derived from the Entropy Weight Method for the L1 loss and SSIM loss of detector d , respectively.

By jointly optimizing with this hybrid loss, our model is guided to produce images that are not only photorealistic in texture but also geometrically faithful to the ground-truth structure, thereby achieving a superior balance within an efficient one-step generation framework.

4 Experiments

4.1 Experiment Settings

Implementation Details

We train SCEESR with the AdamW optimizer at a learning rate of $5e-5$ upon SD 2.1-base for the $\times 4$ SR task. The rank of LoRA in the diffusion network is set to 4. We adopt the degradation-aware prompt extraction (DAPE)[29] module in SeeSR[18] to extract text prompts. The canny edge detector[2] applies Gaussian blur (7×7 kernel) and CLAHE with 2.0 cliplimit and uses the median pixel intensity and morphological closing (5×5 elliptical kernel). We use the deploy prototxt and caffemodel from Holistically-Nested Edge Detection[3] to implement HED detector. The network structure of the gated function MLP consists of three layers and the output size is 2×1 . The batch size is 8 and the training patch size is 512×512 . The modules on 4 NVIDIA V100 GPUs undergo 8K training iterations with no inserting of the ControlNet firstly to avoid early interference and ensure complete convergence and undergo 8K training iterations with the ControlNet on the frozen UNet. The AME loss function is used in

the second training part to enhance edges and the loss weights calculate through Entropy Weight Method based on the training datasets.

Training and Testing Datasets

We train SCEESR on a combination of the LSDIR[30] dataset and the first 10,000 face images from FFHQ[31]. Low-quality (LQ) and high-quality (HQ) training pairs are synthesized using the degradation pipeline from Real-ESRGAN[11]. For evaluation, we compare SCEE-SR against several competing methods on the test set introduced by StableSR[16], which comprises both synthetic and real-world data. The synthetic subset consists of 3,000 images of size 512×512 ; their ground-truth (GT) counterparts are randomly cropped from DIV2K-Val[32] and degraded via the RealESRGAN[11] pipeline. The real-world evaluation data include LQ-HQ pairs from DRealSR[33] and RealSR[34], with image sizes of 128×128 and 512×512 , respectively.

Compared methods

We compare SCEESR with one-step DM-based SR methods OSediff[24], multi-step DM-based SR methods ResShift[15], StableSR[16], SeeSR[18] and PASD[17] and GAN-based SR methods BSRGAN[10], RealESRGAN[11] and FeMaSR[12]. All comparative results are obtained using officially released codes or models.

Evaluation Metrics

To ensure a comprehensive and holistic evaluation of different methods, we employ a diverse set of full-reference and no-reference image quality assessment metrics. The full-reference metrics include: PSNR and SSIM, which assess pixel-wise fidelity to the ground-truth image. LPIPS[4] and DISTS[35], which measure perceptual similarity in a deep feature space. For distribution-level consistency, we use the Fréchet Inception Distance (FID)[36] to quantify the similarity between the distributions of ground-truth and restored images. Additionally, we incorporate several no-reference metrics to evaluate perceptual quality without relying on ground-truth references, including NIQE[37], MUSIQ[38], CLIPQA[39] and MANIQA[40].

4.2 Comparison with State-of-the-Arts

Quantitative Comparisons

Table 1 compares the performance of our proposed SCEESR-s1 model with various GAN-based and DM-based approaches. The following findings can be made. Traditional methods like BSRGAN[10],

Table 1 Quantitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks. The best results of each metric are highlighted in red and second best results are highlighted in blue. ‘s’ denotes the number of diffusion reverse steps in the method.

Datasets	Methods	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DIST $\downarrow$	FID $\downarrow$	NIQE $\downarrow$	MUSIQ $\uparrow$	CLIPQA $\uparrow$	MANIQA $\uparrow$
DIV2K-Val	BSRGAN	24.56	0.6264	0.3356	0.2279	44.25	4.7519	61.15	0.5242	0.5065
	RealESRGAN	24.27	0.6366	0.3117	0.2146	37.67	4.6788	61.01	0.5271	0.5496
	FeMaSR	23.05	0.5883	0.3133	0.2064	35.89	4.7415	60.79	0.5993	0.5068
	StableSR-s200	23.23	0.5718	0.3119	0.2053	24.48	4.7585	65.88	0.6764	0.6187
	ResShift-s15	24.62	0.6173	0.3354	0.2218	36.15	6.8217	61.01	0.6065	0.5448
	PASD-s20	23.11	0.5496	0.3578	0.2213	29.22	4.3621	68.91	0.6782	0.6477
	SeeSR-s50	23.65	0.6033	0.3203	0.1974	25.94	4.8107	68.62	0.6931	0.6235
	OSDiff-s1	23.68	0.6098	0.2948	0.1983	26.36	4.7101	67.93	0.6674	0.6142
	SCEESR-s1	23.78	0.6082	0.2904	0.1959	25.68	4.6687	68.93	0.6852	0.6242
DrealSR	BSRGAN	28.73	0.8025	0.2989	0.2168	155.69	6.5196	57.10	0.4911	0.4872
	RealESRGAN	28.62	0.8047	0.2965	0.2095	147.67	6.6931	54.13	0.4415	0.4901
	FeMaSR	26.89	0.7568	0.3175	0.2242	157.83	5.9077	53.67	0.5459	0.4413
	StableSR-s200	28.01	0.7531	0.3288	0.2275	149.05	6.5245	58.47	0.6351	0.5593
	ResShift-s15	28.43	0.7666	0.4013	0.2663	172.33	8.1256	50.54	0.5336	0.4575
	PASD-s20	27.32	0.7064	0.3767	0.2538	156.19	5.5478	64.82	0.6801	0.6162
	SeeSR-s50	28.14	0.7687	0.3196	0.2323	147.45	6.3978	64.86	0.6798	0.6035
	OSDiff-s1	27.89	0.7829	0.2975	0.2172	135.36	6.4912	64.58	0.6956	0.5892
	SCEESR-s1	28.13	0.7819	0.2963	0.2167	133.93	6.3457	65.41	0.6965	0.6054
RealSR	BSRGAN	26.37	0.7649	0.2795	0.2125	141.36	5.6575	63.15	0.4997	0.5394
	RealESRGAN	25.67	0.7611	0.2832	0.2098	135.23	5.8299	60.12	0.4442	0.5482
	FeMaSR	25.06	0.7352	0.2948	0.2293	141.11	5.7891	58.90	0.5264	0.4861
	StableSR-s200	24.67	0.7077	0.3025	0.2295	128.59	5.9131	65.71	0.6172	0.6215
	ResShift-s15	26.29	0.7416	0.3467	0.2507	141.77	7.2643	58.35	0.5437	0.5278
	PASD-s20	25.18	0.6792	0.3388	0.2269	124.38	5.4145	68.68	0.6613	0.6481
	SeeSR-s50	25.14	0.7211	0.3015	0.2229	125.64	5.4092	69.39	0.6604	0.6435
	OSDiff-s1	25.13	0.7336	0.2931	0.2135	123.56	5.6483	69.01	0.6687	0.6318
	SCEESR-s1	25.35	0.7386	0.2793	0.2091	123.73	5.5861	69.41	0.6698	0.6447

RealESRGAN[11] and FeMaSR[12], while achieving competitive PSNR and SSIM in some cases, often fall short in no-reference metrics such as FID[36], MUSIQ[38], and CLIPQA[39]. Among DM-based approaches, StableSR-s200[16], despite achieving the best FID[36] on DIV2K-Val, does not consistently lead in other perceptual metrics. ResShift-s15[15] shows limited performance across both reference and no-reference metrics, often ranking lower than other DM methods. PASD-s20[17] and SeeSR-s50[18] achieves improved no-reference metric scores such as MUSIQ[38], CLIPQA[39], and MANIQA[40], but their multi-step inference makes them less efficient. Moreover, their relatively higher LPIPS[4] and DIST[35] scores compared to SCEESR-s1 suggest a less faithful reconstruction to the ground truth. OSDiff-s1[24] efficiently distills multi-step generation into a single step, yielding strong LPIPS[4] and DIST[35] scores. However, its no-reference metrics, while good, are often surpassed by methods like PASD[17], SeeSR[18] and SCEESR-s1. Our proposed SCEESR-s1 distinguishes itself by maintaining high efficiency, while consistently achieving impressive perceptual quality. Notably, SCEESR-s1 excels not only in reference-based perceptual metrics such as LPIPS[4] and DIST[35] but also consistently leads in no-reference metrics such as CLIPQA[39], MUSIQ[38], and MANIQA[40], demonstrating its superior ability to generate perceptually realistic and high-fidelity images with high efficiency across all evaluated datasets.

Qualitative comparisons

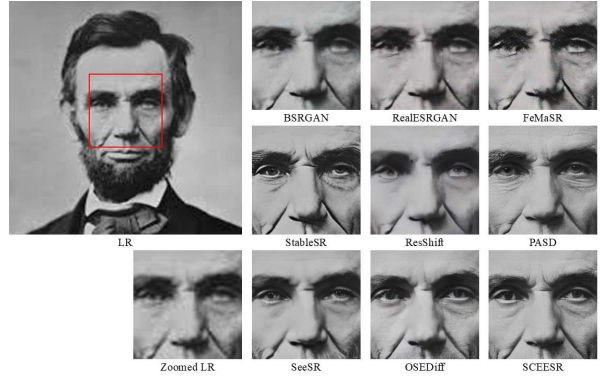


Fig. 3 Vision comparisons of different GAN-based and DM-based SR methods on Lincoln’s face.

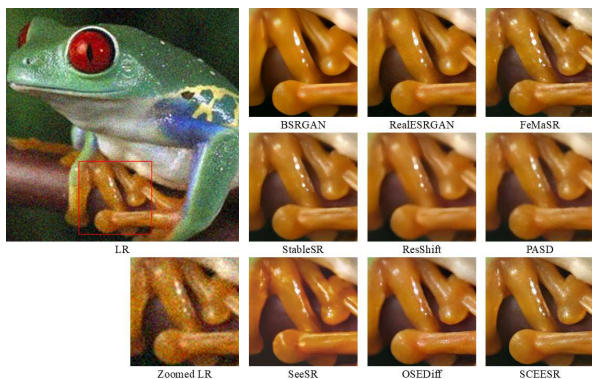
Qualitative comparisons of real-world image super-resolution methods are provided in Figure 3 and Figure 4.

The first example, the face of Lincoln, reveals that the conventional GAN methods fail to reconstruct sharp facial details. Methods built upon Stable Diffusion (SD), such as ResShift[15] and SeeSR[18] blur the facial details, especially the wrinkles on the face. StableSR[16] and PASD[17] often introduces overly exaggerated and unnatural textures. OSDiff[24] achieves realistic facial details. However, compared with OSDiff[24], our SCEESR is more powerful in restoring edge information, such as wrinkles and boundaries.

Table 2 The inference time and the number of total parameters of DM-based SR methods.

	StableSR	ResShift	PASD	SeeSR	OSEDiff	SCEESR
Inference Steps	200	15	20	20	1	1
Inference Time(s)	11.40	0.72	2.70	4.30	0.12	0.15
#Total Parm(M)	1510	179	2210	2524	1775	2028

This superiority is further confirmed in the second example, the toes of a frog. BSRGAN[10], RealESRGAN[11] and FeMaSR[12] lack of color gradation and three-dimensionality due to their frameworks. StableSR[16], lacking explicit semantic guidance and SeeSR[18], lacking extensive pre-trained image priors are constrained in generating rich and accurate textures. Although PASD[17] incorporates text prompts, its non-robust prompt extraction under degradation leads to worse semantic generation. SeeSR[18] attempts to address this with degradation-aware cues, but the resulting color appear unnatural, possibly due to instability from random noise sampling. OSEDiff[24] succeeds in producing finely detailed and natural textures, but the edges of the toes are blurry. SCEESR ensures the rationality of the colors and enhances the clarity of the edges.

**Fig. 4** Vision comparisons of different GAN-based and DM-based SR methods on a frog's toes.

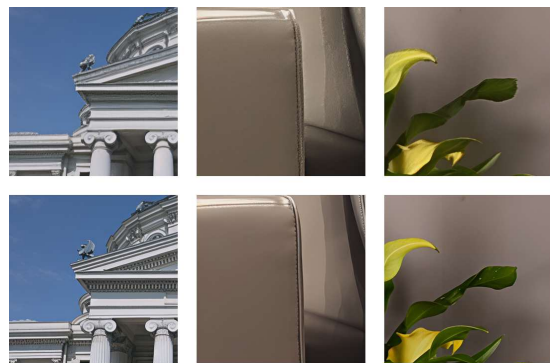
Complexity Comparisons

A complexity comparison of diffusion-based SR models is presented in Table 2, with metrics collected on an NVIDIA V100 GPU for 512×512 inputs. The results confirm the efficiency of the one-step inference paradigm. SCEESR maintains a significant speed advantage over multi-step models. When compared to the faster one-step model, OSEDiff[24], SCEESR introduces a modest increase in computational cost, which is a deliberate design trade-off that directly contributes to its notably higher reconstruction quality.

4.3 Ablation Study

Effectiveness of Controlnet and AME Loss

To validate the effectiveness of our proposed Semantic Control Edge Enhancement ControlNet and Adaptive Multi-Detector Edge Loss (AME), we compare two training phases. The first phase uses only L2 and LPIPS[4] losses without ControlNet, while the second phase incorporates the AME loss along with our ControlNet module. Quantitative results presented in Table 3 demonstrate that the key components of SCEESR play a crucial role in improving the quality of (SR images across multiple metrics. Furthermore, qualitative results in Figure 5 clearly show enhanced details and sharper edges in the reconstructed images. These findings confirm that the integration of the SCCE ControlNet and AME loss effectively boosts the perceptual and structural quality of SR outputs.

**Fig. 5** Qualitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks between the results of the first phase of SCEESR with only L2 loss and LPIPS loss(the first row) and the results of the second phase of SCEESR with semantic control edge enhancement ControlNet and AME loss(the second row).

5 Conclusion and Limitation

In this work, we introduce SCEESR, a novel super-resolution (SR) framework that enhances a pre-trained Stable Diffusion (SD) model via Low-Rank Adaptation (LoRA) modules and incorporates a semantic control edge enhancement module based on ControlNet. By leveraging edge-related information as input, SCEESR effectively improves both structural and perceptual quality. The framework employs the commonly used L2 loss to minimize pixel-wise regression errors and adopts

Table 3 Quantitative comparison with state-of-the-art methods on both synthetic and read-world benchmarks between the results of the first part of SCEESR with only L2 loss and LPIPS[4] loss and the results of the second part of SCEESR with semantic control edge enhancement ControlNet and AME loss.

Datasets	SCEESR Phase	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DIST $\downarrow$	FID $\downarrow$	NIQE $\downarrow$	MUSIQ $\uparrow$	CLIPQA $\uparrow$	MANIQA $\uparrow$
DIV2K-Val	1st	23.63	0.6087	0.2925	0.1976	26.64	4.7162	68.46	0.6713	0.6135
	2nd	23.78	0.6082	0.2904	0.1959	25.68	4.6687	68.93	0.6852	0.6242
DrealSR	1st	27.83	0.7832	0.2971	0.2169	134.78	6.5139	63.24	0.6947	0.5886
	2nd	28.13	0.7819	0.2963	0.2167	133.93	6.3457	65.41	0.6965	0.6054
RealSR	1st	25.06	0.7354	0.2823	0.2114	123.68	5.6114	69.43	0.6696	0.6399
	2nd	25.35	0.7386	0.2793	0.2091	123.73	5.5861	69.41	0.6698	0.6447

the LPIPS[4] loss to reduce perceptual discrepancies. SCEESR demonstrates strong performance in terms of both effectiveness and efficiency. Moreover, with the proposed AME loss, it further enhances structural quality, thereby simultaneously improving perceptual performance.

Despite its competitive performance in enhancing structural and perceptual quality, SCEESR introduces a modest increase in inference time and memory overhead. Furthermore, extreme input images may lead to the generation of meaningless edge maps. In such cases, the erroneous information introduced by ControlNet could adversely affect the diffusion process. To address these limitations, future work will focus on developing more robust edge detectors and semantic control mechanisms with higher tolerance levels, thereby improving the reliability and effectiveness of the integrated ControlNet.

Acknowledgements

We would like to express our sincere gratitude to Zhuang Yun Kai at ShanghaiTech University.

Statements and Declarations

Competing Interests: The authors have no financial or proprietary interests in any material discussed in this article.

Funding: No funds, grants, or other support was received.

Data Availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

- [1] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847 (2023). <https://doi.org/10.48550/arXiv.2302.05543>
- [2] Canny, J.: A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence (6), 679–698 (2009). <https://doi.org/10.1109/TPAMI.1986.4767851>
- [3] Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403 (2015). <https://doi.org/10.48550/arXiv.1504.06375>
- [4] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018). <https://doi.org/10.48550/arXiv.1801.03924>
- [5] Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015). <https://doi.org/10.1109/TPAMI.2015.2439281>
- [6] Li, H., Li, S.: Advancing scene text image super-resolution via edge enhancement priors. Signal, Image and Video Processing **18**(11), 8241–8250 (2024). <https://doi.org/10.1007/s11760-024-03467-9>
- [7] Zhang, L., Wan, Y.: Partial convolution residual network for lightweight image super-resolution. Signal, Image and Video Processing **18**(11), 8019–8030 (2024). <https://doi.org/10.1007/s11760-024-03448-y>
- [8] Xie, Y., Ou, J., Zhong, J., Jiang, T., Ma, T.: Multi-scale feedback residual network for image super-resolution. Signal, Image and Video Processing **19**(8), 635 (2025). <https://doi.org/10.1007/s11760-025-04255-9>
- [9] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 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, pp. 4681–4690 (2017). <https://doi.org/10.48550/arXiv.1609.04802>
- [10] Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4791–4800 (2021). <https://doi.org/10.48550/arXiv.2103.14006>
- [11] Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1905–1914 (2021). <https://doi.org/10.48550/arXiv.2107.10833>
- [12] Chen, C., Shi, X., Qin, Y., Li, X., Han, X., Yang, T., Guo, S.: Real-world blind super-resolution via feature matching with implicit high-resolution priors. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 1329–1338 (2022). <https://doi.org/10.1145/3503161.3547833>
- [13] Yao, Y., Cui, Z., Wang, D., Zhang, M.: Efrg-srgan: combining augmented features for real-world super-resolution. Signal, Image and Video Processing **18**(6), 5173–5187 (2024). <https://doi.org/10.1007/s11760-024-03224-y>
- [14] Kong, L., Zhong, H.: Rgsrgan: Image super-resolution reconstruction using generative adversarial networks based on recursive gated convolution: L. kong, h. zhong. Signal, Image and Video Processing **19**(13), 1126 (2025). <https://doi.org/10.1007/s11760-025-04711-6>
- [15] Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by

- residual shifting. *Advances in Neural Information Processing Systems* **36**, 13294–13307 (2023). <https://doi.org/10.48550/arXiv.2307.12348>
- [16] Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision* **132**(12), 5929–5949 (2024). <https://doi.org/10.1007/s11263-024-02168-7>
- [17] Khan, M.A., Nofresteri, M., Ezzati-Jivan, N.: Pasd: A performance analysis approach through the statistical debugging of kernel events. In: 2023 IEEE 23rd International Working Conference on Source Code Analysis and Manipulation (SCAM), pp. 151–161 (2023). IEEE. <https://doi.org/10.1109/SCAM59687.2023.00025>
- [18] Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seers: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 25456–25467 (2024). <https://doi.org/10.48550/arXiv.2311.16518>
- [19] Gendy, G., Sabor, N.: Fusion of transformer and diffusion features for real-world image super-resolution: G. gendy, n. sabor. *Signal, Image and Video Processing* **19**(9), 734 (2025). <https://doi.org/10.1007/s11760-025-04350-x>
- [20] Liu, X., Zhang, K., Yan, Z., Huang, S., Gu, Z., Zhang, Y., Wang, L.: Sdgdif: A spectral-domain guided diffusion model for underwater image super-resolution. *Signal, Image and Video Processing* **19**(9), 756 (2025). <https://doi.org/10.1007/s11760-025-04373-4>
- [21] Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.-P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: Sinsr: diffusion-based image super-resolution in a single step. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 25796–25805 (2024). <https://doi.org/10.48550/arXiv.2311.14760>
- [22] Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems* **37**, 92529–92553 (2024). <https://doi.org/10.48550/arXiv.2406.08177>
- [23] Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems* **36**, 8406–8441 (2023). <https://doi.org/10.48550/arXiv.2305.16213>
- [24] Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6613–6623 (2024). <https://doi.org/10.48550/arXiv.2311.18828>
- [25] Bevacqua, A., Singha, T., Pham, D.-S.: Enhancing semantic segmentation with synthetic image generation: A novel approach using stable diffusion and controlnet. In: 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 685–692 (2024). IEEE. <https://doi.org/10.1109/DICTA63115.2024.00103>
- [26] Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., Jampani, V.: Ziplora: Any subject in any style by effectively merging loras. In: European Conference on Computer Vision, pp. 422–438 (2024). Springer. https://doi.org/10.1007/978-3-031-73232-4_24
- [27] Stracke, N., Baumann, S.A., Susskind, J., Bautista, M.A., Ommer, B.: Ctrloralter: conditional loradapter for efficient 0-shot control and altering of t2i models. In: European Conference on Computer Vision, pp. 87–103 (2024). Springer. https://doi.org/10.1007/978-3-031-73232-4_6
- [28] Jones, M., Wang, S.-Y., Kumari, N., Bau, D., Zhu, J.-Y.: Customizing text-to-image models with a single image pair. In: SIGGRAPH Asia 2024 Conference Papers, pp. 1–13 (2024). <https://doi.org/10.1145/3680528.3687642>
- [29] Zheng, C., Gao, Y., Shi, H., Huang, M., Li, J., Xiong, J., Ren, X., Ng, M., Jiang, X., Li, Z., et al.: Dape: Data-adaptive positional encoding for length extrapolation. *Advances in Neural Information Processing Systems* **37**, 26659–26700 (2024). <https://doi.org/10.48550/arXiv.2405.14722>
- [30] Li, Y., Zhang, K., Liang, J., Cao, J., Liu, C., Gong, R., Zhang, Y., Tang, H., Liu, Y., Demandolx, D., et al.: Lsdrr: A large scale dataset for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1775–1787 (2023). <https://doi.org/10.1109/CVPRW59228.2023.00178>
- [31] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019). <https://doi.org/10.1109/TPAMI.2020.2970919>
- [32] Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126–135 (2017). <https://doi.org/10.1109/CVPRW.2017.150>
- [33] Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: European Conference on Computer Vision, pp. 101–117 (2020). Springer. https://doi.org/10.1007/978-3-030-58598-3_7
- [34] Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3086–3095 (2019). <https://doi.org/10.48550/arXiv.1904.00523>
- [35] Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence* **44**(5), 2567–2581 (2020). <https://doi.org/10.1109/TPAMI.2020.3045810>
- [36] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017). <https://doi.org/10.5555/3295222.3295408>
- [37] Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012). <https://doi.org/10.1109/LSP.2012.2227726>
- [38] Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5148–5157 (2021). <https://doi.org/10.48550/arXiv.2108.05997>
- [39] Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 2555–2563 (2023). <https://doi.org/10.1609/aaai.v37i2.25353>
- [40] Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1191–1200 (2022). <https://doi.org/10.48550/arXiv.2204.08958>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterialtoSCEESRSemanticControlEdgeEnhancementforDiffusionBasedSuperResolution.pdf](#)