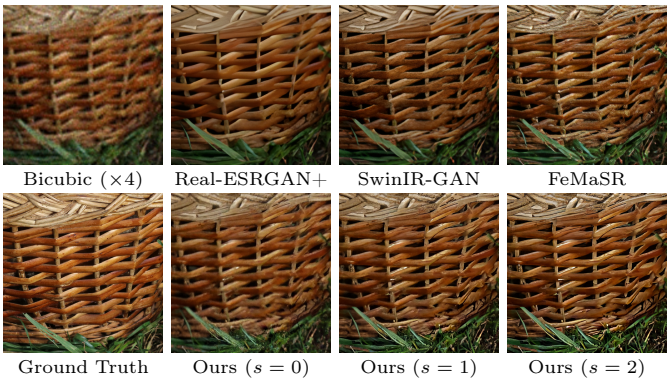
Real-SRGD: Enhancing Real-World Image Super-Resolution with Classifier-Free Guided Diffusion

Kenji Doi, Shuntaro Okada, Ryota Yoshihashi, and Hirokatsu Kataoka

LY Corporation



Elo rating scores from human subject study.

Methods	Rating Score ↑
Ours ($s = 2$)	**1751.91**
Ground Truth	1703.17
Ours ($s = 1$)	1673.21
Real-ESRGAN+	1566.90
SwinIR-GAN	1553.26
FeMaSR	1531.02
StableSR	1505.90
DiffBIR	1466.71
Ours ($s = 0$)	1436.97
RealDAN	1423.54
Swin2SR	1302.02
Bicubic (x4)	1068.58

Fig. 1: Comparison of real-world image super-resolution (RISR) outcomes from existing methods and our Real-SRGD, both without and with classifier-free guidance (CFG) [14]. CFG is a technique for scaling the output of a conditional generative model towards a desired objective, in this case, RISR. The Elo [10] rating scores from a human subject study are displayed in the right-hand table. (**Best view in zoom**)

Abstract. Real-world image super-resolution (RISR) aims to reconstruct high-resolution (HR) images from degraded low-resolution (LR) inputs, addressing challenges such as blurring, noise, and compression artifacts. Unlike conventional super-resolution (SR) approaches that typically generate LR images through synthetic downsampling, RISR confronts the complexity of real-world degradation. To effectively handle the intricate challenges of RISR, we adapt classifier-free guidance (CFG), a technique initially developed for multi-class image generation. Our proposed method, Real-SRGD (Real-world image Super-Resolution with classifier-free Guided Diffusion), decomposes RISR challenges into three distinct sub-tasks: Blind image restoration (BIR), conventional SR, and RISR itself. We then train class-conditional SR diffusion models tailored to these sub-tasks and use CFG to enhance the super-resolution performance in real-world settings. Our empirical results demonstrate that Real-SRGD surpasses existing state-of-the-art methods in both quantitative metrics and qualitative evaluations, as demonstrated by user studies. Moreover, our method demonstrates exceptional generalizability across

a range of conventional SR benchmark datasets. The code can be found at `https://github.com/yahoojapan/srgd`.

Keywords: Super-resolution · Diffusion model · Classifier-free guidance

1 Introduction

Single image super-resolution (SISR) is a fundamental and widely studied field in low-level computer vision, focusing on restoring high-resolution (HR) images from low-resolution (LR) inputs. Significant progress has been made in SISR [22], which focuses on precise image upsampling. However, real-world image super-resolution (RISR) remains challenging due to hurdles such as image degradation, including but not limited to blurring, noise, and compression artifacts.

Conventional super-resolution (SR) methods typically use pairs of LR images synthetically downsampled using methods such as bicubic downsampling, and their corresponding HR ones for training datasets. However, these datasets do not capture the complex degradation patterns found in real-world scenarios [26].

A primary obstacle in advancing RISR techniques is the need to address this compound degradation process. Notably, Real-ESRGAN [42] uses self-supervised learning and a data augmentation pipeline with multiple types of degradations, including image blurring, noise addition, JPEG compression, and downsampling. This method introduces high-order degradations by repeating sequences of degradation types, an approach that has been shown to significantly affect out-of-distribution (OOD) generalization [42].

Our work contributes to the further advancement of RISR by adapting classifier-free guidance (CFG) [14], which is one of the key techniques behind the success of diffusion models. CFG was initially developed for multi-class image generation to enable class-conditioning control [14]. However, as a bonus additional to the controllability, it has a pleasant side effect of suppressing the diversity in generation and consequently enhancing its adherence to generative conditions. This nature of CFG operates advantageously, particularly in super-resolution tasks where consistency is given precedence over diversity. We design our model to perform super-resolution processes based on multiple class conditions, and we ensure that these conditions include RISR. As depicted in Table 1, we decompose the RISR challenges into three sub-tasks: Blind Image Restoration (BIR), conventional SR, and RISR itself. Our novel method, Real-SRGD, leverages class-conditional SR diffusion models, each tailored to these sub-tasks, and uses CFG to enhance performance in real-world settings. Empirical results affirm that Real-SRGD outperforms current state-of-the-art methods in quantitative evaluation and user studies, additionally exhibiting remarkable generalizability across various conventional SR benchmarks.

Fig. 1 shows the comparison of existing methods and our Real-SRGD, both without and with CFG. As CFG scale increases, Real-SRGD enhances image resolution and often appears superior to other methods. Despite the simple approach, Real-SRGD results demonstrate a level of resolution that is comparable

to, if not better than, the ground truth image. The results of the human subject study suggest participants perceived both Real-SRGD and ground truth images as high-resolution, often favoring Real-SRGD for appearing more realistic.

Our contributions are summarized as follows:

1) We have devised a framework for applying CFG to super-resolution tasks for the first time, based on task decomposition. Implementing CFG for image processing tasks poses challenges due to the absence of readily available guidance information such as class labels and prompt texts. **2)** We establish our Real-SRGD as the new state-of-the-art through comprehensive benchmarking across diverse datasets, outperforming existing methods and corroborating superior perceptual image quality via studies involving human participants. **3)** While our primary focus is on RISR, our method outperforms existing RISR methods in generalizing conventional SR benchmarks, owing to the task decomposition implemented during training. This demonstrates that our method can serve as a versatile tool for image upsampling.

2 Related Work

2.1 Perceptual Quality in Super-Resolution

Benchmarks for assessing these SR methodologies have included metrics for both pixel-level accuracy and perceptual quality. While pixel-wise metrics, such as PSNR and SSIM [44], evaluate fidelity, perceptual metrics such as NIQE [31] provide insights into visual quality as perceived by humans. GAN-based models have improved perceptual quality at the expense of pixel-based metrics such as PSNR and SSIM. This phenomenon is known as the Perception-Distortion Trade-off [2], which highlights the inability to simultaneously achieve high pixel-based quality and perceptual quality. This necessitates the use of perceptual metrics such as NIQE over distortion-based metrics in RISR. Additionally, perceptual metrics like CLIP-IQA [40] and MUSIQ [19] are being utilized. CLIP-IQA takes advantage of the visual and linguistic understanding inherent in CLIP [32] to assess image quality. MUSIQ employs a Vision Transformer [9] to capture features at multiple scales within an image and understands their relationships to conduct quality assessment.

2.2 Background on Diffusion Models

Inspired by non-equilibrium thermodynamics, diffusion models are generative models that show state-of-the-art performance in density estimation and sample quality [20,8]. The central idea of the diffusion model is to define a diffusion process that gradually adds random noise to data (referred to as the forward diffusion process), and then learns to reverse this diffusion process to generate desired data samples from noise (referred to as the reverse process). In practice, given data $\mathbf{x}_0$ sampled from a real (possibly conditional) distribution $q(\mathbf{x}, \mathbf{c})$, a small level of Gaussian noise is added to the sample during the T steps of the

forward diffusion process, generating a series of noisy samples $\mathbf{x}_1, ..., \mathbf{x}_T$. Each step size, denoted by $\beta_1, ..., \beta_T$, follows the variance schedule.

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right). \tag{1}$$

Training is performed by optimizing the variational upper bound on negative log-likelihood, where the step size β_t can be learned through reparameterization [21] or held fixed as a hyperparameter. If you can sample data from $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$ by reversing the above process, the true sample can be re-generated from the Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. If β_t is sufficiently small, it is well-established that $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$ approximates a Gaussian distribution.

The reverse diffusion process is defined as a Markov chain with learned Gaussian transitions starting from $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ as follows:

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}\right) := \mathcal{N}\left(\mu_\theta\left(\mathbf{x}_t, \mathbf{c}, t\right), \Sigma_\theta\left(\mathbf{x}_t, \mathbf{c}, t\right)\right). \tag{2}$$

For super-resolution, we used a conditional diffusion model. The data distribution $q(\mathbf{x}, \mathbf{c})$ consists of HR images $\mathbf{x}$ and their corresponding LR images $\mathbf{c}$. In our method, we employ a class-conditional diffusion model that considers conditions from task classes that segment the larger RISR task into multiple sub-categories, in addition to the condition of the LR image.

2.3 Diffusion Model-Based Super-Resolution

The effectiveness of diffusion models for super-resolution tasks has been substantiated through various studies [36,35,6,18,43]. SR3 [36] and SR3+ [35] are indicative of the significant strides made by applying diffusion models to SR and RISR. Building on these foundational works, our Real-SRGD innovates further.

Recently, RISR techniques that involve foundation models like Stable Diffusion [33] have emerged. StableSR [41] exploits the features of LR images in conjunction with Stable Diffusion's intermediate features to steer the generative process, and DiffBIR [25] employs a two-stage pipeline, first incorporating a restoration module to address degradation and subsequently utilizes the generative capabilities of diffusion models.

In methods incorporating diffusion models like Stable Diffusion, the focus is on leveraging the generative capabilities of pre-trained diffusion models, utilizing classifier-free guidance or classifier guidance to control the balance between fidelity and realism of the generated images. However, our proposed method improves quality by using classifier-free guidance with models trained with different characteristics under class conditions.

3 Methodology

Fig. 2 shows an overview of our proposed method. To demonstrate the superiority of our method through a fair comparison, we repurposed the data degradations used in Real-ESRGAN [42] as seen in prior RISR work. However, as depicted in Table 1, we decompose real-world degradations into the four categories and define the RISR task that includes all of them, along with conventional SR and Blind image restoration (BIR) tasks, which incorporate subsets of these degradations. While BIR potentially includes degradations such as blur and resolution change, we have chosen to distinguish the three tasks in this study. Thus, we define BIR as a degradation that consists solely of noise and JPEG compression degradations.

Table 1: RISR task decomposition.

Task class	Blurring	Resolution change	Noise addition	Compression (JPEG)
RISR	✓	✓	✓	✓
SR	✓	✓		
BIR			✓	✓

3.1 Class-Conditional Training Pipeline

In our proposed method, we implement class-conditional training based on these sub-tasks. Thus, we generate low-resolution (LR) images by applying only a subset of the degradations in the pipeline, depending on the randomly selected task. In the case depicted in the Fig. 2, the SR task is chosen and only blur and resolution change degradations are used.

The conditions of an LR image are concatenated with input noise along the channel axis before being fed into the model. Simultaneously, the task class condition is encoded as a vector embedding with the same dimension as the timestep embedding, and it is added to the timestep embedding before being integrated into the residual blocks of the model. During inference, the LR condition remains unchanged, whereas the noisy image ($\mathbf{x}_t$) is iteratively denoised. Furthermore, during the training of our models, the task class condition is dropped with a certain probability to enable classifier-free guidance during inference.

After training, the model is utilized as an RISR model during inference by consistently specifying the RISR task class. Therefore, no additional labeling or annotation is required for the input LR images.

3.2 Classifier-Free Guidance for SR Diffusion Model

Our model is trained in a class-conditional manner, with the classes randomly selected from the previously mentioned three tasks. Furthermore, during training, at a given probability (for instance, 10%), we drop the class conditions, thereby co-training both class-conditioned and class-unconditioned models. The model without class conditions tends to be conservative, aiming to accommodate any of the randomly chosen tasks. Conversely, the class-conditioned model morphs into a version specialized towards the selected task. When using CFG

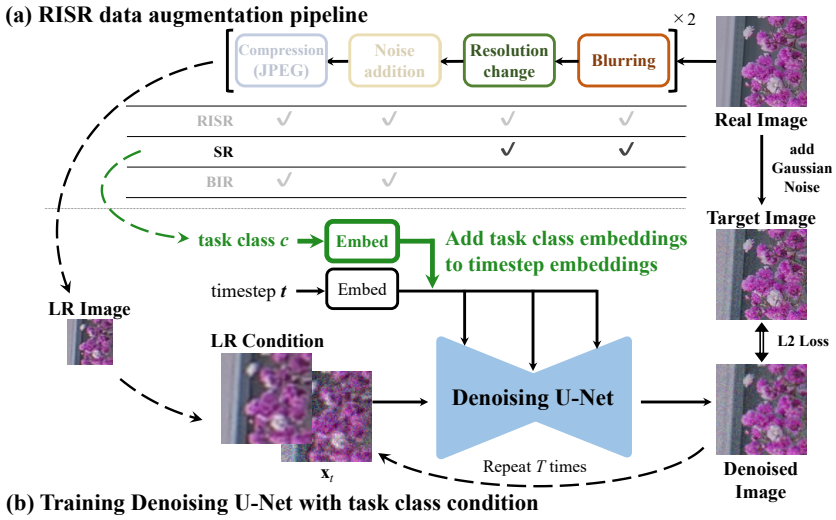


Fig. 2: Overview of our Real-SRGD. **(a) RISR data augmentation pipeline:** Degradation is applied twice to real images according to the randomly selected task class, following the sequence indicated in this figure. **(b) Training the Denoising U-Net with task class condition:** LR images are superimposed with noise and then fed into our model along with task-class conditions—either RISR, SR, or BIR—converted into embeddings and merged with timestep embeddings. To utilize classifier-free guidance, during the training, the task class condition is dropped with a certain probability.

for a RISR task, we compute the discrepancy between the predictions of the RISR-specialized model and the conservative model without class conditions. The difference between these predictions is considered to be the key factor that specializes our method towards RISR. Scaling the predictions of the model conditioned on RISR in this direction allows us to further specialize the model's predictions for the RISR task.

In terms of a specific implementation, we include task conditions to equation 2. It can be re-written as follows, with $\mathbf{c}_{LR}$ denoting the LR image condition and $\mathbf{c}_T$ the task class conditions:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}_{LR}, \mathbf{c}_T) := \mathcal{N}\left(\mu_\theta\left(\mathbf{x}_t, \mathbf{c}_{LR}, \mathbf{c}_T, t\right), \Sigma_\theta\left(\mathbf{x}_t, \mathbf{c}_{LR}, \mathbf{c}_T, t\right)\right). \quad (3)$$

Then, we integrate CFG into our super-resolution diffusion model. Given a model $\epsilon_\theta(\mathbf{x}, \mathbf{c}_T, t)$ that executes class-conditional generation, CFG is accomplished by correcting the denoising process at each timestep as follows: (the LR condition, $\mathbf{c}_{LR}$, is constantly supplied as an input to the model, so we omit it from the function to streamline notation.)

$$\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}_T, t) = \epsilon_\theta\left(\mathbf{x}_t, \mathbf{c}_T, t\right) - s\left(\epsilon_\theta\left(\mathbf{x}_t, \varnothing, t\right) - \epsilon_\theta\left(\mathbf{x}_t, \mathbf{c}_T, t\right)\right) \tag{4}$$

$$= (1+s)\epsilon_\theta\left(\mathbf{x}_t, \mathbf{c}_T, t\right) - s\epsilon_\theta\left(\mathbf{x}_t, \varnothing, t\right). \tag{5}$$

In the equation above, $\varnothing$ represents the absence of the class condition in the model input, and s, short for scale, is a hyperparameter used to modulate the strength of the CFG. When $s = 0$, the model operates devoid of any form of guidance.

3.3 Comparison wtih Classifier Guidance

Classifier guidance (CG) is a technique that guides the generation results of a diffusion model towards a specific class by using an extra trained classifier [8].

We also compared the performance of our method when using CG instead of CFG. For a detailed discussion on the experimental settings, see Section 4.10. As per the results in Table 7, the guidance provided by CG did not reach the performance of CFG. Moreover, combining CFG and CG resulted in poorer performance than using CFG alone. Based on these results, we decided to adopt CFG in our proposed method.

3.4 Architecture

Our U-Net-based [34] architecture, inspired by the DDPM [13] model, incorporates LR conditions into the noise input by enlarging the size of the input channels. To embed the LR image within our system, we upscale it to the target resolution using bicubic interpolation. We then align it to the channel axis with the noise input, while consistently producing a 3-channel output. The network's residual blocks incorporate timestep embeddings, which are translated into sinusoidal positional embeddings used in Transformers [39], and the task conditions are transformed into vector embeddings with the same dimensionality as the timestep embeddings.

4 Experiments

4.1 Training and Testing Datasets

For training, we utilized the DIV2K [24], DIV8K [11], Flickr2K [37] and OutdoorSceneTraining (OST) [48] datasets.

Our method was evaluated on synthetic and real-world datasets. The synthetic dataset utilized was the DIV2K Realistic-Wild dataset [38] (hereafter DIV2K-Wild), while the real-world datasets include DPED-iPhone [16], RealSRv3 [3], and DRealSR [45]. DIV2K-Wild, part of the NTIRE 2018 Super-Resolution Challenge (Track 4) [38], comprises DIV2K images subjected to various real-world degradations that differ across images. The DPED-iPhone dataset

includes 100 LR images captured by smartphone cameras. The RealSRv3 and DRealSR datasets are designed for RISR, and are composed of images taken with several cameras at different settings. We test our models with an upscaling factor of 4, which is the standard in this field.

4.2 Evaluation Metrics

We evaluate the quality of generated images using various perceptual metrics including LPIPS [51], NIQE [31], CLIP-IQA [40], and MUSIQ [19] over traditional pixel-based metrics due to the perceptual distortion trade-offs [2] inherent in realistic enhancements. Additionally, we report PSNR and SSIM scores for reference. Furthermore, we used the Fréchet Inception Distance (FID) [12] score to evaluate the RealSRv3 and DRealSR datasets. Considering the necessity for a large set of images in the FID calculation, we cropped 10,000 non-overlapping patches (sized 256×256 pixels) from the training sets of these datasets and calculated the FID scores—referred to as FID10K [12]. This approach follows the precedent set by the SR3+ evaluation strategy [35].

4.3 Diffusion Model Selection for RISR

To determine the best diffusion models for our proposed method, we experimented with various models, timesteps, and noise schedules. Following extensive trials, we selected two models: the continuous timestep DDPM [13] model with a linear noise schedule, hereafter referred to as CDM, as the model prioritizing perceptual image quality, and the EDM [17] model, chosen for prioritizing processing speed.

4.4 Training and Inference Details

Real-SRGD models were trained using the AdamW optimizer and L2 loss as the optimization criterion, with the task class condtion being dropped at a probability of 10% during training. For the appropriate selection of the drop probability, refer to section 4.10. During inference, we discerned that 250 generation timesteps yielded an optimal balance between quality and efficiency in our CDM model. The EDM model used a 32-step setting for both training and generation, serving as a lightweight and efficient alternative. The EDM's generation speed is increased through the use of the DPM++ sampler [27], enabling a twofold increase in generation speed over the standard setting.

4.5 Comparison with Baseline, EDM and CDM Models

We compared the performance of our proposed methods (EDM and CDM) with a baseline model— without class-conditional learning on the DIV2K-Wild and DPED-iPhone datasets. Table 2 shows the results. CFG was found to enhance the perceived quality when applied (i.e., when $s \geq 1.0$), as indicated by the improved NIQE, CLIP-IQA and MUSIQ scores, despite a marginal drop in PSNR and SSIM in line with perception-distortion trade-off principles.

Table 2: Comparison of baseline, EDM (efficiency-oriented) and CDM (quality-oriented) on DIV2K-Wild and DPED-iPhone datasets. (**Bold** and underline number are best and second best performance in all tables.)

Methods	DIV2K-Wild						DPED-iPhone		
	PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑
baseline	17.55	0.4554	0.4035	3.269	0.5881	55.40	3.918	0.3831	45.58
EDM ($s = 0$)	16.98	0.3454	0.5630	3.389	0.5080	46.68	3.993	0.2697	34.61
EDM ($s = 1$)	16.47	0.3381	0.5092	2.928	0.6373	55.55	3.495	0.3672	46.99
EDM ($s = 2$)	15.96	0.3265	0.4951	2.729	0.6844	59.19	**3.213**	0.4322	52.00
CDM ($s = 0$)	**17.73**	**0.4600**	0.4367	3.732	0.4707	47.51	4.148	0.3273	39.81
CDM ($s = 1$)	17.00	0.4275	**0.4000**	2.866	0.7125	62.24	3.522	0.4887	53.13
CDM ($s = 2$)	16.34	0.3984	0.4183	**2.722**	**0.7711**	**65.85**	3.365	**0.5757**	**57.75**

Table 3: RISR results on DIV2K-Wild and DPED-iPhone datasets.

Methods	DIV2K-Wild						DPED-iPhone		
	PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑
Real-ESRGAN+	17.50	0.4852	0.4047	4.365	0.6126	54.82	5.105	0.3008	43.89
SwinIR-GAN	17.23	0.4740	0.3858	3.813	0.6711	57.64	4.773	0.3248	42.10
FeMaSR	16.92	0.4045	0.3969	4.156	0.7694	62.7	5.117	0.4860	48.29
RealDAN	17.69	0.4806	0.4388	5.359	0.4813	47.96	6.349	0.2571	35.54
Swin2SR	17.80	**0.5021**	0.5017	7.574	0.4776	44.05	8.291	0.2999	33.00
StableSR	17.71	0.4853	0.4182	5.268	0.4890	48.20	5.708	0.3038	40.84
DiffBIR	**17.88**	0.4679	0.4173	5.352	0.6463	53.67	5.608	0.5167	43.63
DiffIR	17.69	0.4827	**0.3586**	4.853	0.6289	56.32	5.688	0.2082	38.64
SeeSR	17.33	0.4744	0.4203	5.044	0.7267	52.98	5.674	0.5459	41.44
SUPIR	17.19	0.4517	0.3860	3.520	**0.8051**	64.42	5.499	0.4200	43.41
Ours$_{CDM}$ ($s = 2$)	16.34	0.3984	0.4183	**2.722**	0.7711	**65.85**	**3.365**	**0.5757**	**57.75**

4.6 Comparison with Existing Methods

We evaluated our proposed methods and ten representative RISR methods: Real-ESRGAN+ [42], SwinIR-GAN [23], FeMaSR [4], RealDAN [28], Swin2SR [7], StableSR [41], DiffBIR [25], DiffIR [47], SeeSR [46], and SUPIR [49].

RISR on DIV2K-Wild and DPED-iPhone Datasets We evaluate our CDM ($s = 2.0$) and existing methods. Table 3 shows the scores for the RISR results on DIV2K-Wild and DPED-iPhone datasets. Since the DPED-iPhone dataset does not have ground truth data, PSNR, SSIM and LPIPS scores cannot be used. Our method was inferior to existing methods in PSNR, SSIM and LPIPS; however, it outperformed existing methods in the scores for NIQE, CLIP-IQA, and MUSIQ. Fig. 3 shows comparison of RISR results on heavily degraded sample from DIV2K-Wild dataset.

RISR on RealSRv3 and DRealSR Datasets We evaluate six variants of our method (three patterns of classifier-free guidance (CFG) scale: $s = 0, 1, 2$, for both EDM and CDM models) and existing methods on RealSRv3 and DRealSR

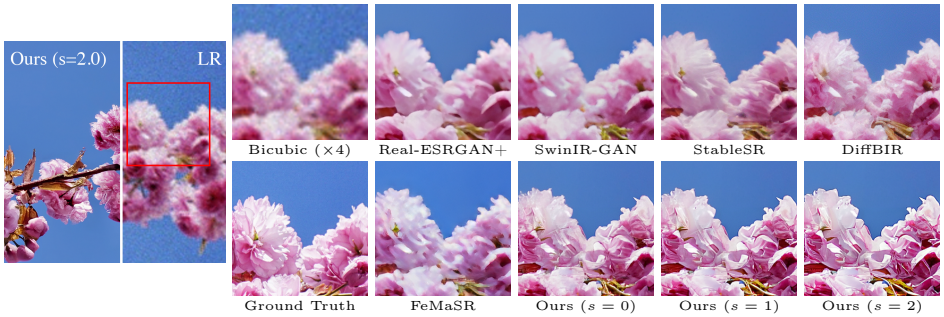


Fig. 3: Comparison of RISR results on heavily degraded sample from DIV2K-Wild.

Table 4: RISR results on RealSRv3 and DRealSR datasets.

Methods	RealSRv3				DRealSR			
	FID10K ↓	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑	FID10K ↓	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑
Real-ESRGAN+	33.41	9.550	0.3643	45.82	28.92	9.652	0.3743	41.36
SwinIR-GAN	28.44	9.215	0.4422	44.84	27.21	9.282	0.4512	41.16
FeMaSR	35.30	8.898	0.4728	45.34	32.21	8.492	0.4686	42.83
RealDAN	54.07	9.918	0.3471	36.87	60.92	10.24	0.3092	33.24
Swin2SR	63.64	14.02	0.2904	38.09	72.93	14.29	0.2748	31.12
StableSR	24.28	9.249	0.5737	50.33	35.45	9.163	0.5252	46.04
DiffBIR	35.60	9.564	0.5723	<u>52.80</u>	53.32	9.523	0.5690	**51.92**
DiffIR	<u>24.07</u>	9.179	0.2728	41.43	**20.56**	9.025	0.2643	38.91
SeeSR	41.76	11.95	0.4879	48.21	49.09	12.64	0.5108	48.21
SUPIR	37.34	10.46	<u>0.6398</u>	41.74	41.68	10.61	0.5662	41.74
Ours$_{EDM}$ ($s = 0$)	33.71	**7.207**	0.2730	30.21	25.53	<u>7.574</u>	0.3043	33.10
Ours$_{EDM}$ ($s = 1$)	25.34	<u>7.298</u>	0.3845	39.89	23.76	**7.545**	0.4325	41.49
Ours$_{EDM}$ ($s = 2$)	25.63	7.650	0.4354	44.63	26.80	7.709	0.4877	44.94
Ours$_{CDM}$ ($s = 0$)	**23.72**	9.541	0.4326	39.44	<u>23.20</u>	9.514	0.4216	35.75
Ours$_{CDM}$ ($s = 1$)	27.27	9.473	0.6134	52.56	26.49	9.299	<u>0.5884</u>	47.57
Ours$_{CDM}$ ($s = 2$)	33.72	9.800	**0.6453**	**54.39**	33.90	9.567	**0.6342**	<u>50.64</u>

datasets. Table 4 shows the results. Our methods achieved the highest scores across all evaluation metrics, except for MUSIQ on the DRealSR dataset where it ranked second, using the FID10K, NIQE, CLIP-IQA, and MUSIQ metrics. For FID10K, the CDM model achieved the best score with no CFG, and when CFG was applied, the score worsened. In the EDM model, the score improves with CFG scale at $s = 1.0$, but after that, it gets worse. This could likely be attributed to the fact that the FID metric focuses not on the perceptual quality of an image, but on the “difference in distribution" between the ground truth data and the data generated by our method. Considering these results, in Section 4.8, we will conduct an evaluation of RISR quality through a human subject study.

Conventional SR on Classical Benchmarks We also tested our method on four classical benchmark datasets: Set5 [1], Set14 [50], BSD100 [30], and Urban100 [15]. The results are shown in Table 5. Our model can specify tasks other

Table 5: Conventional SR results on Set5, Set14, BSD100, and Urban100.

Methods	Type	Set5			Set14			BSD100			Urban100		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SwinIR	SR	<u>30.89</u>	<u>0.8877</u>	0.1663	<u>27.01</u>	<u>0.7787</u>	0.2664	<u>26.60</u>	<u>0.7474</u>	0.3533	<u>25.87</u>	<u>0.8197</u>	0.1840
HAT	SR	**30.98**	**0.8903**	0.1598	**27.13**	**0.7831**	0.2587	**26.68**	**0.7517**	0.3464	**26.73**	**0.8381**	<u>0.1668</u>
Real-ESRGAN+	RISR	24.33	0.7322	0.1695	23.30	0.6505	0.2339	23.57	0.6277	0.2819	21.06	0.6574	0.2140
SwinIR-GAN	RISR	24.84	0.7232	0.1659	23.26	0.6532	0.2267	23.51	0.6305	0.2598	21.02	0.6673	0.2014
FeMaSR	RISR	23.25	0.7126	<u>0.1507</u>	21.82	0.6065	<u>0.2162</u>	21.81	0.5869	0.2517	20.22	0.6381	0.1983
RealDAN	RISR	25.55	0.7772	0.1922	24.13	0.6955	0.2920	24.33	0.6827	0.3761	21.86	0.6844	0.2922
Swin2SR	RISR	26.95	0.8006	0.2224	25.04	0.7119	0.3339	25.28	0.6882	0.4167	23.10	0.7324	0.2740
StableSR	RISR	23.37	0.6966	0.1824	22.01	0.6036	0.2447	22.41	0.5955	0.2740	20.56	0.6326	0.2032
DiffBIR	RISR	24.92	0.7299	0.1606	22.85	0.5896	0.2177	23.79	0.6043	**0.2351**	21.41	0.6282	0.2120
Ours_{CDM} $(s = 2)$	SR	28.77	0.8452	**0.1488**	24.77	0.7070	**0.1890**	24.99	0.6867	<u>0.2424</u>	23.54	0.7482	**0.1584**

than RISR, such as BIR and SR; hence, we also investigated the performance of CDM model when used as a conventional SR model. For evaluation metrics, we used PSNR, SSIM, and LPIPS, which are reference-based metrics that consider the Ground Truth images. We compared our model with SwinIR [23], a representative in the field of conventional SR, HAT [5], the current state-of-the-art method, and also the existing methods for RISR that we have been comparing with. From the PSNR and SSIM scores, it is evident that SwinIR and HAT, designed specifically for conventional SR tasks, performed best. However, our model is also producing competitive scores, not inferior to these conventional SR methods. It surpassed the existing methods designed specifically for RISR in all cases except for the LPIPS score on the BSD100 dataset, where it ranked second.

4.7 Generation Randomness in Diffusion-Based Super-Resolution

In this section, we evaluated the degree of randomness in RISR results generated by diffusion-based methods, specifically DiffBIR, StableSR, and our proposed method, which includes the use of a CFG scheme. Generally, the use of CFG is known to suppress the diversity in generation, thereby enhancing its conformity with generative conditions. This works advantageously, especially in super-resolution tasks where consistency is prioritized over diversity. This decrease in diversity does not hinder our method but instead improves the quality of the generated images. For each method, we changed the random seed and performed RISR processing five times for the DIV2K-Wild dataset (a total of 100 images). We analyzed the variations in NIQE scores for the resulting 500 images.

The experimental results are shown in Fig. 4. Even without CFG, our proposed method exhibited a lower average NIQE and less variability compared with the existing models. Upon applying the CFG, both the average NIQE and variability decreased further. This demonstrates that our proposed method not only outperforms the existing ones in the absence of CFG but also improves the quality and stability of RISR when CFG is applied.

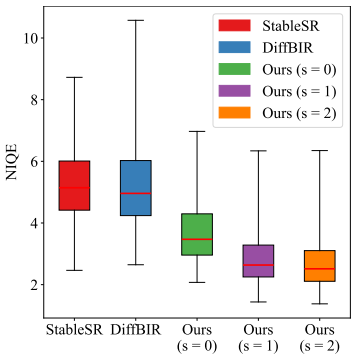


Fig. 4: Distributions of NIQE scores obtained by methods that include generation randomness.

Table 6: Processing speed, along with DIV2K-Wild results, sorted by Elo rating score.

Methods	sec/ sample	DIV2K-Wild			Rating Score ↑
		NIQE ↓	CLIP-IQA ↑	MUSIQ ↑	
Ours$_{CDM}$ ($s = 2$)	437	**2.722**	0.7711	**65.85**	**1751.91**
Ground Truth	-	3.072	**0.7754**	65.55	1703.17
Ours$_{CDM}$ ($s = 1$)	437	2.866	0.7125	62.24	1673.21
Ours$_{EDM}$ ($s = 2$)	105	2.729	0.6844	59.19	1641.61
Real-ESRGAN+	0.52	4.365	0.6126	54.82	1566.90
SwinIR-GAN	2.10	3.813	0.6711	57.64	1553.26
Ours$_{EDM}$ ($s = 1$)	105	2.928	0.6373	55.55	1531.84
FeMaSR	1.96	4.156	0.7694	62.77	1531.02
StableSR	268	5.268	0.4890	48.20	1505.90
DiffBIR	81.6	5.352	0.6463	53.67	1466.71
Ours$_{CDM}$ ($s = 0$)	194	3.372	0.4707	47.51	1436.97
RealDAN	**0.36**	5.359	0.4813	47.96	1423.54
Ours$_{EDM}$ ($s = 0$)	54.9	3.389	0.5080	46.68	1343.34
Swin2SR	12.9	7.574	0.4776	44.05	1302.02
Bicubic (4x)	-	8.047	0.3583	21.72	1068.58

4.8 Human Subject Study

We conducted a human subject study to validate our method's qualitative results. We evaluated six variants of our method and seven RISR methods, as well as bicubic upsampled and ground truth (GT) images. DiffIR, SeeSR, and SUPIR were not included as they were quantitatively evaluated after this study. For the experiments, we utilized images from the RealSRv3, DRealSR, and the DIV2K-Wild datasets.

In studies examining which assessment methods should be chosen for subjective image quality evaluation, it has been concluded that the forced-choice pairwise comparison method results in the smallest measurement variance, thereby producing the most accurate results [29]. Based on this, we adopted this method for our user study.

Participants were presented with image pairs and asked to select the one with perceived better quality. To facilitate comparison, we used the top 2,000 images with the largest variance after super-resolution from the 10,000 RealSRv3 and DRealSR images. For the DIV2K-Wild dataset, we used a 256×256 crop with the maximum variance between methods from each of the 100 samples after super-resolution. A web-based system was developed for the random presentation of image pairs, through which we collected 2,100 votes from 14 participants. To calculate ratings from the forced-choice pairwise comparison results, we utilized the Elo rating system [10], often used for evaluating players in paired competitive games, due to the reliability of the ratings, and it is also employed in quality evaluation of super-resolution models.

Final ratings are referred to in the Rating Score column of Table 6. As per the rankings, our proposed method achieved higher rankings than the existing methods. Particularly, the CDM model (with $s = 2.0$) significantly surpassed the top-rated existing method, which is Real-ESRGAN+, with a considerable

Table 7: Comparison of Classifier Guidance with Classifier-Free Guidance.

Methods	RealSRv3				DRealSR			
	FID10K ↓	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑	FID10K ↓	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑
baseline	<u>22.59</u>	9.141	0.4935	45.05	**22.32**	9.133	0.4724	41.84
baseline + CG ($s = 1$)	**22.56**	<u>8.797</u>	0.5107	47.17	<u>22.46</u>	<u>8.678</u>	0.4962	44.83
baseline + CG ($s = 8$)	23.08	**8.643**	0.5261	48.63	23.91	**8.525**	0.5209	46.84
baseline + CG ($s = 32$)	26.52	9.122	0.5394	50.23	32.02	9.199	0.5415	48.87
baseline + CG ($s = 64$)	32.18	10.463	0.5351	50.46	42.71	11.140	0.5275	49.11
CDM + CFG ($s = 2$) + CG ($s = 32$)	40.30	14.299	<u>0.6041</u>	52.70	54.69	16.834	<u>0.5830</u>	50.35
CDM + CFG ($s = 2$) + CG ($s = 64$)	42.34	15.111	0.6028	<u>52.78</u>	59.31	15.111	0.5826	<u>50.63</u>
CDM + CFG ($s = 2$)	33.72	9.800	**0.6453**	**54.39**	33.90	9.567	**0.6342**	**50.64**

margin in the confidence interval. Moreover, an increase in the scale of the CFG resulted in a corresponding rise in the rankings. Surprisingly, our CDM model surpasses the ranking of the ground truth. While we only compared up to CFG scale 2.0 for our method to limit the number of methods compared, the NIQE, CLIP-IQA and MUSIQ scores rises for CFG scales larger than 2.0, indicating the possibility of further improvements in human subjectivity-based evaluations.

4.9 RISR Quality vs. Computational Cost

Our proposed method comprises an efficiency-oriented EDM model and a quality-oriented CDM model. We measured the processing time for processing 100 images from the DIV2K-Wild dataset, including existing methods, to investigate the trade-off between RISR quality and computational cost.

Results are shown in Table 6. Considering that the adoption of CFG doubles the inference cost to twice as much, the CDM model, which applied the CFG that achieved the best performance, significantly outweighed the computational cost of the pre-existing methods. On the flip side, while the EDM model, which does not use CFG, was the least computationally hungry among our proposed methods, if we consider the DIV2K-Wild's quantitative performance scores and the Elo Rating Score, it fell short of delivering performance proportional to its computational cost. With a relatively high Elo rating score that stood second in computational cost only to RealDAN, Real-ESRGAN+ can be called a well-balanced method. To trim down the computational cost of our methods, future research issues include using distillation methods for diffusion models considering CFG or utilizing a faster sampler.

4.10 Ablation Studies

Comparison of Classifier Guidance with Classifier-Free Guidance We evaluated our method with classifier guidance (CG) [8] instead of CFG. The implementation of CG was based on the proposed paper [8], utilizing a U-Net-based classifier that was trained to classify images with noise added according to the timestep. We compared the results of CG with our proposed method that employs CFG on RealSRv3 and DRealSR datasets. Table 7 shows the results.

Table 8: Comparison of condition drop probability.

Methods	drop rate	RealSRv3				DRealSR			
		FID10K ↓	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑	FID10K ↓	NIQE ↓	CLIP-IQA ↑	MUSIQ ↑
CDM ($s = 0$)	0.1	<u>28.51</u>	9.281	0.4926	44.05	<u>29.07</u>	9.275	0.4785	40.95
CDM ($s = 1$)	0.1	33.33	8.620	0.6012	49.57	36.78	8.557	0.5828	46.83
CDM ($s = 2$)	0.1	39.10	**8.376**	**0.6365**	**52.34**	46.23	**8.434**	**0.6276**	**49.92**
CDM ($s = 0$)	0.2	**28.12**	9.263	0.4720	43.08	**28.60**	9.053	0.4659	40.15
CDM ($s = 1$)	0.2	30.98	9.126	0.5758	48.18	33.13	8.846	0.5568	45.69
CDM ($s = 2$)	0.2	34.59	<u>8.620</u>	<u>0.6130</u>	<u>50.76</u>	40.52	<u>8.529</u>	<u>0.5997</u>	<u>48.61</u>

Similar to CFG, CG has a scale parameter that controls the degree of guidance. Even after considerably increasing the scale, gradual improvements in the perceptual metrics were observed, so we investigated the scale parameter up to the range where no further improvements in scores were found. However, the results did not match the performance achieved with CFG. Moreover, combining CFG and CG led to poorer performance than using CFG alone.

Table 7 shows CFG s=2 improves CLIP-IQA and MUSIQ but degrades FID10K and NIQE compared to CG (s=1 or 8). Strong CFG (s=2) may enhance perceptual quality but also deviate from real data distribution, impacting FID10K and NIQE. This highlights the trade-offs of generative models and the need for diverse evaluation metrics, including human perception, for comprehensive image quality assessment.

Comparison of Condition Drop Probability We also examined the influence of the task class condition drop probability on super-resolution quality during training. An ablation in the foundational CFG paper identified a 10% drop probability as optimal, we evaluated this rate against 20% in our Real-SRGD method on RealSRv3 and DRealSR datasets. Table 8 shows the results. Based on the results, we adopted a 10% drop probability over 20% as it yielded better outcomes when applying the CFG.

5 Conclusion and Discussion

This work introduces Real-SRGD, a novel approach for real-world image super-resolution (RISR) using Classifier-Free Guidance within a diffusion framework. Real-SRGD achieves state-of-the-art performance, outperforming existing methods on benchmarks and human evaluations. Our task-specific training, which decomposes RISR into multiple components, enables the effective use of Classifier-Free Guidance and can be easily integrated into other diffusion-based methods. While we utilized a simple model and existing data augmentation techniques, future work could explore more advanced models and degradations for further improvement. Our findings suggest that the proposed approach of task decomposition and reconstruction has the potential to benefit a wide range of image processing tasks beyond RISR.

References

1. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi Morel, M.L.: Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In: BMVC (2012)
2. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: CVPR. pp. 6228–6237 (2018)
3. 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: ICCV (2019)
4. 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: ACMMM (2022)
5. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: CVPR. pp. 22367–22377 (2023)
6. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. In: ICCV. pp. 14367–14376 (2021)
7. Conde, M.V., Choi, U.J., Burchi, M., Timofte, R.: Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration. In: ECCV (2022)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS. vol. 34, pp. 8780–8794 (2021)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
10. Elo, A.E., Sloan, S.: The rating of chessplayers : past and present. Ishi Press International (2008)
11. Gu, S., Lugmayr, A., Danelljan, M., Fritsche, M., Lamour, J., Timofte, R.: Div8k: Diverse 8k resolution image dataset. In: ICCV. pp. 3512–3516 (2019)
12. 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. In: NeurIPS. vol. 30 (2017)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS (2021)
15. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR. pp. 5197–5206 (2015)
16. Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., Van Gool, L.: Dslr-quality photos on mobile devices with deep convolutional networks. In: ICCV. pp. 3277–3285 (2017)
17. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: NeurIPS (2022)
18. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. In: NeurIPS (2022)
19. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV. pp. 5128–5137 (2021)
20. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. In: NeurIPS. vol. 34, pp. 21696–21707 (2021)
21. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
22. Li, J., Pei, Z., Zeng, T.: From beginner to master: A survey for deep learning-based single-image super-resolution. ArXiv **abs/2109.14335** (2021)

23. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: ICCV. pp. 1833–1844 (2021)
24. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPR (2017)
25. Lin, X., He, J., Chen, Z., Lyu, Z., Fei, B., Dai, B., Ouyang, W., Qiao, Y., Dong, C.: Diffbir: Towards blind image restoration with generative diffusion prior (2024)
26. Liu, A., Liu, Y., Gu, J., Qiao, Y., Dong, C.: Blind image super-resolution: A survey and beyond. IEEE TPAMI **45**(05), 5461–5480 (2023)
27. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models (2023)
28. Luo, Z., Huang, Y., Li, S., Wang, L., Tan, T.: End-to-end alternating optimization for real-world blind super resolution. IJCV (2023)
29. Mantiuk, R.K., Lewandowska, A., Mantiuk, R.: Comparison of four subjective methods for image quality assessment. Computer Graphics Forum **31** (2012)
30. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV. pp. 416–423 (2001)
31. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Process. Lett. **20**, 209–212 (2013)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. vol. 139, pp. 8748–8763 (2021)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
35. Sahak, H., Watson, D., Saharia, C., Fleet, D.: Denoising diffusion probabilistic models for robust image super-resolution in the wild (2023)
36. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2023)
37. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L., Lim, B., et al.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPR (2017)
38. Timofte, R., Gu, S., Wu, J., Van Gool, L., Zhang, L., Yang, M.H., Haris, M., et al.: Ntire 2018 challenge on single image super-resolution: Methods and results. In: CVPR (2018)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NeurIPS. vol. 30 (2017)
40. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI. vol. 37, pp. 2555–2563 (2023)
41. Wang, J., Yue, Z., Zhou, S., Chan, K.C.K., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution (2024)
42. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCV. pp. 1905–1914 (2021)
43. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. ICLR (2023)
44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**, 600–612 (2004)

45. 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: ECCV (2020)
46. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution (2024)
47. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Gool, L.V.: Diffir: Efficient diffusion model for image restoration (2023)
48. Xintao Wang, Ke Yu, C.D., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR (2018)
49. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild (2024)
50. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces. pp. 711–730 (2012)
51. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)