Effective Diffusion Transformer Architecture for Image Super-Resolution

Kun Cheng[1]*, Lei Yu[2]*, Zhijun Tu[2], Xiao He[1], Liyu Chen[2], Yong Guo[3], Mingrui Zhu[1], Nannan Wang[1]†, Xinbo Gao[4], Jie Hu[2]

[1]State Key Laboratory of Integrated Services Networks, Xidian University
[2]Huawei Noah's Ark Lab
[3]Consumer Business Group, Huawei
[4]Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications
kunncheng@stu.xidian.edu.cn, yulei96@huawei.com, nnwang@xidian.edu.cn

Abstract

Recent advances indicate that diffusion model holds great promise in image super-resolution. While latest methods are primarily based on latent diffusion models with convolutional neural networks, there are few attempts to explore transformers, which have demonstrated remarkable performance in image generation. In this work, we design an effective diffusion transformer for image super resolution (DiT-SR) that achieves the visual quality of prior-based methods, but through a training-from-scratch manner. In practice, DiT-SR leverages an overall U-shaped architecture, and adopts uniform isotropic design for all the transformer blocks across different stages. The former facilitates multi-scale hierarchical feature extraction, while the latter reallocate the computational resources to critical layers to further enhance performance. Moreover, we thoroughly analyze the limitation of the widely used AdaLN, and present a frequency-adaptive time-step conditioning module, enhancing the model's capacity to process distinct frequency information at different time steps. Extensive experiments demonstrate that DiT-SR outperforms the existing training-from-scratch diffusion-based SR methods significantly, and even beats some of the prior-based methods on pretrained Stable Diffusion, proving the superiority of diffusion transformer in image super resolution.

Code — https://github.com/kunncheng/DiT-SR

Introduction

Image super-resolution (SR) aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) input. Recently, diffusion models (DMs) (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021; Rombach et al. 2022) have demonstrated superior performance in image generation. Notable works (Wang et al. 2023a; Lin et al. 2024; Yu et al. 2024; Wu et al. 2024) have applied DMs to image super-resolution, achieving exceptional performance, particularly on complex natural scenes. Specifically, diffusion-based SR methods typically fall into two categories: the first group (Saharia et al. 2022; Rombach et al. 2022; Shang et al. 2024; Yue, Wang, and Loy 2024) involves injecting LR images directly into the diffusion model and train it from scratch, while the second group (Wang et al. 2023a; Lin et al. 2024; Yang et al. 2023; Yu et al. 2024; Wu et al. 2024), exploits the generative prior from pre-trained diffusion models, such as Stable Diffusion (SD) (Rombach et al. 2022; Podell et al. 2023), to enhance image super-resolution. Methods that are trained from scratch offer significant flexibility and ease of retraining following architectural modifications, making them ideal for lightweight applications. However, as shown in Fig. 1, these methods typically struggle to match the upper bound performance of prior-based methods, which benefit from the rich generative prior gained through extensive training on vast datasets over thousands of GPU days. A natural question is that can we develop a diffusion architecture trained from scratch while rivaling the performance of prior-based methods, balancing both performance and flexibility?

The advent of the Diffusion Transformer (DiT) (Peebles and Xie 2023) has made this idea feasible. This isotropic, full-transformer architecture, which maintains constant resolution and channel dimensions, shows remarkable performance and scalability, establishing a new paradigm in diffusion archi-

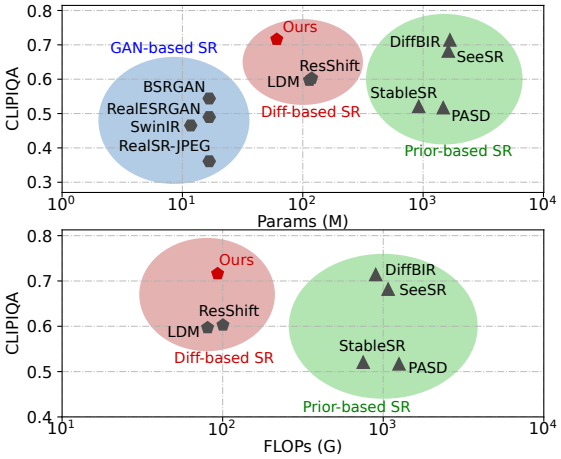


Figure 1: Comparisons of the proposed method with recent SR methods on the RealSR dataset. Top: CLIPIQA vs. Parameters. Bottom: CLIPIQA vs. FLOPs. Specifically, "Diff-based SR" refers to diffusion-based methods trained from scratch.

*These authors contributed equally.
†Corresponding author.

tecture design (Brooks et al. 2024; Esser et al. 2024; Li et al. 2024; Gao et al. 2024; Hatamizadeh et al. 2023). In contrast, early diffusion works (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021; Rombach et al. 2022) typically employed U-shaped denoiser architecture, which also remains popular in low-level tasks (Wang et al. 2022; Zamir et al. 2021) due to its hierarchical feature extraction capability and inductive bias conducive to denoising (Williams et al. 2024). In this paper, we propose a diffusion transformer model for image super-resolution, namely DiT-SR. Instead of applying the standard diffusion transformer architecture directly, DiT-SR is a U-shape encoder-decoder network, but with isotropic designs for all the transformer blocks at different stages. Specifically, DiT-SR adopts the U-shaped global structure with incrementally wider channel dimensions at deeper layers, which helps recover more image details at multi-scale resolutions. Besides, inspired by the observations that (1) The transformer architecture with same depth and channels could process tokens of different lengths well, *e.g.*, DiT-XL/2, DiT-XL/4 and DiT-XL/8. (2) High-resolution DiTs (*e.g.*, DiT/2) benefit more from scaling up than low-resolution DiTs (*e.g.*, DiT/8), thus we introduce the isotropic designs of DiT into the multi-scale framework. DiT-SR mandates the same channel number for all transformer modules in different stages, and set the channel number bigger than original setting of high resolution in U-Net, but much smaller than the low resolution. By allocating computational resource to critical layers, DiT-SR can greatly boost the capacity of the transformer architecture in multi-scale paradigms with the given computation budget.

Furthermore, we observe that DiT-based denoisers encounter a common issue related to the inefficient mechanism of time-step conditioning. As illustrated in Fig. 2, the diffusion-based SR model attends to different frequency components at distinct denoising phases. Consequently, there should be a direct correlation bewteen the time step and frequency. However, the widespreadly used Adaptive Layer Normalization (AdaLN), which modulates features solely in a channel-wise manner, does not effectively capture the unique temporal dynamics of the denoising process. To overcome this limitation, we propose Adaptive Frequency Modulation (AdaFM) module, conditioning on frequency domain. This highly efficient module, replacing AdaLN after each normalization layer, requires significantly few parameters while boosting performance. The time step adaptively reweights different frequency components, making it especially suitable for image super-resolution.

We summarize the primary contributions as follows:

- We propose DiT-SR, a diffusion transformer architecture specifically designed for image super-resolution, the first work that seamlessly combining the advantages of U-shaped and isotropic designs.
- We introduce an efficient yet effective frequency-wise time step conditioning module AdaFM, augmenting the diffusion model's ability to emphasize specific frequency information at varying time steps.
- Extensive experiments demonstrate that the proposed architecture outperforms existing training-from-scratch SR methods dramatically, and even supasses some prior-based methods with about only 5% of the parameters.

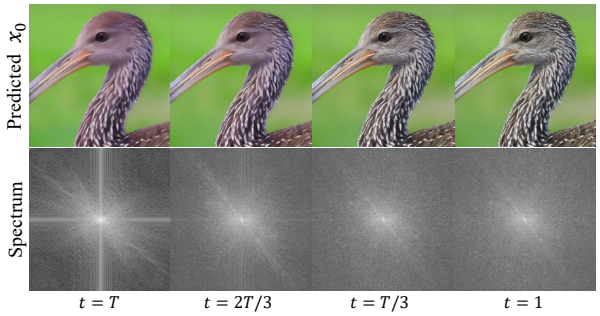


Figure 2: Analysis of images generated at different stages with a diffusion-based super-resolution model (Yue, Wang, and Loy 2024). The first row shows the predicted clean images at various steps, while the second row displays the Fourier spectrums of each predicted clean image. The diffusion model initially generates low-frequency components (center part of spectrums) and subsequently generates high-frequency components (peripheral part of spectrums).

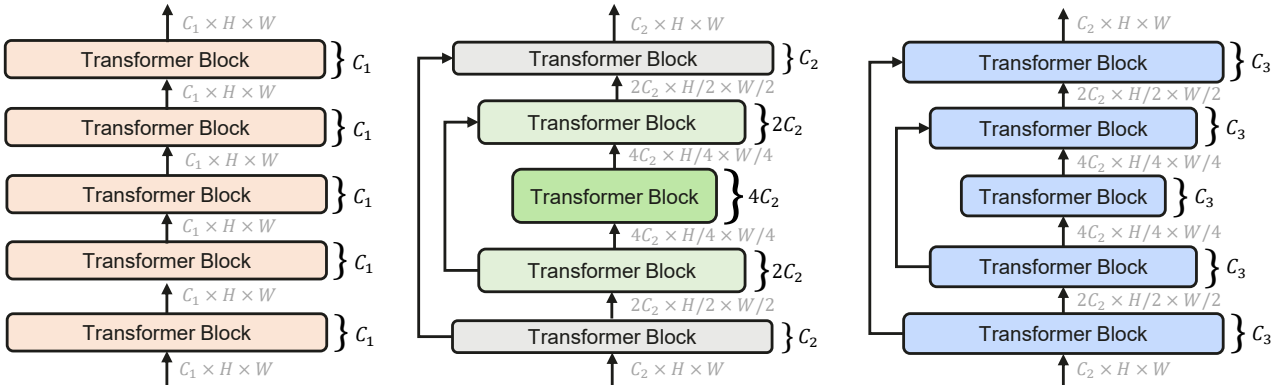
Related Works

Diffusion-based Image Super-Resolution

Recently, diffusion models (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021) have exhibited substantial benefits in image generation tasks, which generally fall into two categories: train-from-srcatch methods and prior-based methods. SR3 (Saharia et al. 2022) is the pioneer in introducing the diffusion model to the image super-resolution. LDM (Rombach et al. 2022) enhances efficiency by performing the diffusion process in latent space. ResShift (Yue, Wang, and Loy 2024) reformulate the diffusion process resulting a shortened markov chain that reduces the number of denoising steps to 15. These methods offer significant flexibility and ease of retraining following architectural modifications, making them ideal for lightweight applications. Inspired by the remarkable potential of Stable Diffusion (Rombach et al. 2022; Podell et al. 2023) for text-to-image tasks, several methods (Wang et al. 2023a; Lin et al. 2024; Yang et al. 2023; Wu et al. 2024) have exploited its generative prior to guide real-world image super-resolution. While these prior-based methods yield remarkable results, their deployment is limited by slow inference speeds, which arise from the redundant denoiser architecture and the multi-step iterative denoising process. Despite SinSR (Wang et al. 2023b) and AddSR (Xie et al. 2024) employing knowledge distillation for one-step denoising, their diffusion architectures typically cannot be altered without massive retraining. Orthogonal to the efforts to reduce denoising steps, we concentrate on developing an effective diffusion architecture trained from scratch while rivaling the performance of prior-based methods.

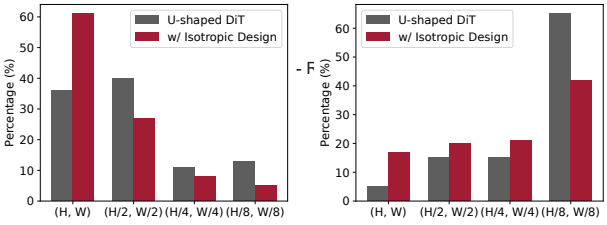
Diffusion Model Architecture

Previous diffusion studies (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Dhariwal and Nichol 2021; Nichol and Dhariwal 2021; Rombach et al. 2022) have predominantly utilized the U-Net (Ronneberger, Fischer, and Brox 2015) architecture for denoising, incorporating ele-



(a) Standard DiT (b) U-shaped DiT (c) Ours

Figure 3: The comparision from the standard DiT to the proposed DiT-SR. (a): The standard DiT. (b):U-shaped DiT, incorporating downsampling and upsampling to standard DiT and increasing the channel dimension in deep layers. (c): The proposed DiT-SR. This architecture employs a U-shaped global structure, yet maintains same inside channel dimension for all transformer blocks in different stages, allocating computational resource to high-resolution layers ($4C_2 > C_3 > C_2$) to boost the model capacity. Gray $C_i \times H \times W$ indicate the feature shape between layers, and the black C_i indicates the number of feature channels inside a layer.



(a) FLOPs (b) Parameters

Figure 4: The percentage of FLOPs and parameters for each stage of the U-shaped DiT, both with and without isotropic design, show that more computational resources are allocated to the high-resolution stages.

ments such as ResBlocks (He et al. 2016) and Transformer blocks (Vaswani et al. 2017). DiT (Peebles and Xie 2023) marks a departure from the U-shaped design by adopting an isotropic full transformer architecture, which showcases enhanced scalability. Subsequent works (Ma et al. 2024; Gao et al. 2023; Lu et al. 2024; Li et al. 2024; Gao et al. 2024; Hatamizadeh et al. 2023) have adopted the standard DiT architecture and shown superior performance across various tasks. U-ViT (Bao et al. 2023) retains the long skip connections typical of U-Net but does not include upsampling or downsampling operations. Our proposed DiT architecture, which merges U-shaped and isotropic designs, achieves remarkable performance on image super-resolution.

Preliminaries

Diffusion Models

Given a LR image $\boldsymbol{y}$ and its corresponding HR image $\boldsymbol{x}_0$, diffusion-based SR methods strive to model the conditional distribution $q(\boldsymbol{x}_0|\boldsymbol{y})$. Typically, these methods define a T-step forward process that gradually introduce random noise to $\boldsymbol{x}_0$, which can be succinctly achieved in one step through the reparameterization trick:

$$q(\boldsymbol{x_t}|\boldsymbol{x_0}) = \mathcal{N}\left(\boldsymbol{x_t}; \sqrt{\bar{\alpha}_t}\boldsymbol{x_0}, (1-\bar{\alpha}_t)\boldsymbol{I}\right) \text{ with } \bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i, \tag{1}$$

where x_t denotes the noised image at time-step t and α_t is the predefined variance schedule. During the reverse process, the model starts from pure Gaussian noise and iteratively generates the preceding state x_{t-1} from x_t using the approximated posterior distribution:

$$p_\theta(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}, \boldsymbol{y_0}) = \mathcal{N}\left(\boldsymbol{\mu_\theta}(\boldsymbol{x_t}, \boldsymbol{y_0}, t), \Sigma(\boldsymbol{x_t}, t)\right), \tag{2}$$

where $\Sigma(x_t, t)$ is a constant that depends on α_t, and $\mu_\theta(x_t, y_0, t)$ is parameterized by a denoiser $\epsilon_\theta(x_t, y_0, t)$.

Residual Shifting

ResShift (Yue, Wang, and Loy 2024) constructs a Markov chain bewteen HR and LR images rather than pure Gaussian noise. Let $\boldsymbol{e}_0 = \boldsymbol{y}_0 - \boldsymbol{x}_0$ represents the resduial between the LR and HR images. Additionally, a shifting sequence $\{\eta_t\}_{t=1}^T$ is introduced, gradually increasing from $\eta_1 \to 0$ to $\eta_T \to 1$ with each timeste. The forward process is then formulated based on this shifting sequence:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}_0 + \eta_t\boldsymbol{e}_0, \kappa^2\eta_t\boldsymbol{I}),\ t = 1, 2, \cdots, T, \tag{3}$$

where $\alpha_t = \eta_t - \eta_{t-1}$ for $t > 1$ and $\alpha_1 = \eta_1$. The hyperparameter κ controls the noise variance. The denoising process, $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}_0)$ is formulated as follows:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}\middle|\frac{\eta_{t-1}}{\eta_t}\boldsymbol{x}_t + \frac{\alpha_t}{\eta_t}f_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t), \kappa^2\frac{\eta_{t-1}}{\eta_t}\alpha_t\boldsymbol{I}\right), \tag{4}$$

where $\boldsymbol{x}_0$ is directly predicted by the denoiser $f_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t)$. This well-designed transfer distribution for image super-resolution effectively reduces the length of Markov chains, thereby reducing the number of required time steps. We follow this paradigm to train our diffusion model.

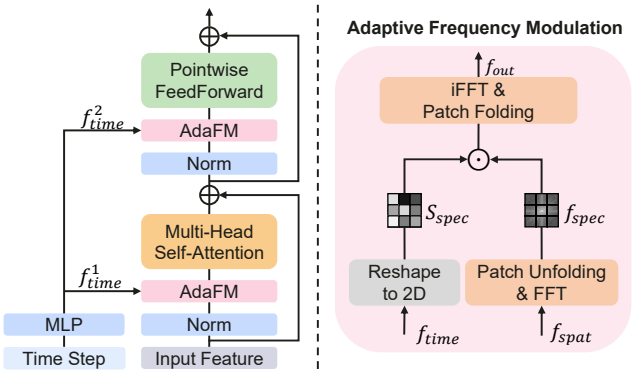


Figure 5: The illustration of transformer block in DiT-SR and Adaptive Frequency Modulation (AdaFM). AdaFM injects the time step to frequency domain and adaptively reweights different frequency components.

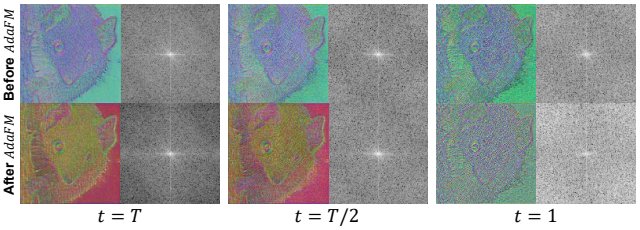


Figure 6: Visualization of the feature maps and their corresponding spectrums before and after applying AdaFM. AdaFM enhances the low frequency components in the early stages of denoising (peripheral part of spectrum getting darker) and the high frequency components in the later stages (peripheral part of spectrum getting brighter), thereby augmenting the diffusion model's ability to emphasize specific frequency at different time steps.

Methodology

Overall Architecture

The proposed DiT-SR, depicted in Fig. 3, aims to be trained from scratch to potentially rival the performance of prior-based methods. This denoiser architecture features a U-shaped encoder-decoder global framework, but with an isotropic design for all the transformer blocks at different stages. It includes several transformer stages in both the encoder and decoder, each with varying feature resolutions. Within each stage, multiple transformer blocks with uniform configurations are employed, reallocating computational resources to high-resolution layers to enhance the transformer architecture's capacity.

The LR image $\boldsymbol{y}$ and noisy image $\boldsymbol{x}_t$ are concatenated along the channel dimension, and together with the time step t, serve as inputs to the denoiser, which predicts $\boldsymbol{x}_0$ and iteratively refines it as outlined in Eq. 4. As shown in Fig. 5, the transformer block consists of a multi-head self-attention (MHSA) mechanism (Liu et al. 2021) that operates as a spatial mixer, and a multi-layer perceptron (MLP) with two fully-connected layers separated by GELU activation, serving as channel mixers. Considering the high computational cost and memory constraints of global self-attention when processing high-resolution inputs, we employ local attention with window shifting as an alternative to the original self-attention (Vaswani et al. 2017). Group normalization layers are applied before both the MHSA and MLP. Additionally, the proposed Adaptive Frequency Modulation (AdaFM) is integrated following each normalization layer to inject the time step. Our transformer block can be formulated as:

$$\begin{aligned} &f^1_{time}, f^2_{time} = \mathrm{MLP_t}(t), \\ &X = \mathrm{MHSA}(\mathrm{AdaFM}(\mathrm{Norm}(X), f^1_{time})) + X, \\ &X = \mathrm{MLP}(\mathrm{AdaFM}(\mathrm{Norm}(X), f^2_{time})) + X. \end{aligned} \tag{5}$$

Subsequent sections will elaborate the design motivation and specific details of DiT-SR, including the integration of U-shaped global architecture and isotropic block design, as well as the frequency-adaptive time-step conditioning mechanism.

Isotropic Design in U-shaped DiT

The U-Net architecture (Ronneberger, Fischer, and Brox 2015), with its encoder-decoder framework, is a popular choice for image generation and restoration tasks. Given the U-shaped architecture's multi-scale feature extraction ability, we propose integrating the U-shaped global architecture into standard DiT to enhance its performance in image super-resolution. The encoder progressively reduces the resolutions of feature maps while increasing their channel dimensions, and the decoder reverses these operations for reconstruction.

We rethink the isotropic design in DiT, and identify two notable characteristics. Firstly, DiTs with consistent channel and depth could effectively handle input with varying patch sizes (*e.g.*, DiT-XL/2, DiT-XL/4, and DiT-XL/8), which is analogous to processing different resolutions in U-Net. Secondly, DiTs at higher resolutions (*e.g.*, DiT/2) benefit more from scaling up compared to those at lower resolutions (*e.g.*, DiT/8). Motivated by these insights, we introduce this straightforward yet effective isotropic design to multi-scale U-shaped DiT in a pioneering way, strategically allocating more computational resources to critical high resolution layers. Specifically, each transformer stage consists of several transformer blocks that operate at the same resolution. Within each stage, we standardize the inside feature's channel dimension to be same across all stages, perform all the transformer blocks in reallocated feature space, and then reassemble them to their original dimensions. Considering that high-resolution stages capture more high-frequency details, which are crucial for image super-resolution and exhibit better scalability, we set the standardized channel dimension larger than the original high-resolution stages in U-Net, yet considerably smaller than low-resolution stages. As depicted in Fig. 4, this isotropic principle allocates computational resources to critical high-resolution layers, avoiding search-based optimizations, greatly boost the capacity of transformer architecture with far fewer parameters than conventional U-Net.

Frequency-Adaptive Time Step Conditioning

Since the diffusion model utilizes the same denoiser across various time steps, it is crucial to explicitly incorporate time

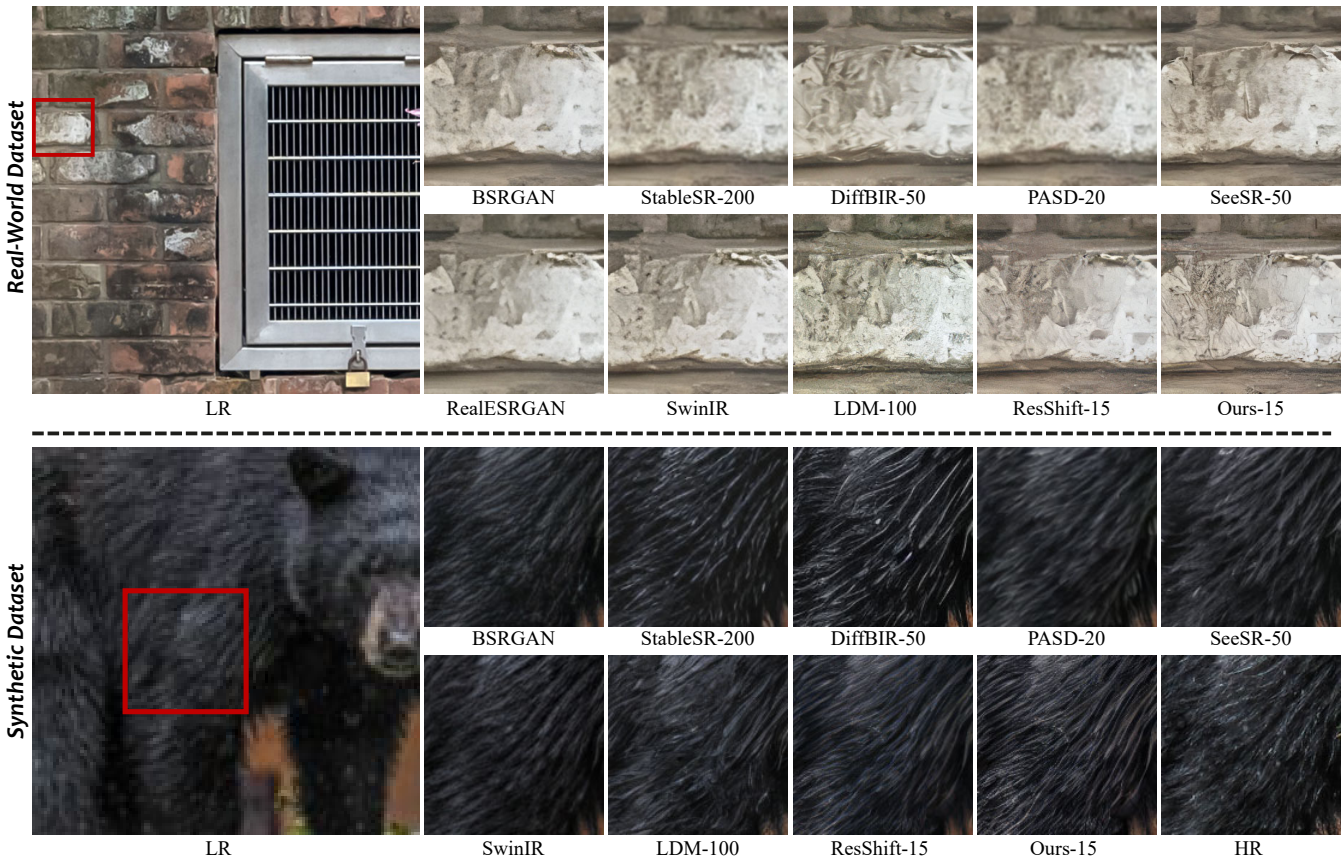


Figure 7: Qualitative comparisons of different methods on both synthetic and real-world datasets.

step as a condition. Adaptive Layer Normalization (AdaLN), first introduced in DiT (Peebles and Xie 2023), has been proven effective on image generation and is widely adopted in subsequent DiT-based models. Nevertheless, unlike image generation task, which starts from pure noise and focuses primarily on semantics, SR task emphasizes the recovery of high-frequency details, necessitating the diffusion model to possess strong frequency perception capabilities.

Our investigation into the temporal evolution of images predicted by the diffusion-based super-resolution model reveals that it focuses on various frequency components at different denoising stages. As shown in Fig. 2, the model initially reconstructs low-frequency elements, corresponding to the image structure, and progressively refines high-frequency details, associated with texture. Consequently, the time step should adaptively modulate different frequency components, using distinct modulation parameters for high and low-frequency regions. However, AdaLN modulates feature maps exclusively in the channel dimension, applying uniform modulation parameters across all spatial locations. This limitation hinders its ability to effectively address the specific frequency requirements of image super-resolution tasks. Moreover, it is challenging to generate modulation spatial-wise parameters from a one-dimensional time-step vector, as it requires adaptively distinguishing between the high and low-frequency components' spatial positions in the input image.

To solve this challenge, we introduce Adaptive Frequency Modulation (AdaFM), replacing AdaLN after each normalization layer and switching the time-step modulation from the spatial domain to the frequency domain, as shown in Fig. 5. Initially, to accommodate various input resolutions and enhance efficiency, we segment the spatial domain feature map $f_{spat} \in \mathbb{R}^{C\times H\times W}$ into $p \times p$ windows. Subsequently, we transform these segments into spectrograms $f_{spec} \in \mathbb{R}^{\frac{H\times W}{p^2}\times C\times p\times p}$ using the Fast Fourier Transform within each window. The time step is mapped to a p^2-dimensional vector f_{time} and reshaped into a frequency scale matrix $S_{spec} \in \mathbb{R}^{p\times p}$, which is then used to adaptively reweight various frequency components, thereby augmenting the diffusion model's ability to emphasize specific frequency at different time steps, as illustrated in Fig. 6. No explicit constraints are imposed on the output of AdaFM. Instead, this adaptivity is learned in an end-to-end manner.

In a spectrum, each pixel at a specific spatial position corresponds to a predetermined frequency components, defined solely by the feature map's spatial dimension, independent of its content. The frequency corresponding to a pixel located at spatial position (u, v) in spectrum $\in \mathbb{R}^{H\times W}$ can be

Methods	#Params	RealSR			RealSet65		
		CLIPIQA↑	MUSIQ↑	MANIQA↑	CLIPIQA↑	MUSIQ↑	MANIQA↑
GAN based Methods							
RealSR-JPEG	17M	0.3611	36.068	0.1772	0.5278	50.5394	0.2943
BSRGAN	17M	0.5438	63.5819	0.3685	0.616	65.5774	0.3897
RealESRGAN	17M	0.4898	59.6766	0.3679	0.5987	63.2228	0.3871
SwinIR	12M	0.4653	59.6316	0.3454	0.5778	63.8212	0.3816
Prior based Methods							
StableSR-200	919M	0.5207	59.4264	0.3563	0.5338	56.9207	0.3387
DiffBIR-50	1670M	<u>0.7142</u>	**66.843**	0.4802	**0.7398**	<u>69.7260</u>	<u>0.5000</u>
PASD-20	1469M	0.5170	58.4394	0.3682	0.5731	61.8813	0.3893
SeeSR-50	1619M	0.6819	<u>66.3461</u>	**0.5035**	0.7030	**68.9803**	**0.5084**
Training-from-Scratch Diff. based Methods							
LDM-100	114M	0.5969	55.4359	0.3071	0.5936	56.112	0.356
ResShift-15	119M	0.6028	58.8790	0.3891	0.6376	58.0400	0.4048
Ours-15	61M	**0.7161**	65.8334	<u>0.5022</u>	<u>0.7120</u>	66.7413	0.4821

Table 1: Performance and denoiser complexity comparison on real-world datasets. The best and second best results are highlighted in **bold** and <u>underline</u>. We denote the number of sampling steps for each difusion-based method using the format "method-steps".

formulated as:

$$f_u = \frac{u - H/2}{H} \times F_s, \quad f_v = \frac{v - W/2}{H} \times F_s, \quad (6)$$

where f_u, f_v denote the vertical and horizontal frequencies separately, and F_s indicates sampling frequency. This consistency allows the same frequency scale matrix S_{spec} to be applied across all windows and channels, significantly enhancing efficiency. In comparison to AdaLN, which requires $dim_{f_{time}} \times C \times 3 \times 2$ mapping parameters ($scale$, $shift$ and $gate$ for both self-attention and MLPs), AdaFM requires only $dim_{f_{time}} \times p^2 \times 2$. The process is formulated as follows:

$$\begin{aligned} S_{spec} &= \text{reshape}(f_{time}, p \times p), \\ f_{spec} &= \text{FFT}\left(\mathcal{P}\left(f_{spat}\right)\right), \\ f'_{spec} &= S_{spec} \odot f_{spec}, \\ f_{out} &= \mathcal{P}^{-1}\left(\text{iFFT}\left(f'_{spec}\right)\right), \end{aligned} \quad (7)$$

where $\mathcal{P}$ and $\mathcal{P}^{-1}$ denote the patch unfolding and folding operations, $\odot$ represents element-wise multiplication, FFT and iFFT indicate Fast Fourier Transform and inverse Fourier Transform. Given that different frequencies correspond to distinct spatial locations on the feature map, the proposed frequency-wise time-step conditioning module actually provides spatial-wise modulation.

Experiments

Experimental Settings

Datasets. We evaluate the proposed model on $\times 4$ real-world SR task. The training data comprises LSDIR (Li et al. 2023), DIV2K (Agustsson and Timofte 2017), DIV8K (Gu et al. 2019), OutdoorSceneTraining (Wang et al. 2018), Flicker2K (Timofte et al. 2017) and the first 10K face images from FFHQ (Karras, Laine, and Aila 2019) datasets. We partition LSDIR into a training set with 82991 images and a test set with 2000 images. Following LDM (Rombach et al. 2022), HR images in our training set are randomly cropped to 256×256 and the degradation pipeline of RealESRGAN (Wang et al. 2021) is used to synthesize LR/HR pairs. The test set images are center-cropped to 512×512 and subjected to the same degradation pipeline used in the training stage to create a synthetic dataset, named LSDIR-Test. Furthermore, we utilize two real-world datasets: RealSR (Cai et al. 2019), which comprises 100 real images captured by Canon 5D3 and Nikon D810 cameras, and RealSet65 (Yue, Wang, and Loy 2024), including 65 low-resolution images collected from widely used datasets and the internet.

Implementation Details. Following LDM (Rombach et al. 2022), the proposed architecture operates in latent space, utilizing the Vector Quantized GAN (VQGAN) (Esser, Rombach, and Ommer 2021) with a downsample factor of 4. We train the proposed model for $300K$ iterations with a batch size of 64 using 8 NVIDIA Tesla V100 GPUs. The optimizer is Adam (Kingma and Ba 2014), and the learning rate is $5e^{-5}$. The FFT window size p is empirically set to 8 (Wallace 1991; Kong et al. 2023). Detailed architectural configurations are provided in the supplementary material.

Evaluation Metrics. We adopt reference-based metrics, including PSNR and LPIPS (Zhang et al. 2018), to evaluate the performance of different models. Additionally, non-reference metrics such as CLIPIQA (Wang, Chan, and Loy 2023), MUSIQ (Ke et al. 2021), and MANIQA (Yang et al. 2022), which are more consistent with human perception in generative SR, are also employed. For assessments on real-world datasets, due to the lack of ground truth, we evaluate their performance using only non-reference metrics.

Comparison with State-of-the-Arts

Comparison Methods. We compared our proposed architecture with several latest SR methods, including GAN-based methods such as RealSR-JPEG(Ji et al. 2020), BSR-

Configuration		#Params	FLOPs	RealSR		RealSet65	
DiT Arch.	Time Conditioning			CLIPIQA↑	MUSIQ↑	CLIPIQA↑	MUSIQ↑
Isotropic	AdaLN	42.38M	122.99G	0.655	64.194	0.664	64.263
U-shape	AdaLN	264.39M	122.87G	0.688	64.062	0.693	65.604
Ours	AdaLN	100.64M(-62%)	93.11G(-24%)	0.700	64.676	0.699	**67.634**
Ours	AdaFM	60.79M(-77%)	93.03G(-24%)	**0.716**	**65.833**	**0.712**	66.741

Table 2: Ablation Study on real-world datasets. The percentage reductions in the number of parameters and FLOPs are compared to the U-shaped DiT. The best results are highlighted in **bold**.

Methods	LSDIR-Test				
	PSNR↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
GAN based Methods					
RealSR-JPEG	22.16	0.360	0.546	59.02	0.342
BSRGAN	<u>23.74</u>	0.274	0.570	67.94	0.394
RealESRGAN	23.15	0.259	0.568	68.23	0.414
SwinIR	23.17	<u>0.247</u>	0.598	68.20	0.414
Prior based Methods					
StableSR-200	22.68	0.267	0.660	68.91	0.416
DiffBIR-50	22.84	0.274	<u>0.709</u>	<u>70.05</u>	0.455
PASD-20	23.57	0.279	0.624	69.07	0.440
SeeSR-50	22.90	0.251	**0.718**	**72.47**	**0.559**
Training-from-Scratch Diff. based Methods					
LDM-100	23.34	0.255	0.601	66.84	0.413
ResShift-15	**23.83**	<u>0.247</u>	0.640	67.74	0.464
Ours-15	23.60	**0.244**	0.646	69.32	<u>0.483</u>

Table 3: Performance comparison on the synthetic LSDIR-Test dataset. The best and second best results are highlighted in **bold** and <u>underline</u>.

GAN(Zhang et al. 2021), RealESRGAN(Wang et al. 2021), and SwinIR(Liang et al. 2021), as well as diffusion-based methods like LDM(Rombach et al. 2022), StableSR(Wang et al. 2023a), ResShift(Yue, Wang, and Loy 2024), DiffBIR (Lin et al. 2024), PASD (Yang et al. 2023) and (Wu et al. 2024). The steps are configured using their default settings. It is worth noting that StableSR, DiffBIR, PASD and SeeSR leverage the generative prior of Stable Diffusion, which is pretrained on large-scale datasets for thousands of GPU days, while LDM and ResShift are trained from scratch like ours.

Comparison on Real-World and Synthetic Datasets. We present the qualitative and quantitative results on Fig. 7, Tab. **??** and Tab. 3. The proposed architecture significantly outperforms existing training-from-scratch methods, and even supasses competitive with state-of-the-art prior-based methods, while utilizing only about 5% of their parameters.

Ablation Study

U-shaped DiT with Isotropic Design. As described above, this paper proposes an evolutionary path from standard DiT to U-shaped DiT, and ultimately introduce isotropic design to multi-scale U-shaped DiT. We reimplement DiT for super-resolution, employing local attention with window shifting to replace the original self-attention. As shown in Table 2, the U-shaped DiT outperforms the standard DiT for the same FLOPs, but having six times more parameters. Notably, by reallocating computational resources to critical layers within the isotropic design, performance is improved even with a 62% reduction in parameters.

Adaptive-Frequency Modulation. The proposed AdaFM operates in the frequency domain, adaptively identifying high and low frequency regions and modulates them separately with distinct parameters. Additionally, due to the nature of the frequency domain and our highly efficient design, the parameter count of AdaFM is only a fraction of that of AdaLN. As shown in Tab. 2, replacing AdaLN with AdaFM reduced the number of denoiser parameters from 100.64M to 60.79M, while also enhancing model performance, demonstrating the effectiveness of AdaFM. Fig. 6 visualizes the feature maps and there spectrums before and after AdaFM, illustrating how it adaptively enhances low-frequency components in the early stages of denoising and high-frequency components in the later stages, thereby establishing a correlation between the time step and frequency.

Discussion and Conclusion

In this work, we introduce DiT-SR, an effective diffusion transformer architerture for image super-resolution that can be trained from scratch to rival the performance of prior-based methods. It integrates U-shaped global architecture and isotropic block designs, reallocating the computational resources to critical high-resolution layers, boosting the performance efficiently. Furthermore, we propose an efficient yet effective time-step conditioning module AdaFM that adaptively reweights different frequency components, augmenting the diffusion model's ability to emphasize specific frequency information at varying time steps.

Future Work. AdaFM holds the potential to establish a new time-step conditioning paradigm for diffusion models, extending its application to various low-level visual tasks and even to text-to-image generation that also adhere to the paradigm of initially generating low frequencies followed by high frequencies.

Limitation and Ethical Statement. Although the proposed denoiser achieves competitive performance with much fewer parameters compared to prior-based models, it still has some way to go before fully surpassing their performance, particularly in terms of semantic structure due to the lack of priors. In addition, Similar to other content generation methods, our approach must be used cautiously to prevent potential misuse.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62441601 and U22A2096, in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23042, in part by the Innovation Fund of Xidian University under Grant YJSJ24017.

References

Agustsson, E.; and Timofte, R. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Bao, F.; Nie, S.; Xue, K.; Cao, Y.; Li, C.; Su, H.; and Zhu, J. 2023. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22669–22679.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

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

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Gao, P.; Zhuo, L.; Lin, Z.; Liu, C.; Chen, J.; Du, R.; Xie, E.; Luo, X.; Qiu, L.; Zhang, Y.; et al. 2024. Lumina-T2X: Transforming Text into Any Modality, Resolution, and Duration via Flow-based Large Diffusion Transformers. *arXiv preprint arXiv:2405.05945*.

Gao, S.; Zhou, P.; Cheng, M.-M.; and Yan, S. 2023. MDTv2: Masked Diffusion Transformer is a Strong Image Synthesizer. *arXiv preprint arXiv:2303.14389*.

Gu, S.; Lugmayr, A.; Danelljan, M.; Fritsche, M.; Lamour, J.; and Timofte, R. 2019. Div8k: Diverse 8k resolution image dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 3512–3516. IEEE.

Hatamizadeh, A.; Song, J.; Liu, G.; Kautz, J.; and Vahdat, A. 2023. Diffit: Diffusion vision transformers for image generation. *arXiv preprint arXiv:2312.02139*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ji, X.; Cao, Y.; Tai, Y.; Wang, C.; Li, J.; and Huang, F. 2020. Real-world super-resolution via kernel estimation and noise injection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 466–467.

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

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

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kong, L.; Dong, J.; Ge, J.; Li, M.; and Pan, J. 2023. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5886–5895.

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

Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; et al. 2024. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. *arXiv preprint arXiv:2405.08748*.

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

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Ouyang, W.; Qiao, Y.; and Dong, C. 2024. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. arXiv:2308.15070.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lu, Z.; Wang, Z.; Huang, D.; Wu, C.; Liu, X.; Ouyang, W.; and Bai, L. 2024. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*.

Ma, N.; Goldstein, M.; Albergo, M. S.; Boffi, N. M.; Vanden-Eijnden, E.; and Xie, S. 2024. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

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

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726.

Shang, S.; Shan, Z.; Liu, G.; Wang, L.; Wang, X.; Zhang, Z.; and Zhang, J. 2024. Resdiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8975–8983.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; and Zhang, L. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 114–125.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wallace, G. K. 1991. The JPEG still picture compression standard. *Communications of the ACM*, 34(4): 30–44.

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

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023a. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*.

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

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 606–615.

Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A. C.; and Wen, B. 2023b. SinSR: Diffusion-Based Image Super-Resolution in a Single Step. *arXiv preprint arXiv:2311.14760*.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17683–17693.

Williams, C.; Falck, F.; Deligiannidis, G.; Holmes, C. C.; Doucet, A.; and Syed, S. 2024. A unified framework for U-Net design and analysis. *Advances in Neural Information Processing Systems*, 36.

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

Xie, R.; Tai, Y.; Zhang, K.; Zhang, Z.; Zhou, J.; and Yang, J. 2024. AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation. *arXiv preprint arXiv:2404.01717*.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1191–1200.

Yang, T.; Wu, R.; Ren, P.; Xie, X.; and Zhang, L. 2023. Pixel-Aware Stable Diffusion for Realistic Image Super-Resolution and Personalized Stylization. In *arXiv:2308.14469v3*.

Yu, F.; Gu, J.; Li, Z.; Hu, J.; Kong, X.; Wang, X.; He, J.; Qiao, Y.; and Dong, C. 2024. Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild. arXiv:2401.13627.

Yue, Z.; Wang, J.; and Loy, C. C. 2024. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14821–14831.

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

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