A Timestep-Adaptive Frequency-Enhancement Framework for Diffusion-based Image Super-Resolution

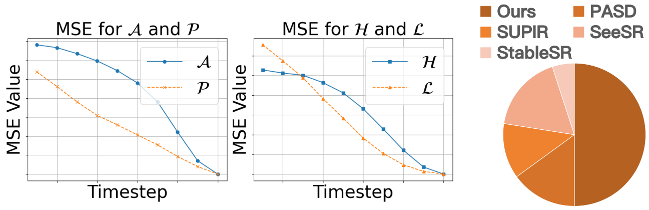
Yueying Li[1,2] , **Hanbin Zhao**[3] , **Jiaqing Zhou**[4] , **Guozhi Xu**[4] ,
Tianlei Hu[2,3] , **Gang Chen**[2,3] and **Haobo Wang**[1,2*]
[1]School of Software Technology, Zhejiang University
[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3]College of Computer Science and Technology, Zhejiang University
[4]ByteDance, Hangzhou
{liyueying, zhaohanbin, htl, cg, wanghaobo}@zju.edu.cn, {jiashu, shuzhi}@bytedance.com

Abstract

Image super-resolution (ISR) is a classic and challenging problem in computer vision because of complex and unknown degradation patterns in the data collection process. Leveraging powerful generative priors, diffusion-based methods have recently established new state-of-the-art ISR performance, but their characteristics in the frequency domain are still underexplored. In this paper, we innovatively investigate their frequency-domain behaviors from a sampling timestep perspective. Experimentally, we find that current diffusion-based ISR algorithms exhibit insufficiency in different frequency components in distinct groups of timesteps during the sampling. To address this, we first propose a Timestep Division Controller that is able to adaptively divide the timesteps into groups based on the performance gradient across different components. Next, we design two dedicated modules — the Amplitude and Phase Enhancement Module (APEM) and the High- and Low-Frequency Enhancement Module (HLEM), to regulate the information flow of distinct frequency-domain features. By adaptively enhancing specific frequency components at different stages of the sampling process, the two modules effectively compensate for the insufficient frequency-domain perception of diffusion-based ISR models. Extensive experiments on three benchmark datasets verify the superior ISR performance of our method, e.g., achieving an average **5.40%** improvement on CLIP-IQA compared to the best diffusion-based ISR baseline.

1 Introduction

Image super-resolution (ISR) is a fundamental task in low-level vision that aims to reconstruct high-resolution (HR) images from their low-resolution (LR) counterparts. It has widespread applications in areas such as medical imaging [Li *et al.*, 2024], satellite imagery [Liu *et al.*, 2017], and surveillance systems [Shermeyer and Etten, 2019], where obtaining high-quality images can naturally be subject to hardware limitations and transmission losses. Early ISR methods [Dong *et al.*, 2016a; Tai *et al.*, 2017] attempt to construct synthetic image pairs through simple handcrafted degradation operations (e.g., bicubic downsampling). However, they fail to generalize well in realistic scenarios since real-world LR images typically involve complex and unknown degradation patterns.



(a) **The visualization of numerical metrics.** (b) **User study.**

Figure 1: (a) illustrates the variation of Mean Squared Error (MSE) for different frequency-domain components (amplitude $\mathcal{A}$, phase $\mathcal{P}$, high-frequency $\mathcal{H}$, and low-frequency $\mathcal{L}$) with respect to timesteps. (b) exhibits the result of user study among existing SOTA methods.

To address this problem, some studies employ diverse network architectures, such as CNN-based [Dong *et al.*, 2016a] and GAN-based methods [Zhang *et al.*, 2021; Wang *et al.*, 2021]. Among them, recent diffusion model-based ISR methods [Wang *et al.*, 2024b; Lin *et al.*, 2023; Yu *et al.*, 2024; Wu *et al.*, 2024; Yang *et al.*, 2024] have gained great attention, which exhibit superior performance levels in image quality assessment (IQA) metrics while maintaining competitive full-reference results. Another mainline of ISR researches [Guan *et al.*, 2024; Li *et al.*, 2023a; Xu *et al.*, 2024; Xie *et al.*, 2021] approaches image restoration from a frequency-domain perspective, aiming to model the relationship with degradation processes for targeted image recovery. But, current frequency-domain enhancement algorithms are predominantly applied to GANs [Fu *et al.*, 2021; Xu *et al.*, 2024] and traditional CNN models [Yu *et al.*, 2022; Huang *et al.*, 2022; Guo *et al.*, 2022]. Only a few methods [Wang *et al.*, 2024c; Moser *et al.*, 2024; Zhao *et al.*, 2024; Luo *et al.*, 2023] explore the potential of diffusion-based models in the frequency domain. Overall, these methods either rely solely on a single frequency decomposition ap-

[*]Corresponding author.

proach [Wang *et al.*, 2024c], or apply frequency information merely during image processing [Moser *et al.*, 2024; Zhao *et al.*, 2024]. For instance, FourierDiff [Lv *et al.*, 2024] merely leverages the amplitude component of the generative prior to ensure enhanced brightness aligns with natural image distributions. WF-Diff [Zhao *et al.*, 2024] employs a wavelet-based dual-branch diffusion framework to refine high-frequency and low-frequency components of the initial enhanced images. To date, the frequency-domain property of the current diffusion methods remains unexplored.

In this paper, we investigate the frequency-domain behavior of diffusion-based ISR models from a novel sampling timestep perspective. More concretely, we are interested in the impact of different frequency-domain components obtained via frequency spectral and band decomposition, on the restoration process as low-resolution images are progressively recovered through diffusion model timesteps. Through our preliminary experiments, we have two main findings: **(1) Impact of Amplitude and Phase Components.** To investigate the variation of frequency spectral components across timesteps, we decompose the amplitude and phase components. By fusing one of the components at the current timestep with the corresponding component of the HR images (see top two rows, Figure 2), we observe that in the early timesteps, containing the phase of current timesteps significantly reduces MSE, while in the later timesteps, containing the amplitude drive more MSE change. This indicates that during the restoration process, the diffusion model first generates the phase component (structural information), and later generates the amplitude component (overall visual information). **(2) Impact of High- and Low-Frequency Components.** Next, to investigate the variation of frequency band components, we decompose the images into high- and low-frequency components and perform a similar fusion procedure (see bottom two rows, Figure 2). The experimental results show that in the early timesteps, containing the low-frequency component significantly reduces the MSE, while in the later timesteps, containing the high-frequency has a greater impact. This indicates that the diffusion model initially reconstructs the low-frequency component with global structures, and later generates high-frequency components with fine details, consistent with its generation mechanism.

Based on this, we believe that enhancing the diffusion model's frequency-domain perception at the appropriate time can improve ISR performance effectively. Therefore, we propose a **T**imestep-adaptive **F**requency-aware **D**iffusion framework for **S**uper-**R**esolution (dubbed TFDSR), which enhances the diffusion ISR models's different frequency components at adaptive timesteps. It comprises three core components: **(1) Timestep Division Controller (TDC)**: This module dynamically determines the enhancement frequency components across different timesteps, strategically selecting phase and low-frequency components for the early stage, while emphasizing amplitude and high-frequency components for later. **(2) Amplitude-Phase Enhancement Module (APEM)**: This module adaptively enhances the missing amplitude and phase components through a channel attention mechanism, thereby optimizing the representation of each frequency component. **(3) High-Low Frequency Enhancement Module (HLEM)**: This module operates on skip connection and adaptively enhances high- and low-frequency information in the skip features using a frequency band modulation. In terms of sampling strategy, TFDSR adaptively enhances different frequency components at different timesteps, requiring only minimal computational resources and time for efficient fine-tuning. Extensive experiments demonstrate that TFDSR significantly outperforms state-of-the-art diffusion models, especially on no-reference metrics, e.g., average $+\mathbf{5.40}\%$ in CLIP-IQA. Our source code and appendix are available at https://github.com/liyueying233/TFDSR.

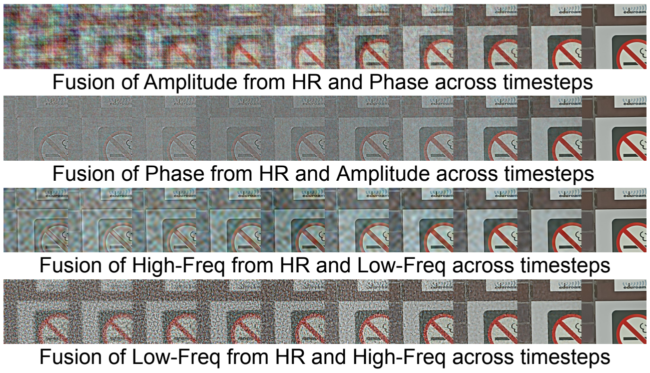

Fusion of Amplitude from HR and Phase across timesteps
Fusion of Phase from HR and Amplitude across timesteps
Fusion of High-Freq from HR and Low-Freq across timesteps
Fusion of Low-Freq from HR and High-Freq across timesteps

Figure 2: Denoising process of each frequency component. Top two rows: amplitude ($\mathcal{A}$) and phase ($\mathcal{P}$) combined with the corresponding frequency components of the HR image. Bottom rows: low- ($\mathcal{L}$) and high-frequency ($\mathcal{H}$) components with their relative HR components after the inverse Fourier Transform, aligned with each timestep. It demonstrates that early timesteps show greater variation in $\mathcal{P}$ and $\mathcal{L}$, while $\mathcal{A}$ and $\mathcal{H}$ change more in later timesteps.

2 Related Work

Image Super-Resolution (ISR). Although deep learning-based ISR techniques have gained widespread adoption, most CNN-based methods [Dong *et al.*, 2016a; Lim *et al.*, 2017; Kim *et al.*, 2016; Dong *et al.*, 2016b; Shi *et al.*, 2016] still suffer from the issue of excessive detail smoothing. To better enhance visual perception, some advances [Zhang *et al.*, 2021; Wang *et al.*, 2021; Liang *et al.*, 2021; Chen *et al.*, 2022; Liang *et al.*, 2022; Wang *et al.*, 2024a] using the GAN-based models in Real-ISR have explored more complex degradation models. For instance, BSRGAN [Zhang *et al.*, 2021] and Real-ESRGAN [Wang *et al.*, 2021] employ realistic degradation modeling techniques. Despite progress in generating perceptually realistic details, GAN-based ISR methods often suffer from unstable training and produce unnatural artifacts. In recent years, the powerful Stable Diffusion (SD) [Rombach *et al.*, 2022] model has been applied to ISR tasks [Wang *et al.*, 2024b; Lin *et al.*, 2023; Yu *et al.*, 2024; Wu *et al.*, 2024; Yang *et al.*, 2024]. For instance, SeeSR [Wu *et al.*, 2024] proposes a semantic-aware approach that better preserves semantic fidelity in reconstructing real-world images. While achieving remarkable performance in Real-ISR tasks, these methods are limited to the spatial domain, failing to fully exploit frequency domain characteristics. In contrast, we discuss the degradation processes of various frequency components and design a timestep-adaptive method to enhance them.

Frequency-based Image Reconstruction. Frequency analysis of image processing has been widely used in computer vision [Yu *et al.*, 2022; Huang *et al.*, 2024; Yang and Soatto, 2020; Cai *et al.*, 2021; Si *et al.*, 2024; Yu *et al.*, 2022]. For ISR tasks, many studies improve images reconstruction quality by applying frequency domain to comprehensively extract features from low-resolution images [Guan *et al.*, 2024; Li *et al.*, 2023a; Xu *et al.*, 2024; Xie *et al.*, 2021]. Among these approaches, most frequency-domain enhancement algorithms are primarily applied to CNN-based and GAN-based models [Yu *et al.*, 2022; Huang *et al.*, 2022; Guo *et al.*, 2022]. Some methods improve reconstruction quality by separating specific components (such as high-frequency or amplitude components) in the frequency domain [Guan *et al.*, 2024; Li *et al.*, 2023a; Xu *et al.*, 2024; Xie *et al.*, 2021; Dai *et al.*, 2024]. Appendix A lists the effects of different frequency components on image quality for other computer vision tasks. Although these existing frequency domain-based ISR methods significantly improve performance, their integration with the increasingly popular diffusion models remains largely unexplored. Only a few methods [Wang *et al.*, 2024c; Moser *et al.*, 2024; Zhao *et al.*, 2024; Luo *et al.*, 2023] explore the potential of diffusion-based models in the frequency domain. For instance, WF-Diff [Zhao *et al.*, 2024] employs a wavelet-based dual-branch diffusion framework to refine frequency components of the initial input images. FourierDiff [Lv *et al.*, 2024] decomposes frequency-domain samples, using generative prior amplitudes to enhance brightness. These methods lack deep integration of frequency components with crucial denoising timesteps of the diffusion. In contrast, our TFDSR framework introduces frequency-domain enhancement through a novel timestep-adaptive approach that leverages the inherent characteristics of diffusion models.

3 Background and Preliminaries

3.1 Diffusion Models for Image Super-Resolution

Diffusion models, like DDPM [Ho *et al.*, 2020] and LDM [Rombach *et al.*, 2022], are latent variable models primarily consisting of a diffusion and denoising process. In the diffusion process, Gaussian noise is gradually added at each timestep t via a Markov chain, using a variance schedule $\beta_1, ..., \beta_t$, resulting in a random noise distribution

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

In the denoising process, given the noisy input $\boldsymbol{x}_t$, the model outputs the clean data $\boldsymbol{x}_{t-1}$ without noise, represented as

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)=\mathcal{N}\left(\boldsymbol{x}_{t-1} ; \boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\boldsymbol{x}_t, t\right)\right). \tag{2}$$

Here, $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ are set by the denoising model. Current diffusion models [Ho *et al.*, 2020; Rombach *et al.*, 2022] are implemented using a U-Net [Ronneberger *et al.*, 2015] to remove noise from data samples, which consists of downsampling and upsampling paths. Each upsampling block concats both the backbone and skip features in the skip connections.

To improve ISR performance, existing diffusion-based methods typically encode LR images and fuse them with the U-Net through cross-attention or a ControlNet module for HR image generation. Through iterative diffusion and reverse processes, these models effectively recover realistic details.

3.2 Fourier Frequency Domain Transformation

The Fast Fourier Transform (FFT) is widely applied in low-level vision tasks, transforming images from the spatial domain to the Fourier frequency domain, denoted as

$$\mathcal{F}(\boldsymbol{x})(u,v)=\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\boldsymbol{x}(h,w)e^{-j2\pi\left(\frac{h}{H}u+\frac{w}{W}v\right)}. \tag{3}$$

Its inverse function (IFFT) is formulated as

$$\mathcal{G}(\boldsymbol{f})(h,w)=\frac{1}{UV}\cdot\sum_{u=0}^{U-1}\sum_{v=0}^{V-1}\boldsymbol{f}(u,v)e^{-j2\pi\left(\frac{u}{U}h+\frac{v}{V}w\right)}, \tag{4}$$

where j is the imaginary unit, and $e^{j\theta}=\cos\theta+j\sin\theta$. $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$ denote 2D Fourier transform and its inverse. The frequency features $\mathcal{F}(\boldsymbol{x})$ in Eq. (3) and $\boldsymbol{f}$ in Eq. (4) are both complex domain tensors, expressed as $\mathcal{F}(\boldsymbol{x})=\mathcal{R}(\boldsymbol{x})+j\mathcal{I}(\boldsymbol{x})$, with $\mathcal{R}(\boldsymbol{x}),\mathcal{I}(\boldsymbol{x})$ being the real and imaginary parts.

In this paper, we explore two frequency-domain decomposition methods (see Appendix A). The first is the frequency spectral decomposition, separating the frequency into the amplitude $\mathcal{A}$ and phase $\mathcal{P}$, which is represented as

$$\mathcal{A}(\boldsymbol{x})(u,v)=\sqrt{\mathcal{R}^2(\boldsymbol{x})(u,v)+\mathcal{I}^2(\boldsymbol{x})(u,v)}, \tag{5}$$

$$\mathcal{P}(\boldsymbol{x})(u,v)=\arctan\left[\frac{\mathcal{I}(\boldsymbol{x})(u,v)}{\mathcal{R}(\boldsymbol{x})(u,v)}\right]. \tag{6}$$

The second is the frequency band decomposition, which divides the frequency domain into high- $\mathcal{H}$ and low-frequency $\mathcal{L}$ parts based on their distance from the frequency center.

4 The Propose Framework

In our preliminary experimental explorations, we observe that in the early denoising stages, diffusion-based ISR models tend to reconstruct the overall structural phase and low-frequency components while generating amplitude and high-frequency details in the later stages. To leverage this dynamic sampling characteristic, we propose a Timestep Division Controller (Section 4.1) to determine which frequency component at the current timestep should be enhanced. For the enhancement modules, shown in Figure 3, we introduce a channel attention-based module (Section 4.2) for frequency spectral enhancement, and an adaptive modulation module (Section 4.3) for frequency band enhancement.

4.1 Timestep Division Controller

To determine the optimal demarcation points for dividing the sampling timesteps into two distinct groups, we propose a computationally efficient timestep division optimization strategy. Recall our frequency-domain analysis (as illustrated in Figure 1a), which reveals rapid phase and low-frequency components changing in early stages, significant

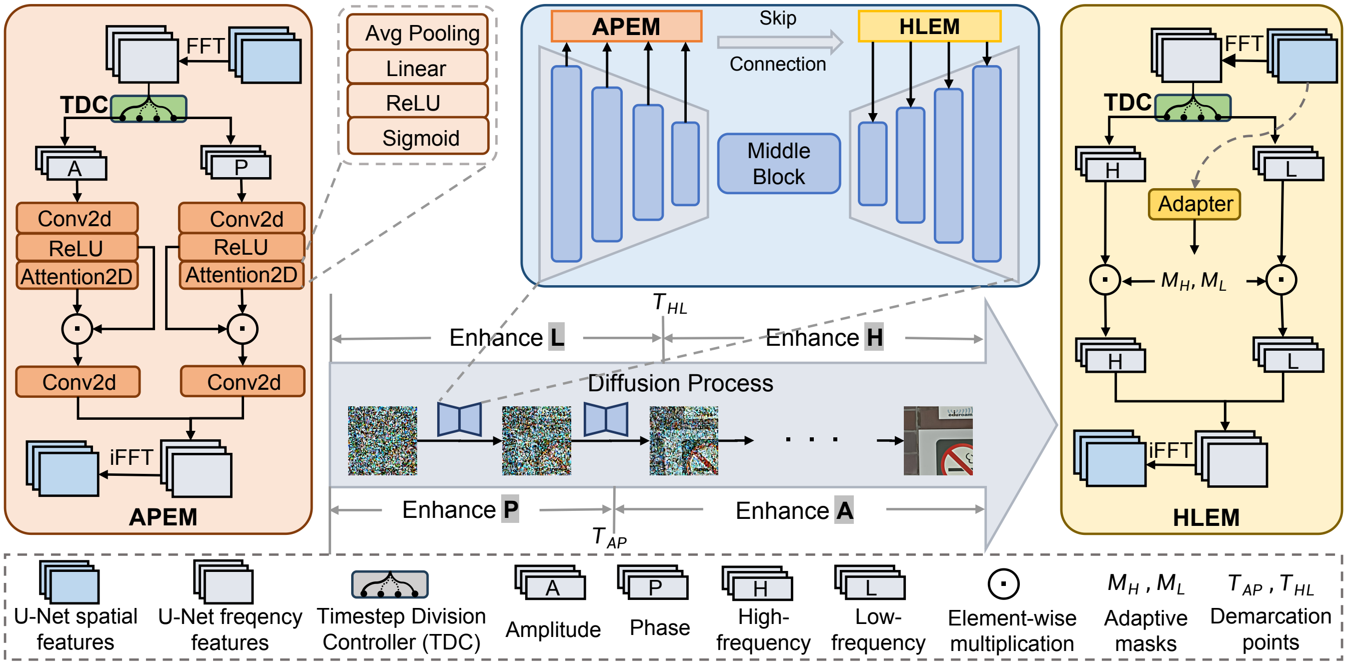


Figure 3: **The overview of TFDSR**, which has three modules, (a) **TDC**: a dynamical timestep division controller, determining which component should be activated across the timesteps. (b) **APEM**: a channel attention module for frequency spectral components that adaptively enhance the missing amplitude and phase components; (c) **HLEM**: a modulation module for frequency band components that enhances high- and low-frequency information in the skip features using, improving detail restoration capabilities.

amplitude and high-frequency components variations in later. Leveraging this temporal divergence, our strategy identifies the demarcation points $\mathcal{T}_{\mathcal{AP}}, \mathcal{T}_{\mathcal{HL}}$ through three steps.

a) MSE Arrays of Frequency Domain Components. To quantify the reconstruction dynamics of different frequency domain components, in the diffusion model-based ISR baseline (SeeSR), we save the sampling results x_t every 5 steps. Then we decomposed them into low-frequency $\mathcal{L}_t$ and high-frequency components $\mathcal{H}_t$, as well as amplitude $\mathcal{A}_t$ and phase components $\mathcal{P}_t$ through FFT. Subsequently, frequency-domain fusion and iFFT are performed, expressed as

$$\mathcal{I}_{\mathcal{A}}(t), \mathcal{I}_{\mathcal{P}}(t) = \mathcal{G}(\mathcal{A}_t \cdot e^{j\mathcal{P}_{\mathrm{HR}}}), \mathcal{G}(\mathcal{A}_{\mathrm{HR}} \cdot e^{j\mathcal{P}_t}); \quad (7)$$

$$\mathcal{I}_{\mathcal{H}}(t), \mathcal{I}_{\mathcal{L}}(t) = \mathcal{G}(\mathcal{H}_t + \mathcal{L}_{\mathrm{HR}}), \mathcal{G}(\mathcal{H}_{\mathrm{HR}} + \mathcal{L}_t); \quad (8)$$

where $\mathcal{A}_{\mathrm{HR}}$, $\mathcal{P}_{\mathrm{HR}}$, $\mathcal{H}_{\mathrm{HR}}$ and $\mathcal{L}_{\mathrm{HR}}$ represent the frequency components of HR images, respectively. The fused results are used to calculate the Mean Squared Error (MSE) with the HR images, yielding a similarity array that changes across sampling timesteps for four different frequency-domain components. The visualization curves are shown in the Figure 1a.

b) Gradient Difference Calculation. For high- and low-frequency components, we calculate the frequency band gradient difference $\delta_{\mathcal{AP}} = \frac{d\mathcal{I}_{\mathcal{A}}(t)}{dt} - \frac{d\mathcal{I}_{\mathcal{P}}(t)}{dt}$. For amplitude and phase, we calculate the spectral gradient difference $\delta_{\mathcal{HL}} = \frac{d\mathcal{I}_{\mathcal{H}}(t)}{dt} - \frac{d\mathcal{I}_{\mathcal{L}}(t)}{dt}$ to capture the dynamic relationship between different frequency components during the denoising process.

c) Timestep Optimization Search. Notably, the gradient difference δ_{comp} follows a positive-to-negative or negative-to-positive pattern, which indicates that $\mathcal{L}/\mathcal{P}$ changes rapidly in the early timestep group $[0, t)$, while $\mathcal{H}/\mathcal{A}$ changes dominate in the later $(t, T-1]$. To identify the point where the sign reverses, we sum the cancellable signs, expressed as

$$\mathcal{S}^t_{\mathrm{comp}} = \Big|\sum_{i=0}^{t-1} \mathrm{sign}(\delta_{\mathrm{comp}}(i))\Big| + \Big|\sum_{i=t+1}^{n-1} \mathrm{sign}(\delta_{\mathrm{comp}}(i))\Big|, \quad \mathrm{comp} \in \{\mathcal{AP}, \mathcal{HL}\}. \quad (9)$$

Through searching across all sampling timesteps, $\mathcal{T}_{\mathcal{AP}}$ and $\mathcal{T}_{\mathcal{HL}}$ with $\max_t \mathcal{S}^t_{\mathrm{comp}}$ are selected as the demarcation points. Then, our controller divides the whole sampling process into two groups $[0, \mathcal{T}_{\mathcal{AP}})$, $[\mathcal{T}_{\mathcal{AP}}, T-1]$ for frequency spectral decomposition, and $[0, \mathcal{T}_{\mathcal{HL}})$, $[\mathcal{T}_{\mathcal{HL}}, T-1]$ for frequency bands, which is applied to determine the enhanced components across diffusion sampling timesteps.

4.2 Frequency Spectral Enhancement

As stated previously, we observe that the amplitude and phase features of images are progressively restored during the ISR reconstruction process. Therefore, we propose that diffusion-based ISR models should actively learn these two types of features during training. Inspired by prior studies [Hu *et al.*, 2018; Zhao *et al.*, 2019], which enhance model performance by adjusting the importance of different channel features in convolutional neural networks (CNNs), we extend this concept to amplitude and phase features. To this end, we design a channel attention module based on CNNs, termed the Amplitude-Phase Enhancement Module (APEM), to enhance these two frequency-domain components. Through this, the diffusion model is able to perceive the phase features that better capture structural information at the early stage, and later amplitude features to improve the visual quality of the image. Technically, this process can be divided into two steps:

a) Initial Feature Extraction. First, we extract the amplitude and phase components of the features from the upsampling blocks of the U-Net. This is implemented through a combination of convolution and ReLU operations, which serve as the initial feature extraction, represented as

$$\mathcal{C}_f = \text{ReLU}(\text{Conv}(\mathcal{C})), \tag{10}$$

$$\text{where } \mathcal{C} := \mathcal{A} \text{ if } t \geq \mathcal{T}_{\mathcal{AP}} \text{ else } \mathcal{P}, \tag{11}$$

where $\mathcal{C}$ is the controller, $\text{Conv}(\cdot)$ denotes convolution operation, := means the replacement of variables. We activate $\mathcal{A}$ when the timestep $t \geq \mathcal{T}_{\mathcal{AP}}$, and activate $\mathcal{P}$ when $t < \mathcal{T}_{\mathcal{AP}}$, hence achieving adaptive control to better complete the frequency perceptron ability of diffusion-based ISR models.

b) Channel Attention Map Generation and Application. Next, we generate channel attention maps for both the amplitude and phase components, which are denoted as

$$\mathcal{M}_{\mathcal{C}}^{\text{attn}} = \sigma(\text{ReLU}(\text{Conv}(\text{AvgPool}(\mathcal{C}_f)))), \tag{12}$$

where $\mathcal{M}_{\mathcal{C}}^{\text{attn}}$ represents the attention maps of amplitude and phase. $\sigma(\cdot)$ and $\text{AvgPool}(\cdot)$ denote the sigmoid and average pooling operations. Then we calculate the enhanced amplitude and phase components $\mathcal{A}_{\text{out}}$ and $\mathcal{P}_{\text{out}}$ by applying the $\mathcal{M}_{\mathcal{C}}^{\text{attn}}$, which is denoted as

$$\mathcal{C}_{\text{out}} = \text{Conv}(\mathcal{M}_{\mathcal{C}}^{\text{attn}} \odot \mathcal{C}_f). \tag{13}$$

The attention map allows the model to better emphasize channel features that contribute significantly to the ISR task.

To apply the enhanced amplitude and phase components, we combine $\mathcal{A}_{\text{out}}$ and $\mathcal{P}_{\text{out}}$ into the frequency domain by $\mathcal{F}' = \mathcal{A}_{\text{out}} \cdot e^{j\mathcal{P}_{\text{out}}}$, and further transfer to the spatial domain by iFFT, which is denoted as $\boldsymbol{x}'_{\text{skip}} = \mathcal{G}(\mathcal{F}(\boldsymbol{x}_{\text{skip}})')$.

4.3 Frequency Band Enhancement

Next, for frequency band, we explore enhancing the high- and low-frequency components at different sampling timesteps. As discussed in FreeU [Si *et al.*, 2024], the skip connections in U-Net blocks can transmit high-frequency, information-rich features to deeper layers of the network, thereby preserving more comprehensive image information. Note that FreeU is designed for text-to-image tasks which only applies two constant scaling transformations to low-frequency features on all layers. However, for diffusion-based ISR problems, the features on U-Net layers with various resolutions convey various semantic information. Therefore, considering the varying richness of information, we propose a high- and low-frequency enhancement module (HLEM) with adaptive masking to enhance different frequency bands across sampling timesteps. It can be divided into the following two steps.

a) Adaptive Mask Construction. To accurately filter and dynamically enhance the frequency components in the skip features, we construct the adaptive mask $\mathcal{M}_{\mathcal{H}}$ and $\mathcal{M}_{\mathcal{L}}$, for high- and low-frequency respectively. Considering that lower-level and smaller-scale features often contain less detailed information, the mask adjusts the enhancement factor based on scale adaptively, to better adapt to the frequency structure of features at different levels, formulated as

$$\mathcal{M}_{\mathcal{C}}(r) = 1 + (\frac{S - S_{\min}}{S_{\max} - S_{\min}} + 0.5) \cdot \frac{P_{\mathcal{C}}}{2} \cdot (r > r_{\text{thresh}}), \tag{14}$$

$$\text{where } \mathcal{C} := \mathcal{H} \text{ if } t \geq \mathcal{T}_{\mathcal{HL}} \text{ else } \mathcal{L}. \tag{15}$$

Here r and r_{thresh} are the radius and the radius threshold relative to the frequency center; $\mathcal{C}$ is the controller. S is the scale of skip features; $P_{\mathcal{C}}$ is the enhancement factor. Similarly, we activate $\mathcal{L}$ when the timestep $t < \mathcal{T}_{\mathcal{HL}}$, and activate $\mathcal{H}$ when $t \geq \mathcal{T}_{\mathcal{HL}}$, obtaining the high- and low-frequency masks.

b) High- and Low-Frequency Component Enhancement. We then perform an element-wise multiplication of the adaptive mask $\mathcal{M}_{\mathcal{C}}$ from Equation 14 with the skip features $\boldsymbol{x}_{\text{skip}}$ in the frequency domain, to amplify and enhance frequency band components, which is represented as

$$\mathcal{F}(\boldsymbol{x}_{\text{skip}})' = \mathcal{F}(\boldsymbol{x}_{\text{skip}}) \odot \mathcal{M}_{\mathcal{C}}, \tag{16}$$

where $\odot$ denotes element-wise multiplication. Finally, the inverse Fourier transformation transfers the enhanced skip features to the spatial domain, denoted as $\boldsymbol{x}'_{\text{skip}} = \mathcal{G}(\mathcal{F}(\boldsymbol{x}_{\text{skip}})')$.

4.4 Overall Training and Sampling Details

Notably, diffusion-based ISR models typically consist of two key stages — the training and sampling process. In practice, our modules are integrated into the whole process as follows:

- For frequency spectral enhancement (APEM), we focus on learning frequency-domain feature parameters without requiring a timestep division controller during the training process. During sampling, we apply the demarcation point $\mathcal{T}_{\mathcal{AP}}$ for adaptive enhancement, prioritizing the phase component in the initial stage with $t \in [0, \mathcal{T}_{\mathcal{AP}})$, and focusing on the amplitude component in the later stage with $t \in [\mathcal{T}_{\mathcal{AP}}, T-1]$.
- For frequency band enhancement (HLEM), which does not contain any trainable layers, we set it as a training-free module and apply it exclusively during the sampling process. Similarly, we enhance the low-frequency component when the timestep $t \in [0, \mathcal{T}_{\mathcal{HL}})$, and regulate the high-frequency when $t \in [\mathcal{T}_{\mathcal{HL}}, T-1]$.

Overall, we make only minor modifications to the network structure, which have minimal impact on the entire training process. Consequently, in practice, the overall training cost of our method is comparable to that of our diffusion-based ISR baseline model (SeeSR); see Appendix E. These two modules are applied to the skip connection features from the U-Net down-blocks (see Figure 3). For more ablation results of the training strategies and places, please refer to Appendix C.

5 Experiments

5.1 Experimental Settings

Training Datasets. We train TFDSR on the first 10K real-world images from LSDIR [Li *et al.*, 2023b] and first 10k face images from FFHQ [Karras *et al.*, 2021], which are cropped into 512 × 512 patches. And we use the degradation model with the setting of Real-ESRGAN [Wang *et al.*, 2021].

Datasets	Metrics	BSR-GAN	Real-ESRGAN	FeMaSR	DASR	SwinIR-GAN	StableSR	SS-MoE	SUPIR	PASD	SeeSR	TFDSR (Ours)
DIV2K-Val	PSNR↑	**24.57**	24.29	23.05	24.46	23.92	23.26	22.11	22.14	24.16	23.71	23.87
	SSIM↑	0.6232	**0.6328**	0.5816	0.6267	0.6235	0.5644	0.5775	0.5180	0.6099	0.6043	0.6004
	LPIPS↓	0.3354	0.3115	0.3125	0.3542	0.3159	0.3119	**0.2881**	0.3930	0.3705	0.3212	0.3251
	MUSIQ↑	61.23	61.11	60.82	55.20	60.22	65.91	64.78	63.60	61.85	68.56	**69.42**
	CLIPIQA↑	0.5255	0.5283	0.5997	0.5036	0.534	0.6771	/	0.713	0.5848	0.6927	**0.7300**
	MANIQA↑	0.3561	0.3823	0.3457	0.3186	0.3656	0.4208	/	**0.5533**	0.4028	0.5040	0.5510
RealSR	PSNR↑	26.37	25.65	25.06	**27.01**	26.3	24.66	24.68	23.64	26.53	25.05	25.20
	SSIM↑	0.7643	0.7592	0.7342	0.7702	**0.7719**	0.7003	0.7352	0.6603	0.7597	0.7394	0.7359
	LPIPS↓	0.2652	0.2720	0.2896	0.3047	**0.2479**	0.3101	0.2719	0.3511	0.2783	0.2862	0.3020
	MUSIQ↑	63.19	60.49	59.20	40.95	58.83	65.24	57.10	61.34	60.61	70.99	**71.15**
	CLIPIQA↑	0.5105	0.4491	0.5450	0.3135	0.4367	0.6169	/	0.6316	0.5030	0.6787	**0.7245**
	MANIQA↑	0.3800	0.3769	0.3648	0.2459	0.3455	0.4302	/	0.4952	0.3894	0.5456	**0.5771**
DRealSR	PSNR↑	28.68	28.61	26.87	**29.74**	28.46	27.93	29.35	24.80	28.96	27.92	27.77
	SSIM↑	0.8021	0.8044	0.7557	**0.8257**	0.8036	0.7442	0.7946	0.6333	0.7919	0.7773	0.7683
	LPIPS↓	0.2885	0.2848	0.3179	0.3143	**0.2801**	0.3280	0.3017	0.4323	0.3142	0.3196	0.3406
	MUSIQ↑	57.25	54.26	53.32	42.43	52.65	58.28	42.32	59.73	52.29	65.37	**67.07**
	CLIPIQA↑	0.5104	0.4525	0.5534	0.3807	0.4389	0.6272	/	0.6880	0.5122	0.6887	**0.7168**
	MANIQA↑	0.3407	0.3422	0.3121	0.2822	0.3265	0.3890	/	0.5040	0.3672	0.5164	**0.5526**

Table 1: Quantitative comparison with SOTA methods on the synthetic benchmark DIV2K-Val. **Red** and blue colors represent the best and second-best performance. ↓ represents the smaller the better, while ↑ represents the opposite. It is evident that the core of GAN-based ISR methods lies in enhancing image fidelity, primarily reflected in higher full-reference metrics (e.g., PSNR). In contrast, Diffusion-based ISR methods focus on improving image quality, mainly demonstrated by higher no-reference metrics (e.g., CLIPIQA). Note that the symbol '/' denotes that these metrics are not provided in the original paper.

Testing Datasets. We employ the StableSR [Wang *et al.*, 2024b] test datasets and evaluate our approach on the following datasets. (1) For the synthetic dataset, we use 3,000 generated pairs of LR-HR images from the DIV2K validation set [Agustsson and Timofte, 2017], where the LR images have a resolution of 128 × 128, and the HR images have a resolution of 512 × 512. (2) For the real-world datasets, we utilize the DRealSR [Wei *et al.*, 2020] and RealSR [Ji *et al.*, 2020] datasets center-cropping the LR images to 128 × 128.

Evaluation Metrics. We adopt a series of full-reference and no-reference metrics to assess the performance of different methods. The full-reference metrics include PSNR, SSIM (evaluated on the Y channel in the YCbCr color space), and LPIPS [Zhang *et al.*, 2018]. For quality evaluation, we employ no-reference image quality assessment (IQA) metrics, including CLIP-IQA [Wang *et al.*, 2023], MUSIQ [Ke *et al.*, 2021], and MANIQA [Yang *et al.*, 2022].

Implementation Details. We employ the SeeSR [Wu *et al.*, 2024], a controlled T2I (Text-to-Image) diffusion-based model, as our pre-trained baseline. Then we train the APEM for 600 iterations with a batch size of 32, a learning rate of 5×10^{-5}, and 512 × 512 resolution on a single A100 GPU. During sampling, we utilize the adaptive frequency sampling strategy using the TDC module, which dynamically selects enhanced frequency components based on the current sampling timestep, with a total sampling step of 50. Hyperparameters $\mathcal{T}_{\mathcal{AP}} = 400$, $\mathcal{T}_{\mathcal{HL}} = 500$, $P_{\mathcal{H}} = 0.05$, and $P_{\mathcal{L}} = 0.9$ are tuned using a validation set composed of 100 randomly selected images from the training set (LSDIR+FFHQ), which are uniformly applied to the three datasets of Table 1, including varying samples. For the ablation results of hyperparameters tuning and the full reproducibility information, please refer to Appendix B and the source code.

Compared Methods. We select several state-of-the-art (SOTA) ISR models, which can be divided into two groups. The first group consists of GAN-based methods, including BSRGAN [Zhang *et al.*, 2021], Real-ESRGAN [Wang *et al.*, 2021], FeMaSR [Chen *et al.*, 2022], DASR [Liang *et al.*, 2022], SwinIR-GAN [Liang *et al.*, 2021]. The second group is diffusion-based methods, including StableSR [Wang *et al.*, 2024b], SS-MoE [Luo *et al.*, 2023], SUPIR [Yu *et al.*, 2024], PASD [Yang *et al.*, 2024], SeeSR [Wu *et al.*, 2024].

5.2 Comparison with Existing Models

Quantitative Comparisons. As shown in Table 1, we first conduct a quantitative comparison between the proposed method and the current state-of-the-art (SOTA) methods on both synthetic and real-world datasets. The results demonstrate that our method achieves the best scores on almost all no-reference metrics. Specifically, on the real-world benchmark RealSR, our TFDSR achieves a CLIP-IQA score of **0.7245**, representing a **6.75%** improvement over our baseline (the second-best method) SeeSR, which fully validates the superiority of TFDSR. Notably, the experimental results indicate that *GAN-based methods outperform almost all based on diffusion models in terms of full-reference metrics* (average PSNR/SSIM metrics, Diffusion: 23.21/0.5791 vs. GAN: 24.06/0.6176). This discrepancy can be primarily attributed to a potential limitation inherent to the training strategies of

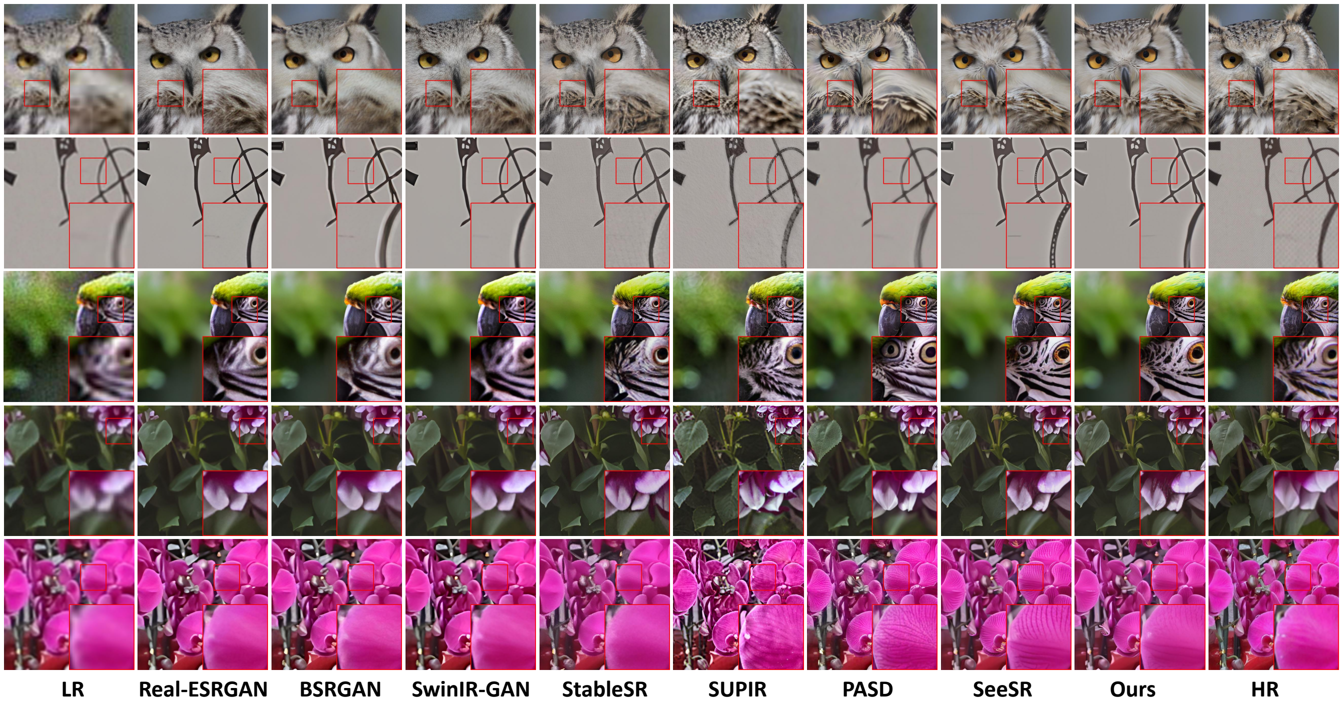


Figure 4: Qualitative comparisons of GAN-based, diffusion model-based ISR methods and our TFDSR. It shows that TFDSR can reconstruct more realistic high-resolution images without unnatural artifacts.

diffusion models. For more detailed discussions, please refer to the Appendix F. ***Overall, it can be shown that our TFDSR does indeed largely outperform current diffusion-based ISR methods in terms of no-reference metric scores while maintaining very competitive full-reference metrics.***

Qualitative Comparisons. To demonstrate the effectiveness of TFDSR, Figure 4 presents a comparison between the existing methods and our TFDSR. It can be observed that our method enhances the quality and fidelity of the image generated by diffusion-based ISR methods, particularly in detailed textures and general visual effects. Specifically, GAN-based approaches tend to produce over-smoothed outputs, whereas diffusion models frequently exhibit unnatural details, particularly manifesting in biological features (e.g., mammalian fur patterns) and complex textures. For instance, the third case in Figure 4 demonstrates some diffusion-based methods will generate incorrect eyes and blurring results. More examples with varying orientations of stripes are in Appendix D.

User Study. We also conduct a user study comparing our method on a set of test datasets, instructing 20 participants to choose the result with high quality and fidelity among these test methods. The results are shown in Figure 1b, revealing superior performance of our approach in perceptual quality.

5.3 Ablation Study

To further demonstrate the effectiveness of our TFDSR method, we conduct ablation results on three key frequency-based modules. We validate the superior performance of APEM shown in Row 2 of Table 2. Next, we evaluate the effectiveness of HLEM shown in Row 3 of Table 2. And we also show the significant potential in ISR of TDC (see Row 4 of Table 2). By applying three modules, we achieve obvious improvement in no-reference metrics over baseline (see Row 1, Table 2), while maintaining competitive full-reference metrics. More ablation results are in Appendix C.

Varients			**Metrics (RealSR)**					
TDC	**APEM**	**HLEM**	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA↑	MANIQA↑
✗	✗	✗	25.05	0.7394	0.2862	70.99	0.6787	0.5456
✓	✓	✗	25.17	0.7435	0.2888	70.29	0.6907	0.5551
✓	✗	✓	25.25	0.7373	0.3013	71.10	0.7239	0.5756
✗	✓	✓	24.96	0.7406	0.2856	70.32	0.6801	0.5435
✓	✓	✓	25.20	0.7359	0.3020	71.15	0.7245	0.5771

Table 2: Ablation studies of TFDSR modules and the relative locations on RealSR. ✓ and ✗ denote the inclusion and exclusion.

6 Conclusion

In this work, we propose a *timestep-adaptive* framework TFDSR for enhancing diffusion-based ISR models from a frequency perspective. To achieve this, we first propose a novel channel attention mechanism for enhancing the frequency spectral components (APEM). Also, we develop a new semantic-aware mask that adaptively determines the thresholds by feature inputs for regulating the frequency band components (HLEM). As shown in the extensive experimental evaluation, we demonstrate the effectiveness of the TFDSR. We also hope our work will draw more attention from the community toward a broader view of addressing diffusion for low-level vision from a frequency perspective.

Acknowledgments

This paper is supported by NSFC under Grants (No. 62402424). Haobo Wang is also partially supported by (No. U24A201401). Hanbin Zhao is supported by NSFC under Grant 62402430, Zhejiang Provincial Natural Science Foundation of China under Grant LQN25F020008.

References

[Agustsson and Timofte, 2017] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proc. of CVPR*, 2017.

[Cai *et al.*, 2021] Mu Cai, Hong Zhang, Huijuan Huang, Qichuan Geng, Yixuan Li, and Gao Huang. Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proc. of ICCV*, 2021.

[Chen *et al.*, 2022] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proc. of ACM MM*, 2022.

[Dai *et al.*, 2024] Tao Dai, Jianping Wang, Hang Guo, Jinmin Li, Jinbao Wang, and Zexuan Zhu. Freqformer: Frequency-aware transformer for lightweight image super-resolution. In *Proc. of IJCAI*, 2024.

[Dong *et al.*, 2016a] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

[Dong *et al.*, 2016b] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Proc. of ECCV*, 2016.

[Fu *et al.*, 2021] Minghan Fu, Huan Liu, Yankun Yu, Jun Chen, and Keyan Wang. DW-GAN: A discrete wavelet transform GAN for nonhomogeneous dehazing. In *Proc. of CVPR*, 2021.

[Guan *et al.*, 2024] Wenxue Guan, Haobo Li, Dawei Xu, Jiaxin Liu, Shenghua Gong, and Jun Liu. Frequency generation for real-world image super-resolution. *IEEE Trans. Circuits Syst. Video Technol.*, 2024.

[Guo *et al.*, 2022] Xin Guo, Xueyang Fu, Man Zhou, Zhen Huang, Jialun Peng, and Zheng-Jun Zha. Exploring fourier prior for single image rain removal. In *Proc. of IJCAI*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*, 2020.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. of CVPR*, 2018.

[Huang *et al.*, 2022] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *Proc. of ECCV*, 2022.

[Huang *et al.*, 2024] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *Proc. of ECCV*, 2024.

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

[Karras *et al.*, 2021] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[Ke *et al.*, 2021] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. In *Proc. of ICCV*, 2021.

[Kim *et al.*, 2016] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proc. of CVPR*, 2016.

[Li *et al.*, 2023a] Ao Li, Le Zhang, Yun Liu, and Ce Zhu. Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. In *Proc. of ICCV*, 2023.

[Li *et al.*, 2023b] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. LSDIR: A large scale dataset for image restoration. In *Proc. of CVPR*, 2023.

[Li *et al.*, 2024] Guangyuan Li, Chen Rao, Juncheng Mo, Zhanjie Zhang, Wei Xing, and Lei Zhao. Rethinking diffusion model for multi-contrast MRI super-resolution. In *Proc. of CVPR*, 2024.

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

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

[Lim *et al.*, 2017] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proc. of CVPR*, 2017.

[Lin *et al.*, 2023] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *CoRR*, 2023.

[Liu *et al.*, 2017] Wu Liu, Xinchen Liu, Huadong Ma, and Peng Cheng. Beyond human-level license plate super-resolution with progressive vehicle search and domain priori GAN. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 2017.

[Luo *et al.*, 2023] Feng Luo, Jinxi Xiang, Jun Zhang, Xiao Han, and Wei Yang. Image super-resolution via latent diffusion: A sampling-space mixture of experts and frequency-augmented decoder approach. *CoRR*, 2023.

[Lv *et al.*, 2024] Xiaoqian Lv, Shengping Zhang, Chenyang Wang, Yichen Zheng, Bineng Zhong, Chongyi Li, and Liqiang Nie. Fourier priors-guided diffusion for zero-shot joint low-light enhancement and deblurring. In *Proc. of CVPR*, 2024.

[Moser *et al.*, 2024] Brian B. Moser, Stanislav Frolov, Federico Raue, Sebastian Palacio, and Andreas Dengel. Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution. In *Proc. of IJCNN*, 2024.

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

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, 2015.

[Shermeyer and Etten, 2019] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proc. of CVPR*, 2019.

[Shi *et al.*, 2016] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proc. of CVPR*, 2016.

[Si *et al.*, 2024] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proc. of CVPR*, 2024.

[Tai *et al.*, 2017] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proc. of ICCV*, 2017.

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

[Wang *et al.*, 2023] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proc. of AAAI*, 2023.

[Wang *et al.*, 2024a] Boyang Wang, Fengyu Yang, Xihang Yu, Chao Zhang, and Hanbin Zhao. APISR: anime production inspired real-world anime super-resolution. In *Proc. of CVPR*, 2024.

[Wang *et al.*, 2024b] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *Int. J. Comput. Vis.*, 2024.

[Wang *et al.*, 2024c] Xingjian Wang, Li Chai, and Jiming Chen. Frequency-domain refinement with multiscale diffusion for super resolution. *CoRR*, 2024.

[Wei *et al.*, 2020] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Proc. of ECCV*, 2020.

[Wu *et al.*, 2024] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proc. of CVPR*, 2024.

[Xie *et al.*, 2021] Wenbin Xie, Dehua Song, Chang Xu, Chunjing Xu, Hui Zhang, and Yunhe Wang. Learning frequency-aware dynamic network for efficient super-resolution. In *Proc. of ICCV*, 2021.

[Xu *et al.*, 2024] Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. Videogigagan: Towards detail-rich video super-resolution. *CoRR*, 2024.

[Yang and Soatto, 2020] Yanchao Yang and Stefano Soatto. FDA: fourier domain adaptation for semantic segmentation. In *Proc. of CVPR*, 2020.

[Yang *et al.*, 2022] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: multi-dimension attention network for no-reference image quality assessment. In *Proc. of CVPR*, 2022.

[Yang *et al.*, 2024] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *Proc. of ECCV*, 2024.

[Yu *et al.*, 2022] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *Proc. of ECCV*, 2022.

[Yu *et al.*, 2024] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proc. of CVPR*, 2024.

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

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

[Zhao *et al.*, 2019] Xiaole Zhao, Yulun Zhang, Tao Zhang, and Xueming Zou. Channel splitting network for single MR image super-resolution. *IEEE Trans. Image Process.*, 2019.

[Zhao *et al.*, 2024] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proc. of CVPR*, 2024.