The Appendix of 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

A Application of Frequency Domain in Vision

In the related work of the main paper, we have briefly summarized existing methods that apply the frequency domain to super-resolution tasks. To further extend the discussion, this section broadens the scope by exploring additional applications of frequency domain components in low-level vision tasks. Numerous studies have investigated the impact of different frequency domain separation methods on image quality [Yang and Soatto, 2020; Yu *et al.*, 2022; Wang *et al.*, 2023; Si *et al.*, 2024; Wang *et al.*, 2024; Cai *et al.*, 2021; Mao *et al.*, 2023; Luo *et al.*, 2023]. (1) **Frequency Spectral Decomposition**: Frequency-domain information can be divided into amplitude and phase spectra through frequency spectral decomposition. FDA [Yang and Soatto, 2020] reduces the distribution discrepancy between the source and target domains by swapping their amplitude spectra. FSDGN [Yu *et al.*, 2022] addresses the dehazing problem by investigating the correlation between amplitude and phase spectra in the frequency domain under foggy degradation. (2) **Frequency Band Decomposition**: The frequency domain can also be divided into high- and low-frequency components by the distance from the frequency center via frequency band decomposition. FreeU [Si *et al.*, 2024] suppresses low-frequency features in the frequency domain to prevent Stable Diffusion from generating overly smooth images. FouriScale [Huang *et al.*, 2024] applies low-pass filtering in the frequency domain to alleviate repetitive patterns and structural distortions in the generation of high-resolution images by pre-trained diffusion models. (3) **Frequency Properties Decomposition**: The frequency domain can be separated into real and imaginary components. DeepRFT [Mao *et al.*, 2023] applies ReLU to the real and imaginary parts of the frequency domain separately to achieve effective image deblurring.

To better understand the semantic information represented by different frequency domain components, we perform visual modeling of the image components, shown in Figure 1. We first transformed the image into the frequency domain, then separated it according to the three aforementioned methods, and finally transformed these components back into the spatial domain. The results show that the frequency domain, along with the phase component, and the low- and high-frequency components, can clearly express the semantic information of the image. Specifically, the amplitude component mainly represents the stylistic features of the image, such as color and contrast; the phase component reflects the con-

Domain	Method	$\mathcal{A}$ and $\mathcal{P}$ Separate	$\mathcal{H}$ and $\mathcal{L}$ Separate	Freq Loss	Adaptive Samping
ISR	FSN	×	✓	×	×
	FDC	×	✓	✓	×
	ARFFT	×	×	✓	×
	FADN	×	✓	✓	×
	CRAFT	×	✓	×	×
	SS-MoE	×	×	✓	×
VSR	DFVSR	×	✓	✓	×
	FTVSR	×	✓	×	×
	VideoGigaGAN	×	✓	×	×
	MFPI	×	×	×	×
FSR	SFMNet	✓	×	✓	×
ISR	Ours	✓	✓	×	✓

Table 1: Classification and Comparison of Frequency-Domain-Based Super-Resolution Methods.

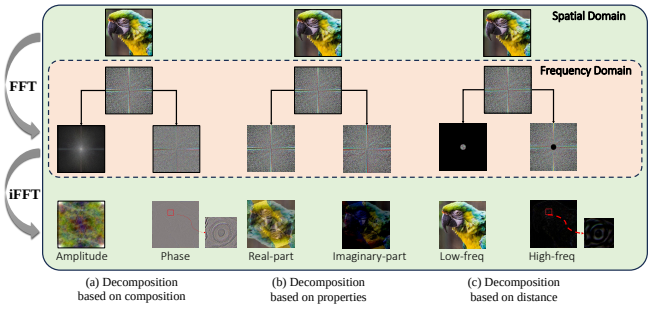

Figure 1: Image modeling methods in the frequency domain. We present the results of 3 decomposition approaches: (a) decomposition based on composition (spectral decomposition), (b) decomposition based on properties, and (c) decomposition based on distance (band decomposition). Compared to (b), (a) and (c) offer better separation of the intrinsic properties of the image.

*Corresponding author.

tour information; the low-frequency component captures the global characteristics of the image, while the high-frequency component encodes edge and texture details. This analysis provides intuitive evidence for the semantic understanding of frequency domain components and further validates the effectiveness of frequency domain methods.

Table 1 integrates and categorizes frequency-based super-resolution methods from multiple perspectives. It can be observed that other methods fail to consider degradation systematically. In contrast, our method addresses degradation by modulating degraded components across timesteps.

B Implementation Details

B.1 Hyperparameter Configuration of HLEM

In the experiment section of the main paper, we have briefly described the hyperparameter construction process. To further validate the rationality of our tuning strategy and parameter selection, this section provides detailed parameter settings, their corresponding performance metrics, and the specific rationale behind the parameter choices. For the high-frequency component enhancement parameter $P_{\mathcal{H}}$, as shown in Table 2, when $P_{\mathcal{H}} = 0$, the MUSIQ metric reaches its peak, while when $P_{\mathcal{H}} = 0.1$, the CLIPIQA and MANIQA metrics achieve their peaks. When $P_{\mathcal{H}} \geq 0.3$, all metrics show a declining trend. Therefore, we select the mean value of $P_{\mathcal{H}} = 0$ and $P_{\mathcal{H}} = 0.1$, which is $P_{\mathcal{H}} = 0.05$, to maintain the full-reference metrics representing fidelity while further improving the non-reference metrics representing image quality. For the low-frequency component enhancement parameter $P_{\mathcal{L}}$, as shown in Table 3, as $P_{\mathcal{L}}$ increases, the full-reference metrics exhibit minor fluctuations, while the non-reference metrics gradually improve. We ultimately set $P_{\mathcal{H}} = 0.9$ as it maximizes the MANIQA metric.

B.2 Discussion on the Module Configuration

We note that FreeU [Si *et al.*, 2024] also includes frequency-related operations. To validate the effectiveness of our adaptive masking for high- and low-frequency components, we replace HLEM with the operations of skip features in FreeU, resulting in a significant decrease in CLIPIQA (see Table 4).

B.3 The Algorithms of HLEM and APEM

As stated in the main paper, APEM and HLEM are two key modules embedded in the skip connections of U-Net within the diffusion model. The detailed algorithmic process for these two modules is shown in Algorithm 1.

Algorithm 1 TFDSR Algorithm

// 1. Choose the Split Timestep $\mathcal{T}_{\mathcal{AP}}, \mathcal{T}_{\mathcal{HL}}$;
for comp in $[\mathcal{AP}, \mathcal{HL}]$ **do**
  $\delta_{\text{comp}} = \frac{d\text{comp}[0](t)}{dt} - \frac{d\text{comp}[1](t)}{dt}$
  $\mathcal{S}^{t}_{\text{comp}} = |\sum_{i=0}^{t-1} \text{sign}(\delta_{\text{comp}}(i))| + |\sum_{i=t+1}^{n-1} \text{sign}(\delta_{\text{comp}}(i))|$
  Select t as $\mathcal{T}_{\text{comp}}$ at the Max $\mathcal{S}_{\text{comp}}$
end for
// 2. Model Structure;
for each $t \in [1, \text{Sampling Steps}]$ **do**
  Initialize the skip features $\boldsymbol{x}_{skip}$ in skip connection;
  $\boldsymbol{f}_{\text{skip}} = \mathcal{F}(\boldsymbol{x}_{\text{skip}})$
  // (1) APEM
  $\mathcal{A}(\boldsymbol{x}_{\text{bone}}), \mathcal{P}(\boldsymbol{x}_{\text{bone}}) = \text{FFTSplit}(\boldsymbol{f}_{\text{bone}})$;
  if $t < \mathcal{T}_{\mathcal{AP}}$ **then**
    $\mathcal{C} = \mathcal{P}$
  else
    $\mathcal{C} = \mathcal{A}$
  end if
  // a) Initial Feature Extraction;
  $\mathcal{C}_f = \text{ReLU}(\mathcal{C} \circledast \mathbf{K}_1^{\mathcal{C}})$;
  // b) Channel Attention Map Generation;
  // c) Amplitude and Phase Modulation;
  $\mathcal{C}_{\text{out}} = (\mathcal{M}_{\mathcal{C}}^{attn} \odot \mathcal{C}_f) \circledast \mathbf{K}_3^{\mathcal{C}}$;
  $\boldsymbol{f}'_{\text{skip}} = \text{FFTCombine}(\mathcal{A}_{\text{out}}, \mathcal{P}_{\text{out}})$, $\boldsymbol{x}'_{\text{skip}} = \mathcal{G}(\boldsymbol{f}'_{\text{skip}})$
  // (2) HLEM
  if $t < \mathcal{T}_{\mathcal{HL}}$ **then**
    $\mathcal{C} = \mathcal{L}$
  else
    $\mathcal{C} = \mathcal{H}$
  end if
  $\mathcal{M}_{\mathcal{C}}(r) = 1 + (\frac{S - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}} + 0.5) \cdot \frac{P_{\mathcal{C}}}{2} \cdot (r > r_{\text{thresh}})$;
  $\mathcal{F}(x_{\text{skip}})' = \mathcal{F}(x_{\text{skip}}) \odot \mathcal{M}(\alpha)$, $\boldsymbol{x}'_{\text{skip}} = \mathcal{G}(\boldsymbol{f}'_{\text{skip}})$;
end for

C More Ablation Result

Training Strategy. We exchange the training strategy of APEM and HLEM, and find that adaptive mask-based HLEM does not need any training strategy while the channel attention-based APEM needs a training strategy to achieve better performance (see Table 6).

The Effectiveness of Placement. We also validate the placement of APEM and HLEM on the RealSR benchmark. Rows 2 to 3 of Table 5 present the quantitative comparison results of applying APEM at different locations in the U-Net, which shows that applying APEM to the skip features with better performance. For the effectiveness of the HLEM, the quantitative comparison results are shown in Rows 4 to 5 of Table 5. The experimental results demonstrate that applying HLEM to the skip connection features improves the performance of full-reference and no-reference metrics compared to the baseline model (see row 1, Table 5).

The Effectiveness of TDC. To further validate the performance of the TDC module and the reasonableness of the variation score, we present the ablation study of $\mathcal{T}_{HL}$, which is shown in Table 7. The results demonstrate both the effectiveness and correctness of our design of the TDC module.

D Additional Visual Results

To show the performance of the structured patterns, we provide more examples on the RealSR dataset with varying orientations of stripes, such as clothing textures and window structures (see Figure 2), which show better performance on generating realistic details than the baseline (SeeSR).

To demonstrate the superiority of TFDSR over other diffusion-based frequency-domain low-level vision denoising algorithms [Zhao *et al.*, 2024; Lv *et al.*, 2024] in ISR task, we provide additional comparative visual results in Figure

Metrics	$P_{\mathcal{H}}$=0	$P_{\mathcal{H}}$=0.1	$P_{\mathcal{H}}$=0.2	$P_{\mathcal{H}}$=0.3	$P_{\mathcal{H}}$=0.4	$P_{\mathcal{H}}$=0.5	$P_{\mathcal{H}}$=0.6	$P_{\mathcal{H}}$=0.7	$P_{\mathcal{H}}$=0.8	$P_{\mathcal{H}}$=0.9	$P_{\mathcal{H}}$=1.0
PSNR	22.75	23.06	23.20	23.24	23.22	23.18	23.10	22.98	22.84	22.69	22.55
SSIM	0.6288	0.6366	0.6399	0.6412	0.6413	0.6409	0.6390	0.6357	0.6323	0.6288	0.6256
LPIPS	0.2972	0.3073	0.3269	0.3536	0.3859	0.4234	0.4633	0.4977	0.5229	0.5425	0.5573
MUSIQ	73.09	71.55	67.78	61.35	51.94	41.52	32.81	27.26	24.01	21.95	20.55
CLIP-IQA	0.7032	0.7137	0.6858	0.6278	0.5196	0.3925	0.3124	0.2610	0.2380	0.2242	0.2131
MANIQA	0.5459	0.5876	0.5824	0.5329	0.4386	0.3161	0.2089	0.1508	0.1239	0.1117	0.1048

Table 2: Hyperparameter $P_{\mathcal{H}}$ tuning results for HLEM in our TFDSR on the random subset of the training dataset.

Metrics	$P_{\mathcal{L}}$=0	$P_{\mathcal{L}}$=0.1	$P_{\mathcal{L}}$=0.2	$P_{\mathcal{L}}$=0.3	$P_{\mathcal{L}}$=0.4	$P_{\mathcal{L}}$=0.5	$P_{\mathcal{L}}$=0.6	$P_{\mathcal{L}}$=0.7	$P_{\mathcal{L}}$=0.8	$P_{\mathcal{L}}$=0.9	$P_{\mathcal{L}}$=1.0
PSNR	23.12	23.11	23.10	23.09	23.08	23.06	23.05	23.04	23.02	23.00	22.98
SSIM	0.6407	0.6403	0.6394	0.6385	0.6376	0.6366	0.6355	0.6343	0.6331	0.6318	0.6305
LPIPS	0.3160	0.3135	0.3123	0.3109	0.3089	0.3073	0.3058	0.3053	0.3048	0.3049	0.3047
MUSIQ	69.53	69.95	70.38	70.83	71.23	71.55	71.85	72.11	72.31	72.52	72.65
CLIP-IQA	0.6833	0.6893	0.6950	0.7018	0.7074	0.7137	0.7166	0.7199	0.7234	0.7267	0.7287
MANIQA	0.5779	0.5788	0.5805	0.5837	0.5860	0.5876	0.5882	0.5880	0.5885	0.5890	0.5876

Table 3: Hyperparameter $P_{\mathcal{L}}$ tuning results for HLEM in our TFDSR on the random subset of the training dataset.

Strategy	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA↑	MANIQA↑
FreeU	25.24	0.7512	0.2825	70.76	0.6786	0.5523
Ours	25.25	0.7372	0.3013	71.08	0.7237	0.5756

Table 4: Supplementary experiments of HPEM on RealSR.

Varients			Metrics (RealSR)					
Bone	**Skip**	**Remark**	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA↑	MANIQA↑
✗	✗	Baseline	25.05	0.7394	0.2862	70.99	0.6787	0.5456
✓	✗	w/ APEM	24.79	0.7343	0.2905	71.52	0.6903	0.5566
✗	✓		25.17	0.7435	0.2888	70.29	0.6907	0.5551
✓	✗	w/ HLEM	22.78	0.6662	0.3630	66.96	0.5682	0.4895
✗	✓		25.25	0.7373	0.3013	71.10	0.7239	0.5756
✗	✓	Default	25.20	0.7359	0.3020	71.15	0.7245	0.5771

Table 5: Ablation studies of TFDSR modules and the relative locations on RealSR. ✓ and ✗ denote training and free-training.

Varients		Metrics (RealSR)					
APEM	HLEM	PSNR↑	SSIM↑	LPIPS	MUSIQ↑	CLIPIQA↑	MANIQA↑
✓	$\mathcal{O}$	24.79	0.7343	0.2905	71.52	0.6903	0.5566
✗	$\mathcal{O}$	25.17	0.7416	0.2843	70.93	0.6784	0.5453
$\mathcal{O}$	✓	22.62	0.6486	0.3243	64.55	0.5232	0.4003
$\mathcal{O}$	✗	25.25	0.7373	0.3013	71.10	0.7239	0.5756
✓	✗	25.20	0.7359	0.3020	71.15	0.7245	0.5771

Table 6: Ablation studies of the training strategy. ✓ and ✗ denoting tuning and free-tuning, $\mathcal{O}$ indicates the absence of this module.

.

3. The results indicate that while methods for different low-level vision tasks share similarities, direct transfer may still encounter significant challenges.

w/ APEM	Metrics (RealSR)					
	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA↑	MANIQA↑
$\mathcal{T}_{HL}$=100	24.89	0.7270	0.2956	69.56	0.7035	0.5351
$\mathcal{T}_{HL}$=300	24.96	0.7275	0.2986	71.04	0.7018	0.5481
$\mathcal{T}_{HL}$=500	25.25	0.7373	0.3013	**71.10**	**0.7239**	**0.5756**
$\mathcal{T}_{HL}$=700	25.30	0.7431	0.2929	70.90	0.6979	0.5629
$\mathcal{T}_{HL}$=900	25.45	0.7468	0.2910	70.45	0.6840	0.5551

Table 7: Hyperparameter $\mathcal{T}_{HL}$ results for TDC in our TFDSR.

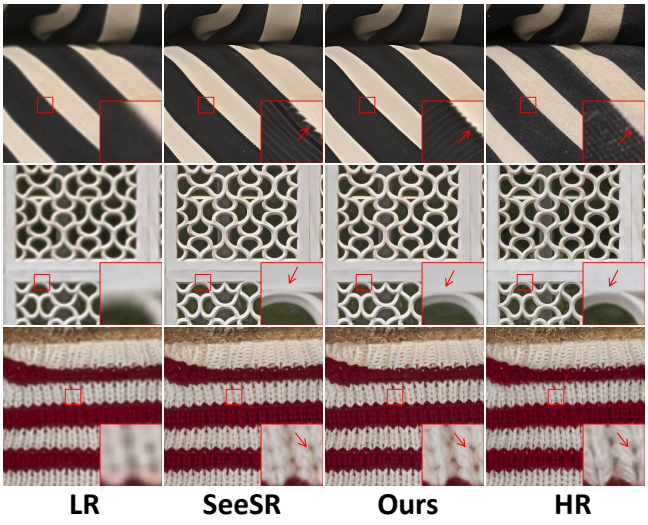

Figure 2: The qualitative comparison of structural examples between our baseline (SeeSR) and our TFDSR. It demonstrates that our approach can generate more realistic images.

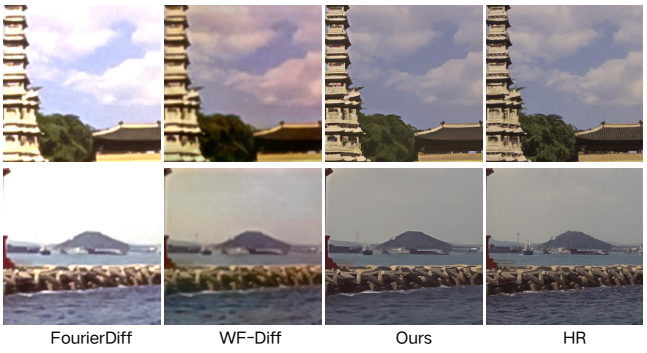

Figure 3: The qualitative comparison between existing diffusion-based frequency-domain low-level vision denoising algorithms and our TFDSR in ISR task. It demonstrates that both FourierDiff and WF-Diff struggle to achieve clarity in the generated images.

E Complexity Analysis

In this section, we conduct a comprehensive evaluation of the computational complexity of our proposed TFDSR method in comparison with the SeeSR baseline. The quantitative analysis, including model parameters and training time in 600 steps, is presented below. The results indicate that our method maintains identical computational statistics to the baseline while achieving superior performance (see Table 8).

Metrics	Params (M)	Time (s)
SeeSR	1615.79	1063.08
TFDSR	1615.79	1120.17

Table 8: The complexity comparison of our TFDSR framework.

F The Trade-off of Metrics

Notably, there is a trade-off between the full-reference metrics and perception metrics. In terms of the full-reference metrics like PSNR/SSIM, our approach achieves performance comparable to other Diffusion-based methods (see Table 1 in the main paper). This shows our method maintains good PSNR/SSIM (average rank of 2 across all metrics among SOTA Diffusion-based methods). On the other hand, it is indeed true that our method—like all Diffusion-based ISR methods—achieves significantly lower PSNR/SSIM metrics compared to GAN-based ISR algorithms (average PSNR/SSIM diffusion 23.21/0.5791 and GAN 24.06/0.6176) for Diffusion-based ISR methods is 23.21/0.5791. This difference highlights **a potential limitation inherent** to the fundamental characteristics of diffusion models. To the best of our knowledge, there is no optimal solution currently to further improve these two metrics.

To explain this, we guess the reason might be that GAN-based methods are typically trained by directly employing L2 and perceptual loss between the predicted output and the GT, which are closely related to the computation of PSNR and LPIPS, respectively. In contrast, Diffusion-based methods optimize solely based on the L2 loss between the predicted noise and GT noise. Such a fundamental difference in loss design may become **a critical factor** contributing to the relatively weaker performance of diffusion models on metrics such as PSNR/SSIM, compared to GAN-based methods. Unfortunately, thoroughly resolving this issue may require fundamentally changing the training paradigm of diffusion models, which goes beyond the scope of this work. Therefore, this study primarily focuses on enhancing perception metrics while maintaining competitive PSNR/SSIM metrics. We will explore this matter more deeply in the future.

G Limitations and Future Work

Although our proposed TFDSR method has achieved significant results, it still has some limitations. Similar to existing studies on natural scene super-resolution, this work focuses solely on natural image datasets and synthetic datasets for super-resolution processing. However, how to achieve large-scale super-resolution on AI-generated datasets remains a direction worthy of further exploration. Additionally, the potential of frequency domain methods in other vision tasks based on diffusion models has not yet been fully explored. In future work, we will expand the application scope and delve deeper into its potential across a wider range of vision-related tasks, such as advanced image editing and image generation.

References

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

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 *ECCV*, 2024.

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*, abs/2310.12004, 2023.

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 *CVPR*, 2024.

Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *AAAI*, 2023.

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

Chenyang Wang, Junjun Jiang, Zhiwei Zhong, and Xianming Liu. Spatial-frequency mutual learning for face super-resolution. In *CVPR*, 2023.

Boyang Wang, Fengyu Yang, Xihang Yu, Chao Zhang, and Hanbin Zhao. APISR: anime production inspired real-world anime super-resolution. In *CVPR*, 2024.

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

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

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