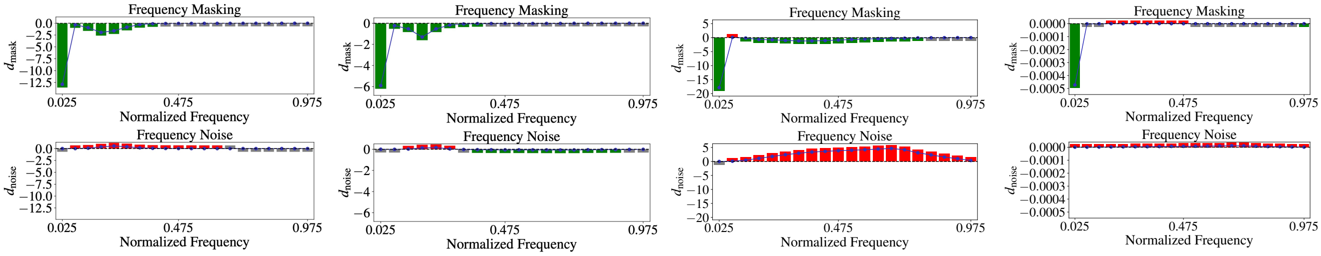
On the Effectiveness of Spectral Discriminators for Perceptual Quality Improvement Supplementary Material

Xin Luo Yunan Zhu Shunxin Xu Dong Liu
University of Science and Technology of China, Hefei, China
{xinluo, zhuyn, sxu}@mail.ustc.edu.cn, dongeliu@ustc.edu.cn
https://github.com/LuooooXin/DualFormer



(a) Spatial Transformer (b) Spatial MLPMixer (c) VGG (d) U-Net

Figure 1: **Robustness behavior of various spatial discriminators**. The Transformer/MLPMixer works better at identifying absence in the middle-frequency range, and the CNN is aware of the higher-range spectrum. Also, CNN works poorly at identifying frequency noise compared to Transformer/MLPMixer. The U-Net has lower spectra perception compared to VGG due to its residual structure [10].

1. Architecture-Related Robustness

In main body of the text, we argue that the spatial discriminator excels at identifying low-frequency masking, while the spectral discriminator is good at identifying high-frequency noise. This section substantiates that the aforementioned phenomenon is independent of the specific network architecture.

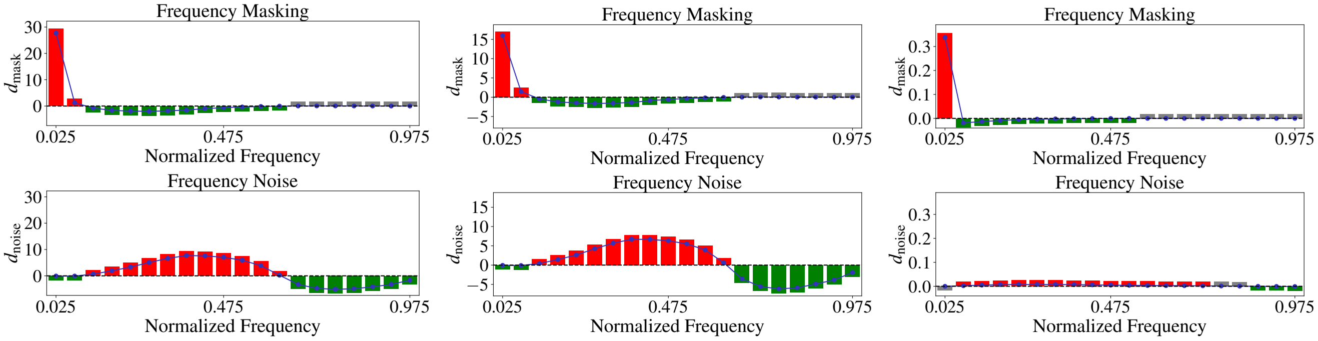
1.1. Spatial Discriminators

Fig. 1 illustrates the robustness of various spatial discriminators under frequency perturbations. All four representative network architectures demonstrate a similar tendency to capture low-frequency masking. Nevertheless, subtle yet critical distinctions exist between these architectures. While Transformer performs like a low-pass filter [5], relying more on low-frequency information, it can identify a narrower range of frequency masking than typical CNNs such as VGG [8]. Similarly, MLPMixer behaves like Transformer due to their similar high-level architecture design. In contrast, VGG, which is a CNN network, has a broader spectrum perception range. The U-Net [6], which is a residual structured network, has a weaker tendency to capture high-frequency components [10], and therefore behaves more like the Transformer [3]/MLPMixer [9]. These phenomena align with those observed in other studies [10, 4, 7, 1].

1.2. Spectral Discriminators

As evidenced by Fig. 2, comparable to the scenario of spatial discriminators, spectral discriminators also exhibit similar behaviors, *i.e.*, they are unable to differentiate the absence of low frequencies. Specifically, just as they do in the spatial domain, both Transformer and MLPMixer exhibit consistent behavior in the frequency domain, as they both effectively learn to discriminate against high-frequency noise. While Spectral MLP performs similarly to Transformer/MLPMixer in terms of frequency masking, it fails to learn to recognize high-frequency noise, further validating the effectiveness of our Spectral Transformer.

In conclusion, there exists a fundamental disparity between the spatial and spectral discriminator. Specifically, the spatial discriminator is an expert at discriminating low-frequency masking, while the spectral discriminator performs better in distinguishing high-frequency noise, and the

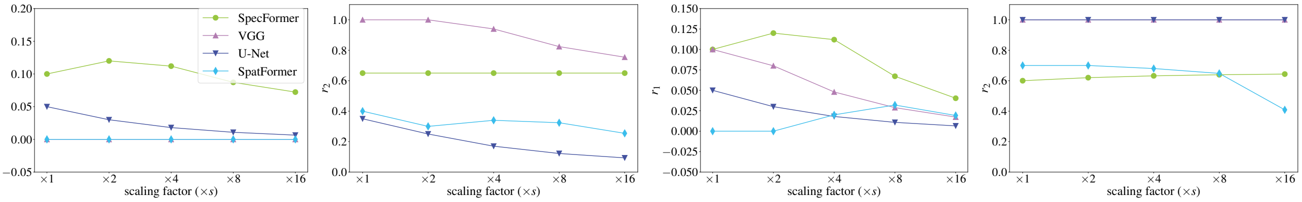


(a) Spectral Transformer

(b) Spectral MLPMixer

(c) Spectral MLP

Figure 2: **Robustness behavior of various spectral discriminators**. Overall, these architectures exhibit similar three-stage behaviors. Specifically, Transformer and MLPMixer perform almost identically, while MLP fails to learn to discriminate high-frequency noise effectively.



(a) r_1 on frequency masking

(b) r_2 on frequency masking

(c) r_1 on frequency noise

(d) r_2 on frequency noise

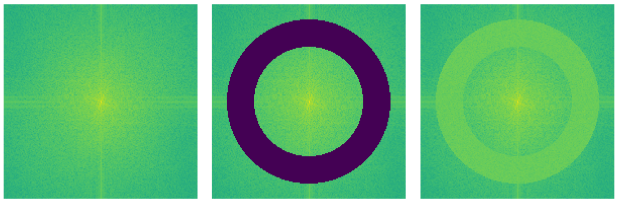
Figure 3: **The shifting behavior of the three frequency ranges varying scaling factors**. The three frequency ranges are $[0, r_1)$, $[r_1, r_2)$, and $[r_2, 1]$. SpatFormer/SpecFormer denotes Transformer applied to the spatial/frequency domain. As the scaling factor grows, the optimization goal of generator migrates from distortion to perception (similar to increasing the weight of the perception term). Consequently, the boundaries of the three frequency ranges shift to the left (r_1 and r_2 decrease). The tiny vibrations may be related to the stochasticity of the training.

architecture contributes to specific behavior. Therefore, it is crucial to consider the specific requirements of the task and the characteristics of the input data when choosing a discriminator architecture. Moreover, our findings can guide future research in developing discriminators that are better suited for specific tasks and data types.

2. The Generalizability of the Three Frequency Ranges Phenomenon

We have demonstrated that both the generator and discriminator exhibit three-range behavior in the frequency domain, and we have explained this phenomenon from the frequency perspective of the PD tradeoff. Nevertheless, in the main body of the text, we conducted our study using a $\times 4$ SR as an example. In order to demonstrate the generalizability of the three-range behavior in the frequency domain, we investigated how the scaling factor of SR influences various discriminators. Specifically, we defined the boundary of the three frequency ranges as r_1 and r_2, where $r_1 \geq 0$, $r_1 \leq r_2$, and $r_2 \leq 1$. These three frequency ranges are in the radius intervals $[0, r_1)$, $[r_1, r_2)$, and $[r_2, 1]$, respectively. Please refer to Fig. 2b for an illustration of the properties of each range.

Let's start by taking a global view. Fig. 3 shows the boundary of the three frequency ranges (r_1 and r_2), which will shift to the left as the scaling factor s increases. This can be explained from the frequency perspective of PD tradeoff. Initially, as the scaling factor s increases, the information accessible in the input image diminishes. Subsequently, the low-frequency part that the generator can perfectly recover also decreases, and the perception term gradually dominates optimization. As a result, r_1 decreases. Moreover, the limited capacity of the generator can cause a decrease in r_2, considering the decrease of r_1, though the decreasing trend of r_2 is relatively mild compared to r_1. Furthermore, when s approaches infinity ($\times\infty$ SR), the input contains scarcely any information, thereby equivalent to an unconditional image generation task. In this scenario, the discriminator can identify a significant amount of frequency noise but can only discriminate a small fraction of the frequency masking. Chen *et al.* [2] also observed this limiting case in image generation.



(a) GT (b) After masking (c) After noise

Figure 4: **Intuitive visualization of frequency masking (b) and noise (c).**

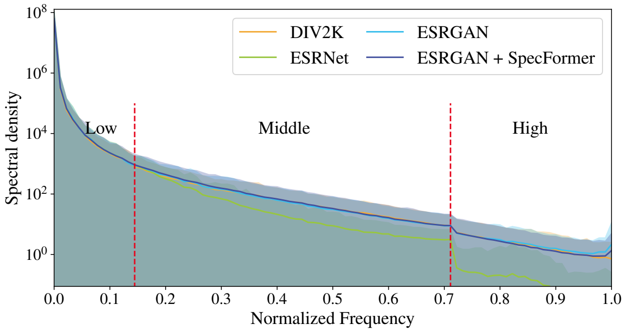


Figure 5: **Spectral Profile of SR Models under Bicubic Degradation**. The spatial discriminator could match the real spectra well under simple degradation (Bicubic downsampling).

3. The Visual Effects of Frequency Perturbations

We investigated the behavior of discriminator by examining its performance under two representative frequency perturbations, and found differences in the capabilities of spatial and spectral discriminators. Here, we provide more comprehensible visualizations, as exemplified in Fig. 4, where frequency masking is the removal of a circular ring with a certain radius from the spectrogram, and frequency noise is the addition of noise within a circular ring with a certain radius in the spectrogram. In addition, we also demonstrate the effect of frequency perturbation on two representative images in Fig. 6. Among them, high-frequency perturbations are relatively difficult for the human eye to perceive, while other perturbations have a significant impact on human perception.

4. Spectral Profile of SR Models under Bicubic Degradation

We have observed that SR networks exhibit poor spectral alignment with real spectra in real-world SR scenarios, which prompted us to introduce a spectral discriminator to improve the spectral alignment of SR networks. Nevertheless, we acknowledge that this issue is not as severe in the case of simple degradation, such as Bicubic degradation. As depicted in Fig. 5, the SR images produced by ESRGAN [12] already match real images in the low and middle-frequency ranges. While our Spectral Transformer mitigates the problem of excessive preference for high-frequency content by the spatial discriminator to some extent, its impact on quantitative metrics and human perception may not be significant.

Discriminator	Params[M]	FLOPs[G]	Activations[G]
VGG [12]	21.1	8728.63	9.57
U-Net [11]	4.4	24776.00	22.56
SpecFormer/8	2.0	2709.60	33.32
SpecFormer/32	2.2	93.41	0.63

Table 1: **Efficiency performance of various discriminators.** Metrics are evaluated on images of size 256×256.

	Ground Truth	Low-Quality	Super-Resolution
Real-ESRGAN [11]	0.77	0.92	0.16
Real-ESRGAN + SpecFormer	0.52	0.43	0.44

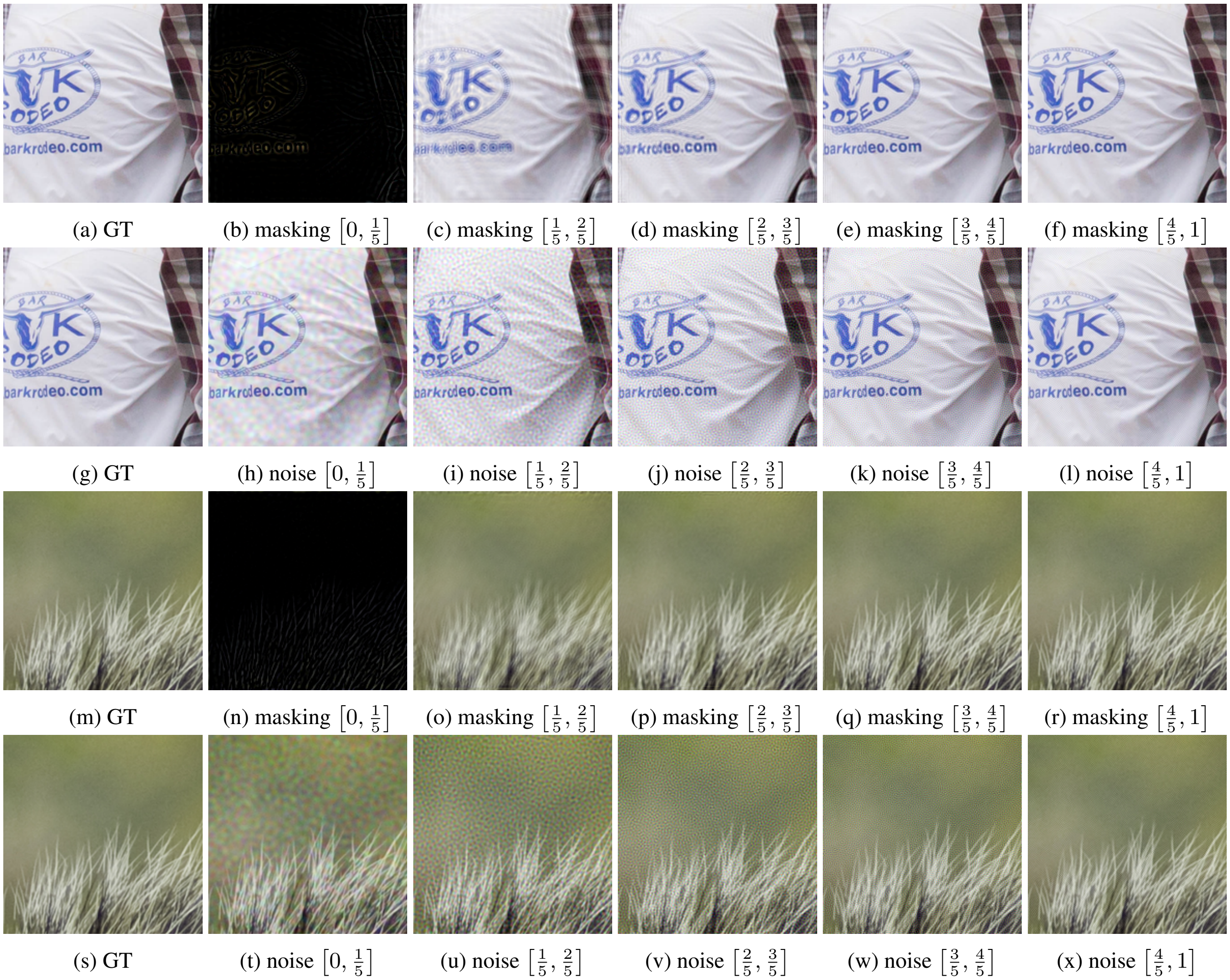
Table 2: **The average scores of discriminators** w.r.t. three types of images. The spectral discriminator, *i.e.*, PSM-T, mitigates the high-frequency flaw of Real-ESRGAN's spatial discriminator.

5. Efficiency Analysis of Discriminators

To compare the efficiency differences between the Spectral Transformer and other commonly used discriminators in SR, we consider three metrics: the total number of parameters, the number of floating point operations (FLOPs), and the number of elements of all outputs of convolutional layers (activations). As is evidenced in Tab. 1, our discriminator is, in fact, highly efficient due to our utilization of a small number of dimensions and a relatively large patch size. For our SR experiments, we employed a patch size of 32×32 for both the Spatial Transformer and the Spectral Transformer. Therefore, the number of parameters in our discriminator is 4.4M, the number of FLOPs is 186.82G, and the number of activations is 1.26G. Among these, the number of parameters in our discriminator is equivalent to that of U-Net, while the FLOPs and Activations are significantly lower than those of VGG and U-Net.

6. The Spectral Discriminator solve the high-frequency noise problem

It is the extreme preference of Real-ESRGAN's spatial discriminator for high-frequency information that motivated us to introduce the spectral discriminator. To confirm that the spectral discriminator can address this problem in practical applications, we introduce spectral Transformer to train Real-ESRGAN. As demonstrated in Tab. 2, our approach yields the highest scores for ground truth im-



(a) GT (b) masking $\left[0, \frac{1}{5}\right]$ (c) masking $\left[\frac{1}{5}, \frac{2}{5}\right]$ (d) masking $\left[\frac{2}{5}, \frac{3}{5}\right]$ (e) masking $\left[\frac{3}{5}, \frac{4}{5}\right]$ (f) masking $\left[\frac{4}{5}, 1\right]$

(g) GT (h) noise $\left[0, \frac{1}{5}\right]$ (i) noise $\left[\frac{1}{5}, \frac{2}{5}\right]$ (j) noise $\left[\frac{2}{5}, \frac{3}{5}\right]$ (k) noise $\left[\frac{3}{5}, \frac{4}{5}\right]$ (l) noise $\left[\frac{4}{5}, 1\right]$

(m) GT (n) masking $\left[0, \frac{1}{5}\right]$ (o) masking $\left[\frac{1}{5}, \frac{2}{5}\right]$ (p) masking $\left[\frac{2}{5}, \frac{3}{5}\right]$ (q) masking $\left[\frac{3}{5}, \frac{4}{5}\right]$ (r) masking $\left[\frac{4}{5}, 1\right]$

(s) GT (t) noise $\left[0, \frac{1}{5}\right]$ (u) noise $\left[\frac{1}{5}, \frac{2}{5}\right]$ (v) noise $\left[\frac{2}{5}, \frac{3}{5}\right]$ (w) noise $\left[\frac{3}{5}, \frac{4}{5}\right]$ (x) noise $\left[\frac{4}{5}, 1\right]$

Figure 6: **The effects of frequency masking and noise on two representative images**.

ages, followed by super-resolution, and the lowest scores for low-quality images, as we had anticipated. Therefore, we can conclude that the spectral discriminator is capable of mitigating the flaw of the spatial discriminator on high frequencies.

References

[1] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of Vision Transformer and MLP-Mixer to CNNs. In *BMVC*, 2021. 1

[2] Yuanqi Chen, Ge Li, Cece Jin, Shan Liu, and Thomas Li. SSD-GAN: Measuring the realness in the spatial and spectral domains. In *AAAI*, 2021. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[4] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, 2021. 1

[5] Namuk Park and Songkuk Kim. How do vision transformers work? In *ICLR*, 2022. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *IEEE MICCAI*, 2015. 1

[7] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021. 1

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[9] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica

Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. MLP-Mixer: An all-MLP architecture for vision. In *NeurIPS*, 2021. 1

[10] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *CVPR*, 2020. 1

[11] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021. 3

[12] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 3