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

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

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.

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 frequency perspective of the PD tradeoff. Nevertheless, in the main, and we have explained this phenomenon from the frequency perspective of PD tradeoff. Consequently, the boundaries of the three frequency ranges shift to the left ($r_1$, $r_2$, and $r_3$ 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.
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 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.

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 $32 \times 32$ 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-

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net [11]</td>
<td>4.4</td>
<td>24778.00</td>
<td>22.56</td>
</tr>
<tr>
<td>SpecFormer/8</td>
<td>2.0</td>
<td>2709.60</td>
<td>33.32</td>
</tr>
<tr>
<td>SpecFormer/32</td>
<td>2.2</td>
<td>93.41</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 1: Efficiency performance of various discriminators. Metrics are evaluated on images of size $256 \times 256$.

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Low-Quality</th>
<th>Super-Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-ESRGAN</td>
<td>0.77</td>
<td>0.92</td>
</tr>
<tr>
<td>Real-ESRGAN + SpecFormer</td>
<td>0.52</td>
<td>0.43</td>
</tr>
</tbody>
</table>

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.
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


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

