Supplementary Material for Feature Modulation Transformer: Cross-Refinement of Global Representation via High-Frequency Prior for Image Super-Resolution

1. Additional Ablation Studies

To further investigate the capabilities of CRAFT, we conducted additional ablation studies. Specifically, we trained several models on the DIV2K dataset [10] and evaluated their performance on five commonly used benchmarks, including Set5 [2], Set14 [11], BSD100 [8], Urban100 [5], and Manga109 [9], all with a magnification factor of ×4. We randomly cropped the images into 64 × 64 sub-image patches and performed data augmentation such as random horizontal flipping and 90° rotation. We set the total number of training iterations to 300K, and used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to minimize the $\ell_1$ loss. The batch size was set to 64, and the initial learning rate was set to $2 \times 10^{-4}$.

**Impact of Channel Number.** To explore the effect of the number of channels on the performance of CRAFT, we conducted four groups of experiments. Specifically, we set the number of channels to 36, 48, 60, and 72 and evaluated each model on the five benchmarks mentioned above. The results in Table 1 indicate that increasing the number of channels leads to improved performance.

**Impact of CRFB Number.** We also investigated the impact of the number of CRFB blocks on the performance of CRAFT. We stacked different numbers of CRFB blocks and evaluated their performance on the five benchmarks. The results in Table 1 demonstrate that adding more CRFB blocks leads to better performance.

**Visualization of the Effectiveness of High-Frequency Prior.** We conducted visual experiments to demonstrate the effectiveness of introducing a high-frequency prior. Figure 1 shows the results, where w/o H indicates the model without the high-frequency prior, and vice versa. We observed that introducing the high-frequency prior led to a better-detailed representation. In addition, we formulated the spectrum of the two models as

$$\Phi(w/H) = FFT(w/H)$$
$$\Phi(w/o H) = FFT(w/o H).$$

After that, we get the residual spectrum map from $\Phi(w/H)$ and $\Phi(w/o H)$. It can be formulated as

$$R(w/H, w/o H) = |\Phi(w/H) - \Phi(w/o H)|,$$

where $R(\cdot)$ denotes the process of generating the residual spectrum map. The residual spectrum map illustrates that including a high-frequency prior in CRAFT results in a stronger high-frequency response, which suggests that the restoration of high-frequency components is enhanced.

2. More Visual Comparisons

We supply more visual comparisons with other methods in Figures 2 and Figure 3.

References

Table 1. Ablation studies on five benchmarks. The total number of training iterations was set to 300K. Params represents the total number of network parameters.

<table>
<thead>
<tr>
<th>Model</th>
<th>Params (M)</th>
<th>Number of Channels</th>
<th>Number of CRFBs</th>
<th>Set5 (PSNR/SSIM)</th>
<th>Set14 (PSNR/SSIM)</th>
<th>BSD100 (PSNR/SSIM)</th>
<th>Urban100 (PSNR/SSIM)</th>
<th>Manga109 (PSNR/SSIM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRAFT-36-4</td>
<td>0.44</td>
<td>36</td>
<td>4</td>
<td>32.57/0.8968</td>
<td>28.74/0.7854</td>
<td>27.64/0.7392</td>
<td>26.34/0.7932</td>
<td>30.89/0.9138</td>
</tr>
<tr>
<td>CRAFT-48-4</td>
<td>0.75</td>
<td>48</td>
<td>4</td>
<td>32.48/0.8891</td>
<td>28.81/0.7867</td>
<td>27.70/0.7409</td>
<td>26.54/0.7987</td>
<td>31.11/0.9155</td>
</tr>
<tr>
<td>CRAFT-60-4</td>
<td>1.16</td>
<td>60</td>
<td>4</td>
<td>32.54/0.8991</td>
<td>28.89/0.7885</td>
<td>27.73/0.7423</td>
<td>26.66/0.8018</td>
<td>31.19/0.9166</td>
</tr>
<tr>
<td>CRAFT-72-4</td>
<td>1.64</td>
<td>72</td>
<td>4</td>
<td>32.65/0.9003</td>
<td>28.86/0.7879</td>
<td>27.76/0.7427</td>
<td>26.67/0.8026</td>
<td>31.35/0.9186</td>
</tr>
<tr>
<td>CRAFT-48-2</td>
<td>0.44</td>
<td>48</td>
<td>2</td>
<td>32.33/0.8964</td>
<td>28.74/0.7845</td>
<td>27.65/0.7389</td>
<td>26.29/0.7914</td>
<td>30.81/0.9178</td>
</tr>
<tr>
<td>CRAFT-48-4</td>
<td>0.75</td>
<td>48</td>
<td>4</td>
<td>32.48/0.8981</td>
<td>28.81/0.7867</td>
<td>27.70/0.7409</td>
<td>26.54/0.7987</td>
<td>31.11/0.9155</td>
</tr>
<tr>
<td>CRAFT-48-6</td>
<td>1.07</td>
<td>48</td>
<td>6</td>
<td>32.50/0.8985</td>
<td>28.86/0.7875</td>
<td>27.72/0.7417</td>
<td>26.65/0.8024</td>
<td>31.27/0.9178</td>
</tr>
<tr>
<td>CRAFT-48-8</td>
<td>1.38</td>
<td>48</td>
<td>8</td>
<td>32.60/0.8998</td>
<td>28.89/0.7884</td>
<td>27.74/0.7425</td>
<td>26.69/0.8035</td>
<td>31.29/0.9179</td>
</tr>
</tbody>
</table>

Figure 1. Comparison of visual quality with and without the HFERB blocks. Models with and without HFERB blocks are denoted as w/ H and w/o H, respectively. The symbols $\Phi(w/o H)$ and $\Phi(w/H)$ represent the spectra of the models without and with HFERB blocks, respectively. The evaluation is conducted using a magnification factor of $\times 4$.


Figure 2. Comparison of visual quality with state-of-the-art methods, evaluated using a magnification factor of $\times 4$.

Figure 3. Comparison of visual quality with state-of-the-art methods, evaluated using a magnification factor of $\times 4$. 