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 $\times 4$. 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 ℓ_1 loss. The batch size was set to 64, and the initial learning rate was set to 2×10^{-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

$$\begin{aligned} \Phi(\textit{w/ } H) &= FFT(\textit{w/ } H) \\ \Phi(\textit{w/o } H) &= FFT(\textit{w/o } H). \end{aligned} \tag{1}$$

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

$$R(\textit{w/ } H, \textit{w/o } H) = |\Phi(\textit{w/ } H) - \Phi(\textit{w/o } H)|, \tag{2}$$

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

[1] Namhyuk Ahn, Byungkon Kang, and Kyung Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 256–272, 2018. 3

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–10, 2012. 1

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 3

[4] Guangwei Gao, Zhengxue Wang, Juncheng Li, Wenjie Li, Yi Yu, and Tieyong Zeng. Lightweight bimodal network for single-image super-resolution via symmetric cnn and recursive transformer. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. 3

[5] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 1

[6] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision Workshop (ICCVW)*, pages 1833–1844, 2021. 3

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.

Model	Params (M)	Number of Channels	Number of CRFBs	Set5 (PSNR/SSIM)	Set14 (PSNR/SSIM)	BSD100 (PSNR/SSIM)	Urban100 (PSNR/SSIM)	Manga109 (PSNR/SSIM)
CRAFT-36-4	0.44	36	4	32.37/0.8968	28.74/0.7851	27.64/0.7392	26.34/0.7932	30.89/0.9138
CRAFT-48-4	0.75	48	4	32.48/0.8981	28.81/0.7867	27.70/0.7409	26.54/0.7987	31.11/0.9155
CRAFT-60-4	1.16	60	4	32.54/0.8991	28.89/0.7885	27.73/0.7423	26.66/0.8018	31.19/0.9166
CRAFT-72-4	1.64	72	4	32.65/0.9003	28.86/0.7879	27.76/0.7427	26.67/0.8026	31.35/0.9186
CRAFT-48-2	0.44	48	2	32.35/0.8964	28.74/0.7845	27.65/0.7389	26.29/0.7914	30.81/0.9118
CRAFT-48-4	0.75	48	4	32.48/0.8981	28.81/0.7867	27.70/0.7409	26.54/0.7987	31.11/0.9155
CRAFT-48-6	1.07	48	6	32.50/0.8985	28.86/0.7875	27.72/0.7417	26.65/0.8024	31.27/0.9178
CRAFT-48-8	1.38	48	8	32.60/0.8998	28.89/0.7884	27.74/0.7425	26.69/0.8035	31.29/0.9179

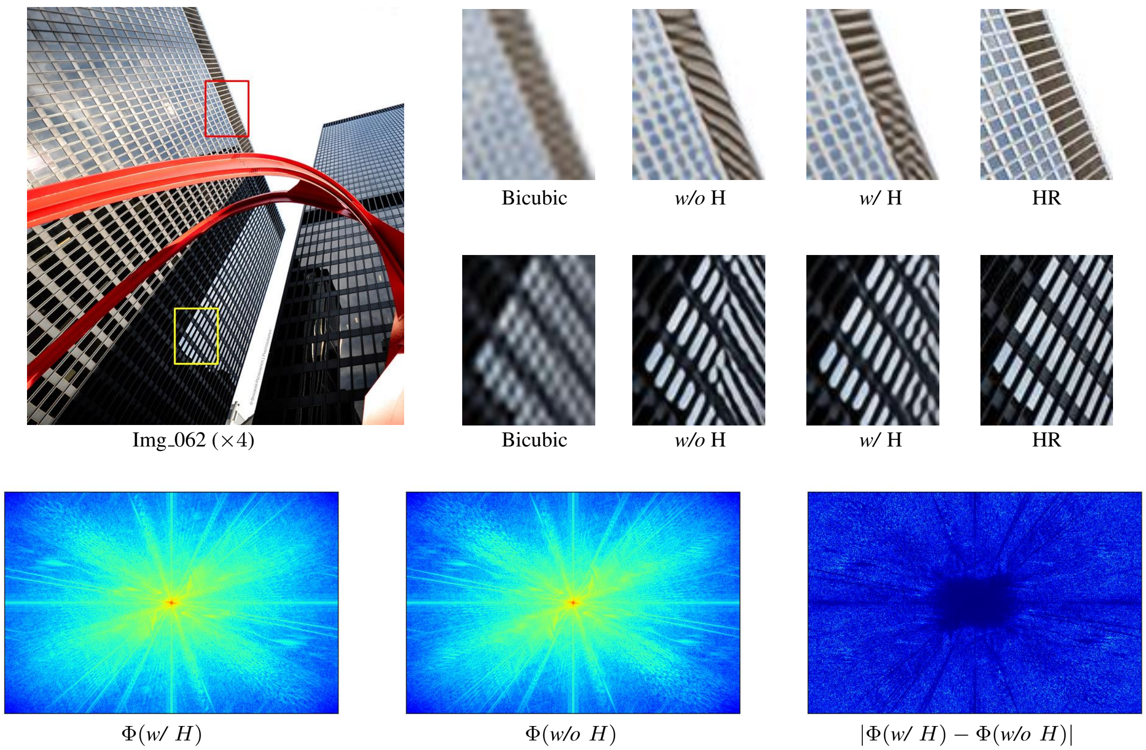

Figure 1. Comparison of visual quality with and without the HFERB blocks. Models with and without HFERB blocks are denoted as *w/ H* and *w/o H*, respectively. The symbols $\Phi(\textit{w/o } H)$ and $\Phi(\textit{w/ } H)$ represent the spectra of the models without and with HFERB blocks, respectively.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

[7] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 457–466, 2022. 3

[8] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, pages 416–423, 2001. 1

[9] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 1

[10] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 114–125, 2017. 1

[11] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces (ICCS)*, pages 711–730, 2010. 1

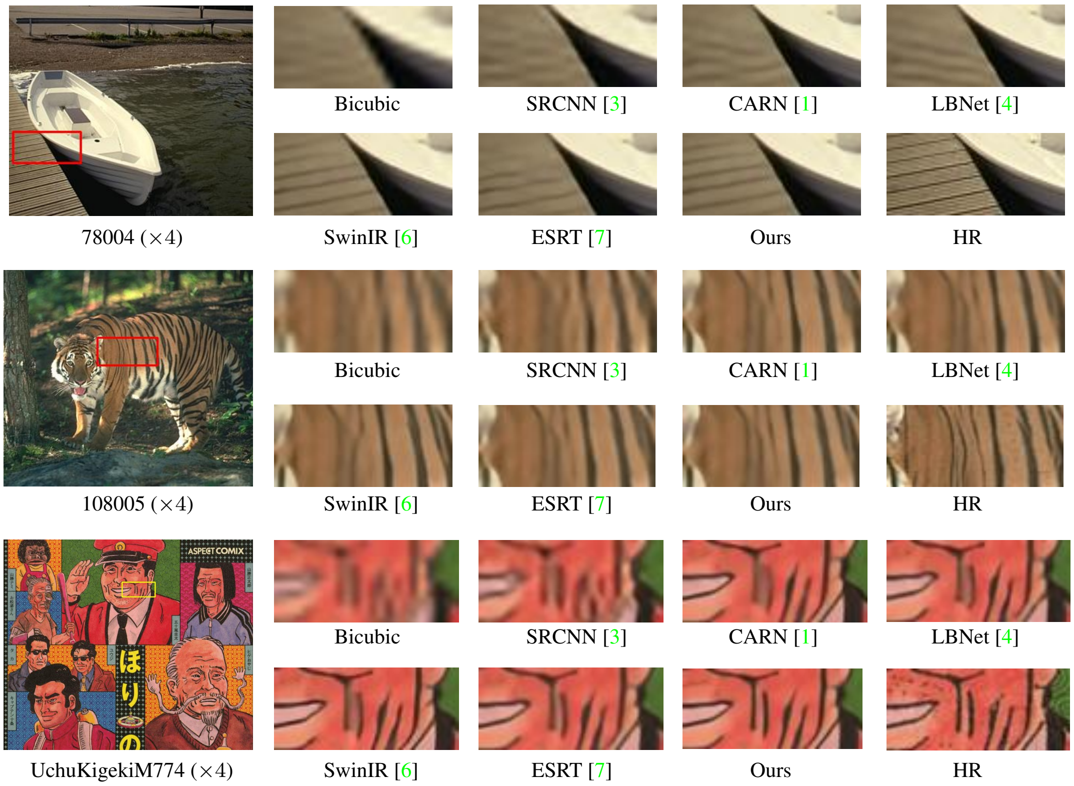

Figure 2. Comparison of visual quality with state-of-the-art methods, evaluated using a magnification factor of $\times 4$.

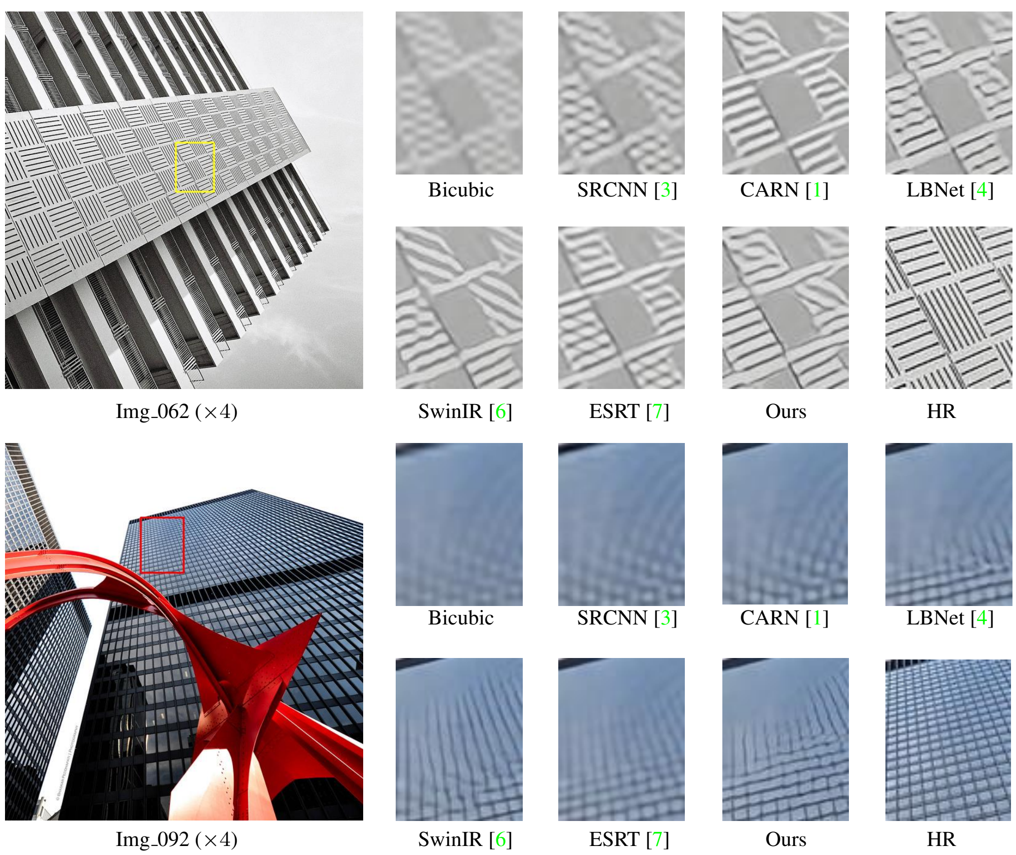

Figure 3. Comparison of visual quality with state-of-the-art methods, evaluated using a magnification factor of $\times 4$.