Supplementary Material for "Efficient Face Super-Resolution via Wavelet-based Feature Enhancement Network"

Anonymous Author(s)

1 2D DISCRETE WAVELET TRANSFORM

2D discrete wavelet transform [9] (2D-DWT) is a mathematical signal processing technique employed to decompose an image into wavelet components of varying frequencies. This technique allows the analysis of image features at different frequency scales. The principle of 2D-DWT is illustrated in Fig. 1, and we explain the process using the commonly used Haar wavelet transform. 2D-DWT can be conceptualized as applying a 1D wavelet transform (1D-DWT) to the image using a filter in both the row and column directions. Assuming the size of the input face image $\boldsymbol{X} \in \mathbb{R}^{H \times W}$, where $\boldsymbol{X}(i,j)$ represents the pixel at position (i,j). Initially, 1D-DWT is applied to $\boldsymbol{X}$ in the row direction, resulting in:

$$\boldsymbol{X}_L(i,j) = \mathcal{L}(k) \sum_{k=1}^{2j} \boldsymbol{X}(i, 2j-k), \tag{1}$$

$$\boldsymbol{X}_H(i,j) = \mathcal{H}(k) \sum_{k=1}^{2j} \boldsymbol{X}(i, 2j-k), \tag{2}$$

Where $\mathcal{L}$ represents low-pass filter, $\mathcal{L} = [1,1]$, $\mathcal{H}$ represents high-pass filter, $\mathcal{H} = [1,-1]$, $\boldsymbol{X}_L \in \mathbb{R}^{\frac{H}{2} \times W}$ is row low-frequency component after 1D-DWT, and $\boldsymbol{X}_H \in \mathbb{R}^{\frac{H}{2} \times W}$ is row high-frequency component after 1D-DWT. Next, 1D-DWT on columns is performed on $\boldsymbol{X}_L$ and $\boldsymbol{X}_H$ to get output:

$$\boldsymbol{A}_{LL}(i,j) = \mathcal{L}(k) \sum_{k=1}^{2i} \boldsymbol{X}_L(2i-k, j), \tag{3}$$

$$\boldsymbol{H}_{LH}(i,j) = \mathcal{H}(k) \sum_{k=1}^{2i} \boldsymbol{X}_L(2i-k, j), \tag{4}$$

$$\boldsymbol{V}_{HL}(i,j) = \mathcal{L}(k) \sum_{k=1}^{2i} \boldsymbol{X}_H(2i-k, j), \tag{5}$$

$$\boldsymbol{D}_{HH}(i,j) = \mathcal{H}(k) \sum_{k=1}^{2i} \boldsymbol{X}_H(2i-k, j), \tag{6}$$

where $\boldsymbol{A}_{LL} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2}}$ denotes the low-frequency part of the image, which contains the overall structure and general shape information in the image, $\boldsymbol{H}_{LH} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2}}$ denotes the high-frequency information in the horizontal direction of the image, including the horizontal variation of edges and details, $\boldsymbol{V}_{HL} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2}}$ denotes the high-frequency information in the vertical direction of the image, including the vertical variation of edges and details, and $\boldsymbol{D}_{HH} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2}}$ denotes the high-frequency information in the diagonal direction of the image, including the diagonal variation of edges and details.

On the contrary, 2D inverse discrete wavelet transform (2D-IDWT) diminishes the four frequency-domain components $\{\boldsymbol{X}_{LL}, \boldsymbol{X}_{LH}, \boldsymbol{X}_{HL}, \boldsymbol{X}_{HH}\}$ back to the input $\boldsymbol{X}$, making the entire process of 2D-DWT and 2D-IDWT closed and lossless. Fig. 2 shows an example of 2D-DWT and 2D-IDWT performed on a face image, where the meaning of the variables is consistent with the above formulation.

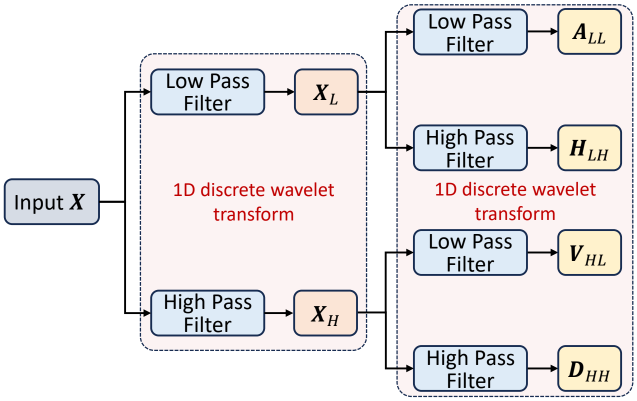

Figure 1: Processing of 2D discrete wavelet transform (2D-DWT), where consists of two 1D discrete wavelet transform (1D-DWT).

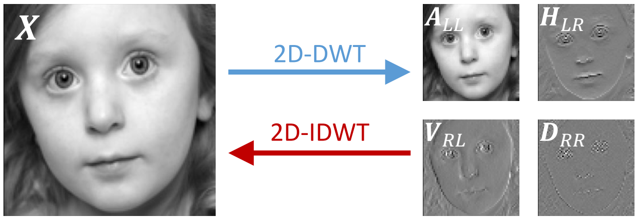

Figure 2: 2D discrete wavelet transform (2D-DWT) and 2D inverse discrete wavelet transform (2D-IDWT) processes are applied to a face image. The 2D-DWT decomposes the input face image X into one low-frequency component A_{LL} and three high-frequency components $\{H_{LR}, V_{RL}, D_{RR}\}$. The 2D-IDWT reduces these frequency domain components back to the original face image X.

2 MORE QUALITATIVE COMPARISONS

In this section, we additionally add a series of qualitative comparisons with existing methods, including prior-based methods like FSRNet [3], FACN [11] and DIC [8], attention-based methods like SPARNet [2] and AD-GNN [1], as well as Transformer-based methods such as Restormer [12], LAAT [6] and SFMNet [10]. Specifically, in Fig 3, we give more qualitative comparisons for ×8 FSR on CelebA [7] and Helen [5] test datasets. In the facial region where the eyes are located, our method excels in recovering finer details

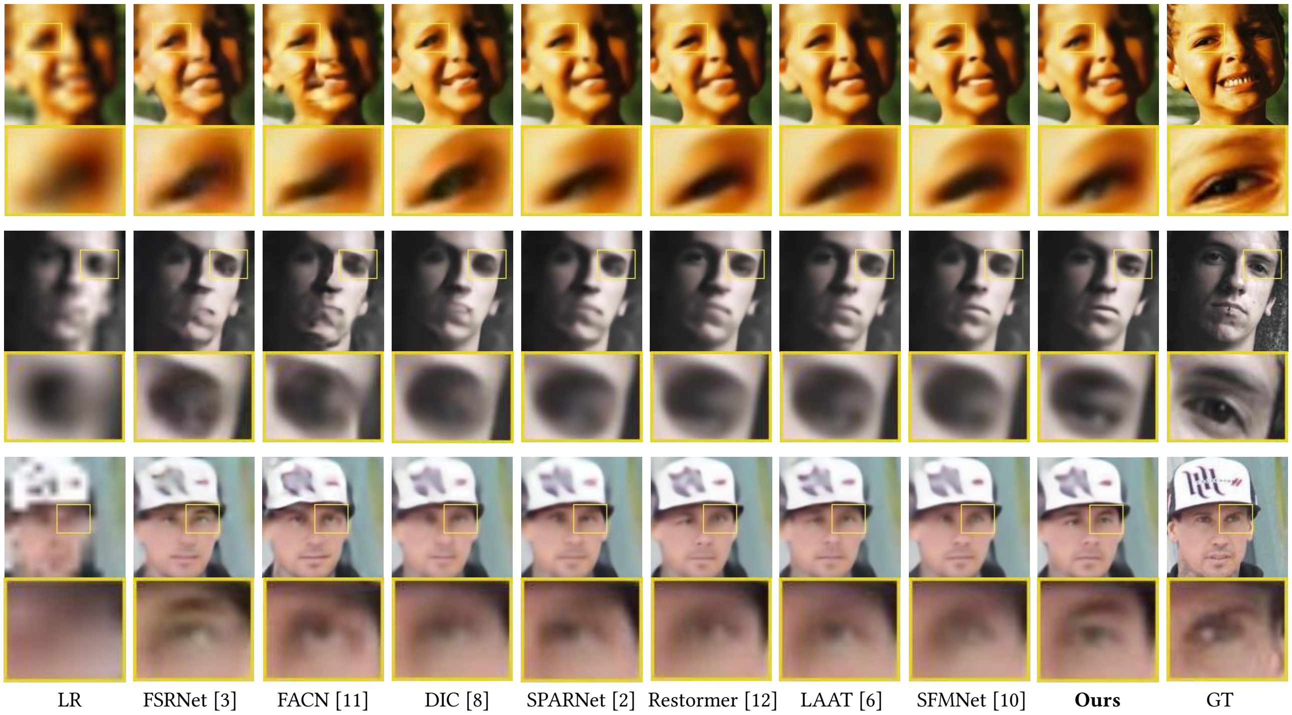

Figure 3: Qualitative comparison for ×8 FSR on CelebA [7] and Helen [5] test datasets. Our method recovers face images that are closer to ground truth and contain more facial details than existing methods.

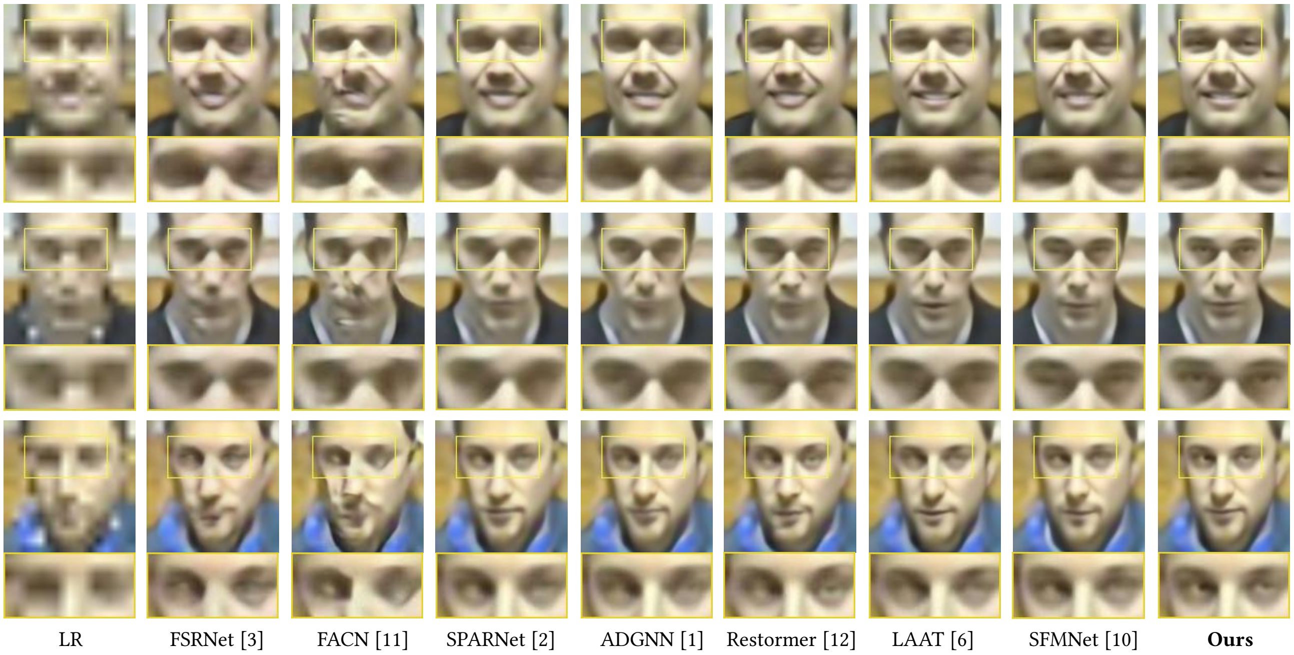

Figure 4: Qualitative comparison for ×8 FSR on SCface [4] test dataset. Our method can restore the clearer face components, especially the eye region, which is critical for downstream face recognition tasks.

such as the pupil of the eye. Moreover, the details we restore are closer to ground truth than existing methods.

In Fig 4, we give more qualitative comparisons for ×8 FSR on SCface [4] test datasets. In the current scenario with high levels of

Table 1: Ablation studies of down-sampling mechanisms in our method on Helen [5] test dataset, where "Stride" denotes stride convolution, "Avgpool" denotes average pool, "Bicubic" denotes bicubic downsample, and "WFD" denotes our proposed wavelet feature downsample module. All downsampling scales are set to 2.

Methods	Stride	Avgpool	Bicubic	WFD	Params	FLOPs	PSNR / SSIM
Case1	✓	✗	✗	✗	0.830M	1.131G	26.22 / 0.7743
Case2	✗	✓	✗	✗	0.830M	1.129G	26.26 / 0.7747
Case3	✗	✗	✓	✗	0.830M	1.129G	26.21 / 0.7731
Ours	✗	✗	✗	✓	0.848M	1.164G	**26.36 / 0.7795**

blurring, prior-based methods exhibit poor performance, showing severe distortion. Transformer-based methods fare better by recovering general contours, yet they struggle with finer details like eye pupils. Our method excels in recovering above facial details, facilitating enhanced accuracy in downstream tasks such as recognition, significantly outperforming existing methods.

3 MORE ABLATION STUDIES

In the main text, we only provide an ablation study using stride convolution instead of our proposed wavelet feature downsample (WFD) module for downsampling. TABLE 1 is additionally supplemented with experimental results of encoder stage downsample using average pool (Avgpool), and bicubic downsample (Bicubic). Compared with these common downsample strategies, our proposed WFD module can minimize the adverse effects of downsample on FSR reconstruction, achieving a performance improvement of more than 0.1dB PSNR with only a small number of params and FLOPs gains. This experiment demonstrates the advantages of our proposed WFD over existing downsample strategies. Additionally, it also confirms that reducing the feature corruption caused by downsampling in the encoding stage can significantly improve the model's performance and further enhance the model's efficiency.

4 LIMITATION

Our method aims to preserve identity accuracy. It can restore face images with higher fidelity compared to existing generative prior-based methods, but its clarity is not as sharp as theirs. In this context, we will further discuss how to integrate the generative prior based on this method to enhance clarity while maintaining the identity accuracy of the restored face images.

REFERENCES

[1] Qiqi Bao, Bowen Gang, Wenming Yang, Jie Zhou, and Qingmin Liao. 2022. Attention-driven graph neural network for deep face super-resolution. *IEEE Transactions on Image Processing* 31 (2022), 6455–6470.

[2] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K Wong. 2020. Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing* 30 (2020), 1219–1231.

[3] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*. 2492–2501.

[4] Mislav Grgic, Kresimir Delac, and Sonja Grgic. 2011. SCface–surveillance cameras face database. *Multimedia Tools and Applications* 51 (2011), 863–879.

[5] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. 2012. Interactive facial feature localization. In *ECCV*. Springer, 679–692.

[6] Guanxin Li, Jingang Shi, Yuan Zong, Fei Wang, Tian Wang, and Yihong Gong. 2023. Learning Attention from Attention: Efficient Self-Refinement Transformer for Face Super-Resolution. In *IJCAI*. 1035–1043.

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. 3730–3738.

[8] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. 2020. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *CVPR*. 5569–5578.

[9] Stephane G Mallat. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 7 (1989), 674–693.

[10] Chenyang Wang, Junjun Jiang, Zhiwei Zhong, and Xianming Liu. 2023. Spatial-Frequency Mutual Learning for Face Super-Resolution. In *CVPR*. 22356–22366.

[11] Jingwei Xin, Nannan Wang, Xinrui Jiang, Jie Li, Xinbo Gao, and Zhifeng Li. 2020. Facial attribute capsules for noise face super resolution. In *AAAI*. 12476–12483.

[12] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*. 5728–5739.