Efficient Video Super-Resolution for Real-time Rendering with Decoupled G-buffer Guidance - Supplemental Material -

Mingjun Zheng* Long Sun* Jiangxin Dong Jinshan Pan†
School of Computer Science and Engineering, Nanjing University of Science and Technology
{mingjunzheng, cs.longsun, jxdong, jspan}@njust.edu.cn

Overview

In this document, we first present the network details in Section 1. We present additional ablation on the effect of sampling size in the DFM layer in Section 2. Then, we provide further comparisons of runtime and temporal consistency in Section 3, and present the detail processing of data generation in Section 4. To more fully illustrate the effect of HFB, we perform more PSD comparisons in Section 5. Finally, we show more visual comparison results in Section 6.

1. Network Details

As stated in Section 3 of the main paper, our method contains a dynamic feature modulator (DFM), a high-frequency feature booster (HFB), and a cross-frame temporal refiner (CTR). We also show the network details of the proposed DFM, HFB, and CTR layers in Figure 2 of the main manuscript. In this document, we show the overall network details in the Figure 1.

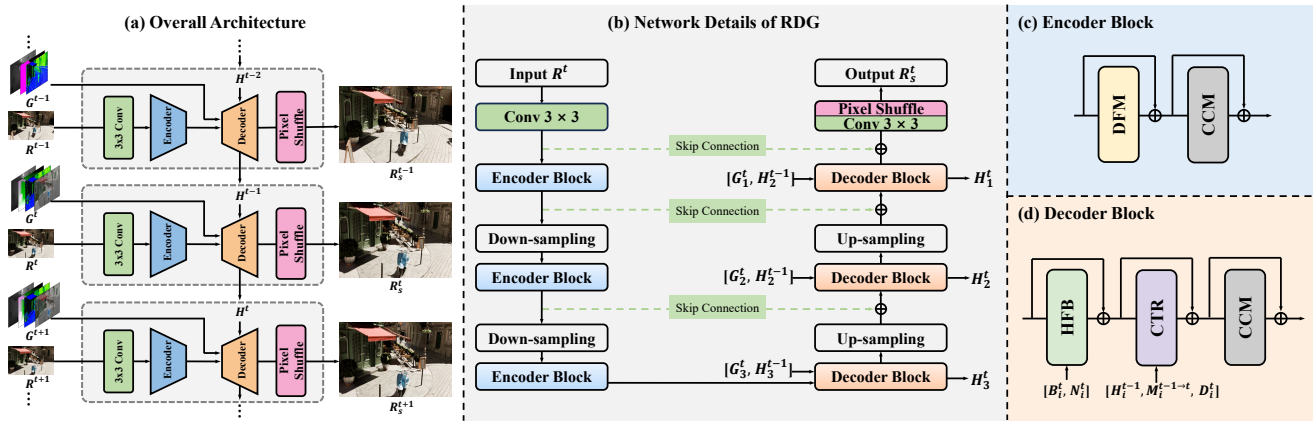


Figure 1. **Network details.** The design of RDG is a 3-level encoder-decoder architecture, where we use a combination of the proposed dynamic feature modulator (DFM) and the CCM layer [4] to encode spatial context information in the encoder part, and incorporate G-buffer components with the high-frequency feature booster (HFB) and cross-frame temporal refiner (CTR) in the decoder sub-network to guide the spatial-temporal reconstruction. We use a 3×3 convolution with stride 2 for down-sampling and a bilinear interpolation followed by a 3×3 convolution for up-sampling.

2. Effect of Sampling Size in the DFM Layer

We quantitatively evaluate the effect of the sampling size in the DFM layer. Table 1 shows that using a relative larger size is able to improve the performance. However, the improvement is not significant when the sampling size is larger than 7.

*Co-first authorship
†Corresponding author

Table 1. **Effect of the sampling size in the DFM layer.** #FLOPs, #Memory and #Latency are measured corresponding to a high-resolution image of the size 1080×1920 pixels.

	#Params	#FLOPs	#Memory	#Latency	PSNR	VMAF
RDG-s	**0.303M**	**8.50G**	**88.48M**	**7.93ms**	**29.37**	**87.22**
Sampling size $7 \rightarrow 5$	0.303M	8.50G	88.48M	7.92ms	29.35	86.75
Sampling size $7 \rightarrow 9$	0.303M	8.50G	88.48M	7.96ms	29.23	87.43
Sampling size $7 \rightarrow 11$	0.304M	8.50G	88.48M	8.14ms	29.25	87.13
Sampling size $7 \rightarrow 15$	0.304M	8.50G	88.48M	8.20ms	29.34	87.08

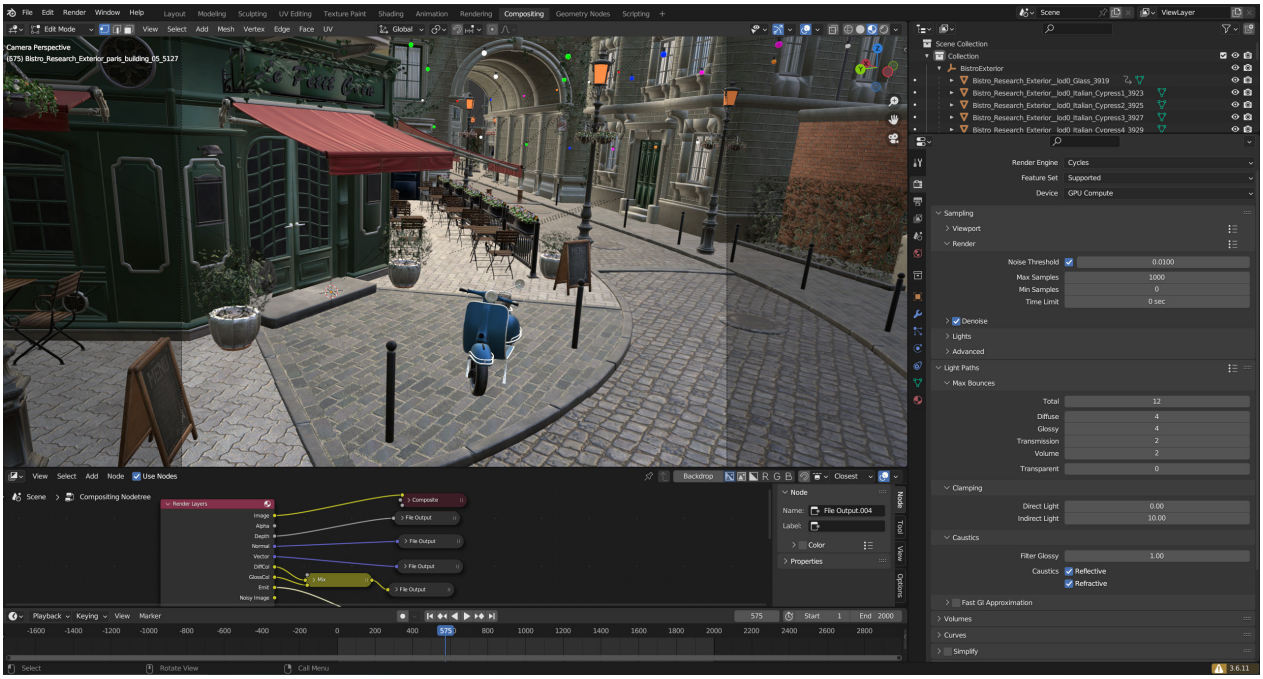


Figure 2. **Blender rendering settings.** We use the Cycles engine to render the scene model and generate the dataset, where we set the maximum number of light samples to 1000, the maximum number of light reflections to 12, the resolution of the HR video is 1920×1080, the resolution of the LR video is set to 480×270, and the frame rate of the camera is set to 24 FPS.

Table 2. Runtime on different devices.

Methods	RTX 3090	RTX 2080 Ti	GTX 1060
FuseSR [6]	33.99ms	74.97ms	154.39ms
RDG	18.78ms	29.85ms	91.24ms
RDG-s	7.93ms	11.68ms	36.95ms

Table 3. Temporal consistency metrics.

Metrics	SAFMN [4]	BasicVSR++ [2]	NSRD [3]	RDG
VMAF ↑	81.62	84.39	67.47	88.96
DOVER ↑	81.47	82.13	81.86	83.25
Flickering ↓	13.28	13.17	13.74	13.06

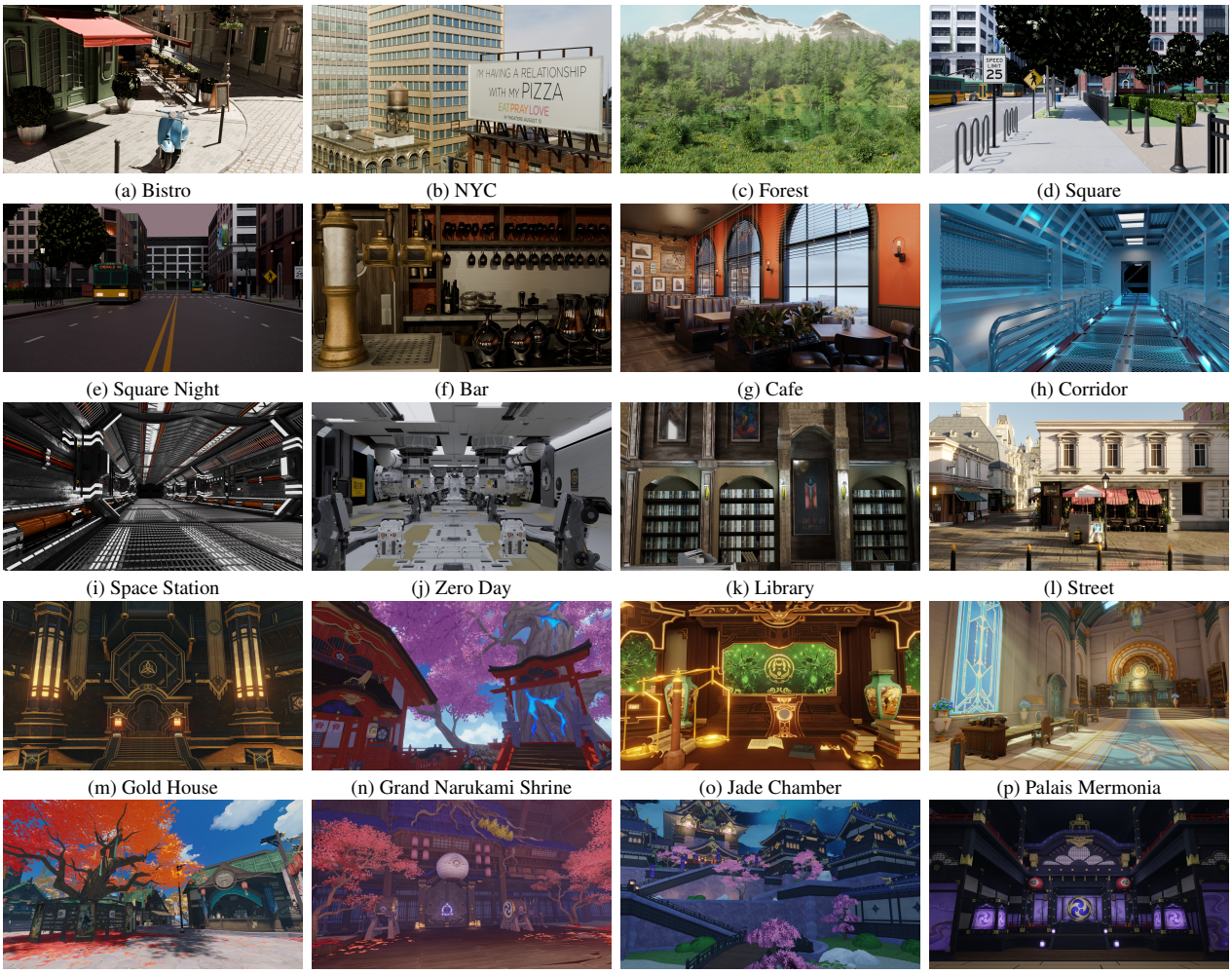
3. Runtime and Temporal Consistency Comparisons.

We further assess the inference time on different devices. Table 2 demonstrates that our RDG achieves real-time 1080P rendering on mid- to high-end GPUs (e.g., RTX 2080 Ti and RTX 3090), and the RDG-s is close to real-time on low-tier device (e.g., GTX 1060).

We compare with several representative methods on all test datasets in terms of temporal stability. Table 3 shows that our RDG achieves a better temporal consistency.

4. Dataset Generation

We use the Cycles [1] engine to render the scene model and generate the dataset. Figure 2 illustrates the detailed rendering settings, where we set the maximum number of light samples to 1000, the maximum number of light reflections to 12, HR



(a) Bistro (b) NYC (c) Forest (d) Square

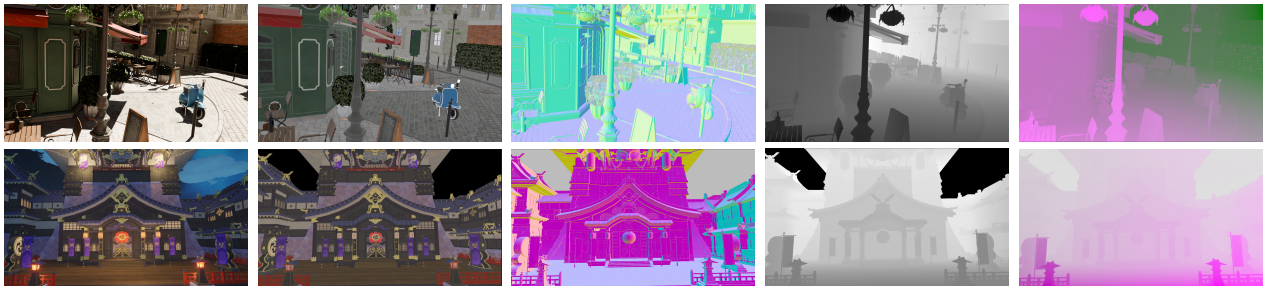
(e) Square Night (f) Bar (g) Cafe (h) Corridor

(i) Space Station (j) Zero Day (k) Library (l) Street

(m) Gold House (n) Grand Narukami Shrine (o) Jade Chamber (p) Palais Mermonia

(q) Ritou (r) Shakkei Pavilion (s) Tenshukaku (t) Tenshukaku Interior

Figure 3. Example scenes. The collected dataset covers different scenarios such as complex textures and geometries, glossy reflections, and fast-moving objects.



(a) Image (b) BRDF (c) Normal (d) Depth (e) Motion Vector

Figure 4. Visualization of frame examples and their corresponding G-buffers. Each frame contains BRDF, normal, depth and motion vector.

videos have a spatial resolution of 1920×1080 and 480×270 for their LR counterparts, and the frame rate of the camera is set to 24 FPS. The collected dataset covers different scenarios such as complex textures and geometries, glossy reflections, and fast-moving objects, as shown in Figure 3. We also generate and save the G-buffer components (*i.e.*, normal, depth, BRDF and motion vector) for each frame, as shown in Figure 4.

5. More PSD Visualization Comparisons

In this section, we provide more power spectral density (PSD) visual comparisons between the proposed HFB layer and the SFT layer [5]. Figure 5 shows our proposed HFB enables moreeffectively transferring high-frequency information from the G-buffer components, resulting in richer details for the reconstruction results.

6. More Experimental Results

In this section, we provide more visual comparisons of the proposed method and state-of-the-art ones on test datasets. Figure 6 shows the comparisons, where our method generates better super-resolved frames.

References

[1] Blender. https://www.blender.org/, 2023. 2

[2] Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 2022. 2

[3] Jia Li, Ziling Chen, Xiaolong Wu, Lu Wang, Beibei Wang, and Lei Zhang. Neural super-resolution for real-time rendering with radiance demodulation. In *CVPR*, 2024. 2, 6

[4] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *ICCV*, 2023. 1, 2

[5] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 4, 5

[6] Zhihua Zhong, Jingsen Zhu, Yuxin Dai, Chuankun Zheng, Guanlin Chen, Yuchi Huo, Hujun Bao, and Rui Wang. Fusesr: Super resolution for real-time rendering through efficient multi-resolution fusion. In *SIGGRAPH Asia*, 2023. 2, 6

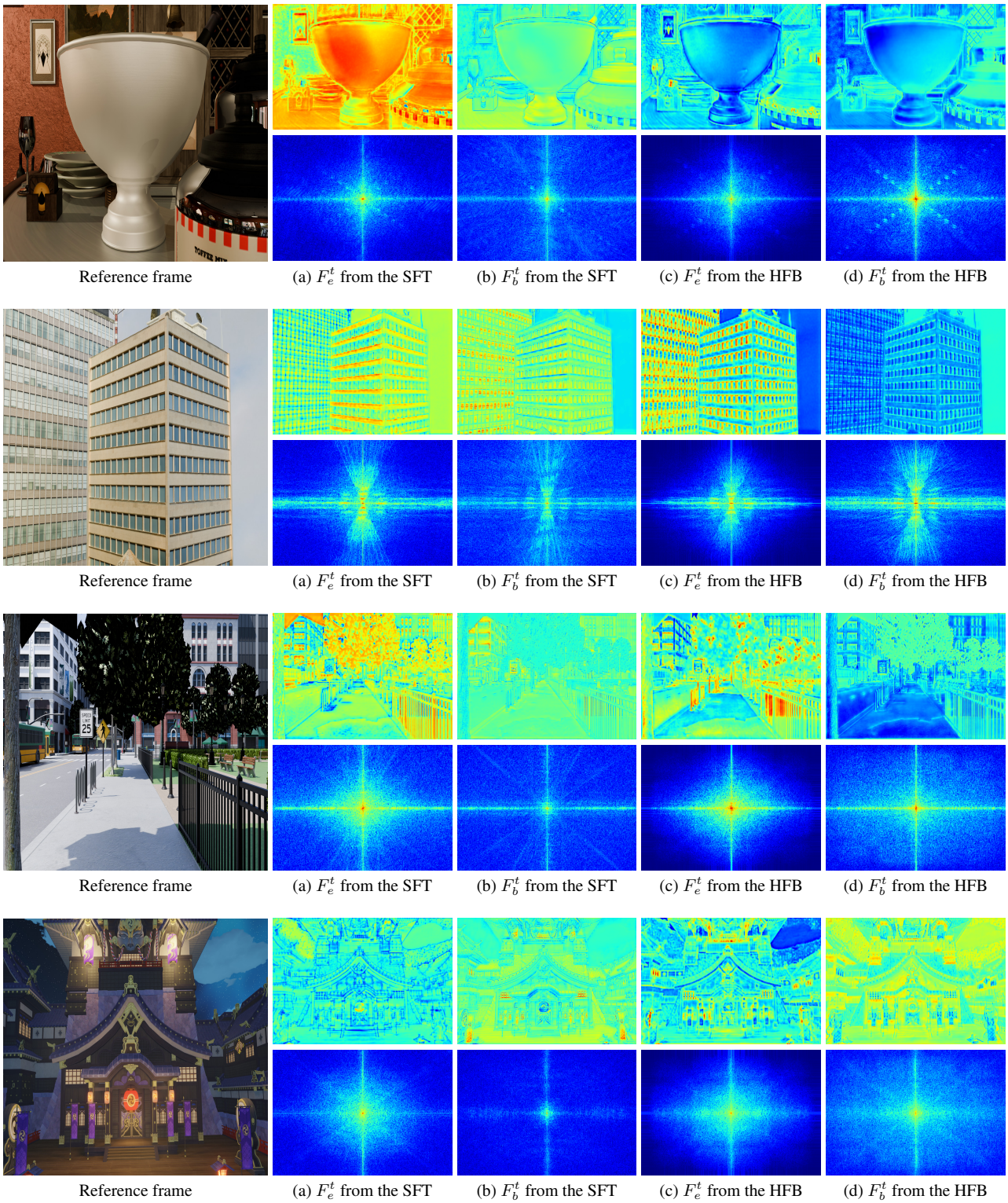
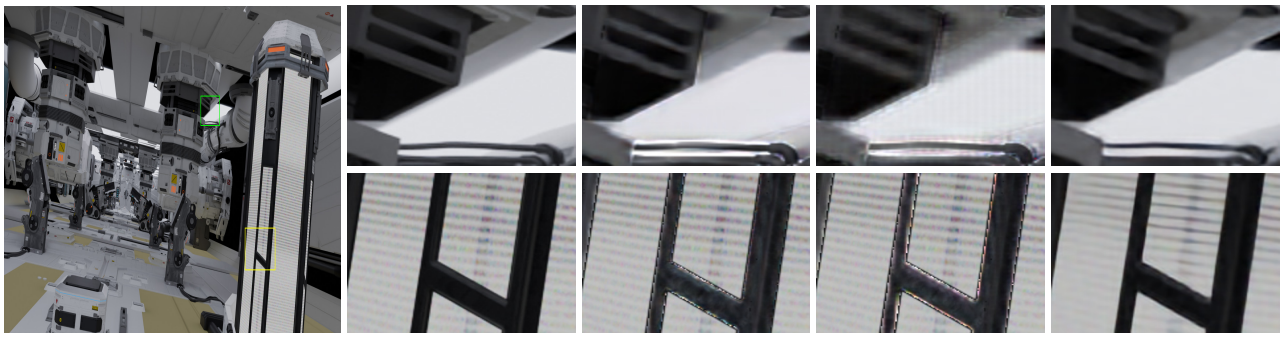


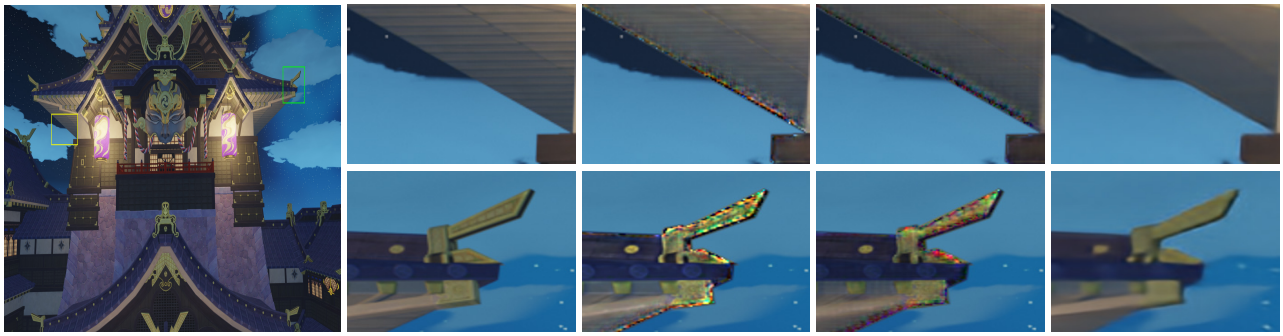
Figure 5. Power spectral density (PSD) comparisons between the SFT layer [5] and the proposed HFB layer. The first row is the visualization of input F_e^t and output F_b^t and the second row is their corresponding power spectral maps. Our proposed HFB enables more effectively transferring high-frequency information from the G-buffer components, resulting in richer details for the reconstruction results.

Frame 062 from Shrink (a) GT patches (b) NSRD [3] (c) FuseSR [6] (d) Ours



Frame 140 from ZeroDay (a) GT patches (b) NSRD [3] (c) FuseSR [6] (d) Ours

Frame 059 from Squre-night (a) GT patches (b) NSRD [3] (c) FuseSR [6] (d) Ours



Frame 121 from Tenshu (a) GT patches (b) NSRD [3] (c) FuseSR [6] (d) Ours

Figure 6. Visual comparisons with recent real-time rendering SR methods on test datasets. The results shown in (b) and (c) contain significant re-modulation artifacts. In contrast, our method generates a better super-resolved frame.