Real-RawVSR: Real-World Raw Video Super-Resolution with a Benchmark Dataset –Supplementary Material–

Huanjing Yue, Zhiming Zhang, and Jingyu Yang*

School of Electrical and Information Engineering, Tianjin University, Tianjin, China
{huanjing.yue,zmzhang,yjy}@tju.edu.cn

This supplementary file provides the details that were not presented in the main paper due to page limitations. First, we give more details about our constructed Real-RawVSR dataset. Second, we give the network details for the temporal fusion module. Third, we provide the ablation study results for 2× SR. Fourth, to verify the effectiveness of our proposed network tailored for raw sequence input, we retrain our network on the synthetic RawVSR dataset, *i.e.,* RawVD [2], and give comparison results with state-of-the-art methods. Hereafter, to verify the generalization of the model trained on our dataset, we provide test results on new frames captured by other devices. Finally, we provide the SR results for large scenes.

1 Real-RawVSR Dataset

We totally captured 450 LR-HR video pairs for 3 magnification ratios. The HR scenes are captured with focal length setting to 72mm and the LR scenes for the three magnification ratios are captured with focal length setting to 36mm, 24mm, and 18mm respectively. Table 1 lists the detailed information in terms of resolutions and scene numbers for our Real-RawVSR dataset. Fig. 1 presents some examples of our captured scenes, including sports, animals, playgrounds, cityscape, indoors, streets, etc.

We would like to point out that there may be small misalignments between the moving area in LR and HR frames when the captured object is moving fast. Even though we utilize an infrared remote control to signal both cameras to capture at the same time, it is difficult to keep synchronization for capturing rapid movements. Fortunately, it only happens in a few scenes.

2 Network Structure

Temporal Fusion. After the interaction module, we get the combined aligned frames $\hat{F}_c^b$ and $\hat{F}_c^s$, whose temporal length is $4N+2$. Since our temporal fusion module is the same for the two branches, in the following, we take the Bayer

* This work was supported in part by the National Natural Science Foundation of China under Grant 62072331. *Corresponding author: Jingyu Yang.*

Table 1. Detailed settings for our Real-RawVSR dataset.

Scale	Pair	Focal Length	Resolution	Frames per Video	Number of Videos		
					Indoor	Outdoor	Total
2×	LR	36mm	320 × 720	120 – 160	19	131	150
	HR	72mm	640 × 1440				
3×	LR	24mm	224 × 480	120 – 160	19	131	150
	HR	72mm	672 × 1440				
4×	LR	18mm	160 × 360	120 – 160	19	131	150
	HR	72mm	640 × 1440				

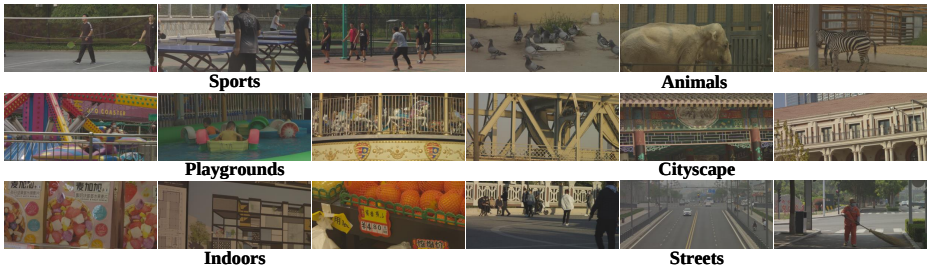

Fig. 1. Examples of our captured scenes.

pattern branch as an example. As shown in Fig. 2, we first utilize a temporal non-local attention, similar to that in [5], to aggregate long-range temporal features of $\hat{F}_c^b$ to enhance the feature representations along the time dimension. Then we utilize temporal-spatial attention (TSA) [4] based fusion method to fuse the $4N + 2$ features along the temporal dimension and $\hat{F}_t^b$ is set as the reference frame.

3 Ablation Study for 2× SR

We further present the ablation experiments for 2× SR to verify the effectiveness of our proposed modules. For 2× SR, our key module achieves larger gains compared with those in 4× SR (see Table 2). For example, 0.55, 0.3, and 0.2 dB gain are achieved by two-branch structure, interaction, and co-alignment modules, respectively.

4 Results on Synthetic RawVSR Dataset

Besides our constructed Real-RawVSR dataset, we further evaluate our proposed network on synthetic RawVSR dataset, *i.e.,* RawVD [2]. The LR raw sequences

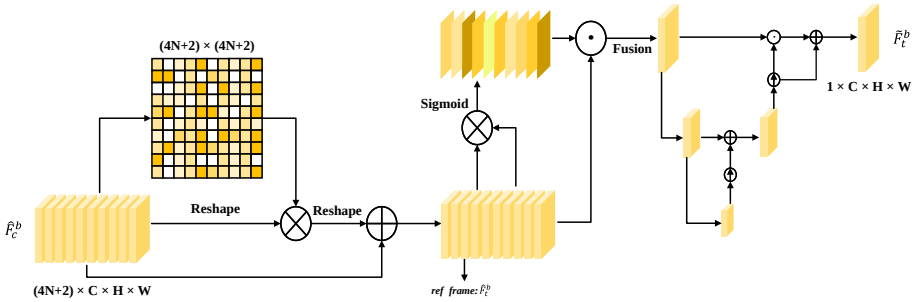

Fig. 2. The temporal fusion module for the Bayer pattern branch, where $N = 2$.

Table 2. Ablation study (2×) for the key modules in our network.

Alignment	Sep-alignment	✗	✓	✗
	Co-alignment	✗	✗	✓
	PSNR/SSIM	37.06/0.9690	37.18/0.9702	**37.38/0.9705**
Interaction	Interaction	✗	✓	
	PSNR/SSIM	37.08/0.9694	**37.38/0.9705**	
Channel Fusion	SKF	✗	✓	
	Concat	✓	✗	
	PSNR/SSIM	37.12/0.9691	**37.38/0.9705**	
Color Correction	Matrix-based	✗	✓	✗
	Channel-based	✗	✗	✓
	PSNR/SSIM	33.56/0.9627	37.23/0.9682	**37.38/0.9705**
Branch	Bayer Branch	✓	✗	✓
	Sub-frame Branch	✗	✓	✓
	PSNR/SSIM	36.65/0.9670	36.83/0.9677	**37.38/0.9705**

in this dataset are obtained by blurring, downsampling and Bayer extraction of the HR demosaiced linear raw sequences. Noise is also introduced to simulate the noise embedded in the LR raw sequences. However, there is no processing to simulate the color and brightness differences between the LR and HR sequences, which usually exist in real LR-HR pairs. Considering this fact, we did not utilize our color correction module designed for real LR-HR pairs and choose to utilize the color correction branch proposed in [2], which is suitable for synthetic pairs. Besides retraining our network, we also retrain RawEDVR for reference. Similarly, the color correction branch for synthetic pairs is also introduced to RawEDVR. As shown in Table 3, compared with the sRGB domain methods, *e.g.* EDVR, our method achieves more than 1.4 dB gain. This demonstrates that raw domain processing is beneficial for video SR. In addition, our method outperforms RawVSR and RawEDVR by 0.5 dB and 0.42 dB for 4× video SR. This demonstrates that our proposed co-alignment and interaction strategy is supe-

Table 3. Quantitative comparison on RawVD [2] for 4× video SR. The best results are highlighted in bold and the second best results are underlined.

	Bicubic	TDAN [3]	EDVR [4]	RawVSR [2]	RawEDVR	Ours
PSNR	23.78	26.85	27.81	28.76	<u>28.84</u>	**29.26**
SSIM	0.6389	0.7422	0.7734	0.7939	<u>0.8001</u>	**0.8043**

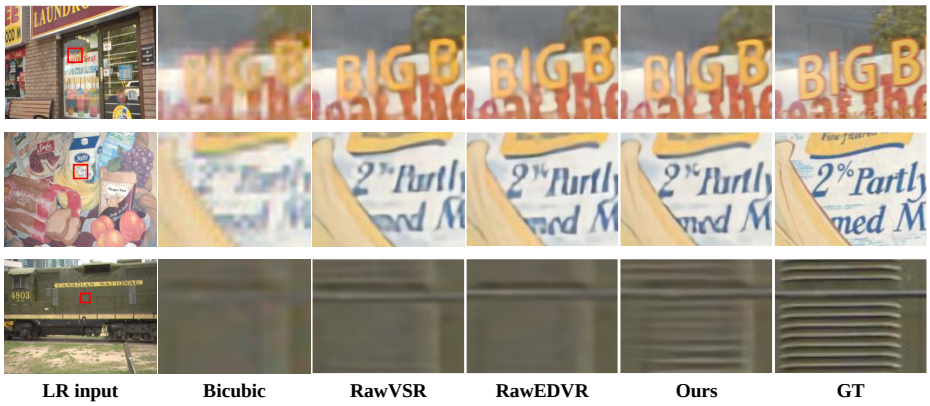

Fig. 3. Visual comparisons on RawVD for 4× video SR.

rior to other modules without specific processing for raw sequence inputs. The visualization comparison results are shown in Fig. 3. It can be observed that our method generates much sharper details compared with state-of-the-arts. Specifically, only our method can recover the dense edges on the wall (see the bottom row). In a word, our network is a strong benchmark network for RawVSR.

5 Generalization

We evaluate the generalization ability of our pre-trained model on the Real-RawVSR dataset by directly testing on LR sequences captured by other different devices. Since most devices cannot capture raw videos directly, we utilize three mobile phones, *i.e.,* Huawei P20, Xiaomi 10 Ultra, and Redmi K50, to capture burst raw images to simulate video sequences. Since there are large differences between the ISP modules on different devices, the testing result is corrected by the color correction coefficients calculated between the output and the LR sRGB image to make the output have similar color tones as that of the LR sRGB image. The results are shown in Fig. 4, Fig. 5 and Fig. 6, respectively. The bicubic results, which are generated by directly upsampling the LR sRGB frame by bicubic interpolation, are given for reference. It can be observed that although the SR model is trained on Canon LR-HR pairs, it can also generate

details for the frames captured by different devices. Note that, our SR results may have a color cast due to the gap between the raw inputs captured by different devices. In this work, we focus on detail generation, which is the main task of SR. Therefore, the readers are encouraged to pay more attention to the sharp edges generated by our method.

Note that, the Bayer patterns of the three phones are BGGR, GRBG, and GBRG, respectively. The Bayer pattern for our captured Canon LR-HR pairs are RGGB. This demonstrates that our network can be extended to different Bayer patterns. For other CFA patterns, such as X-Trans patterns, the raw frame can be packed in terms of 6×6 blocks, similar to that in [1].

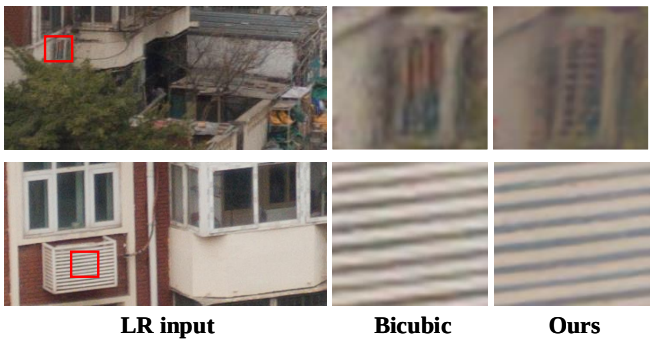

Fig. 4. SR (4×) results on frames captured by Huawei P20.

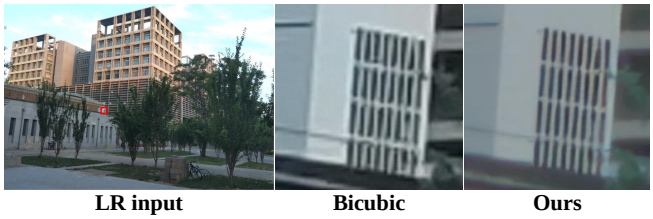

Fig. 5. SR (4×) results on frames captured by Xiaomi 10 Ultra.

Fig. 6. SR (4×) results on frames captured by Redmi K50.

6 SR Results on Wide View Videos

We also test our model on scenes with large view angles, and the resolution of the LR input is 800×1600. Since there are no ground truths for these scenes, we only give the visual results. As shown in Fig. 7 and Fig. 8, our SR results contain rich details, and the noise embedded in the LR input is also reduced. Therefore, our model is an effective tool to get both HR and wide-angle videos.

Bicubic

Ours

Fig. 7. SR (4×) results on a video frame with a large view angle.

Fig. 8. SR (4×) results on a video frame with a large view angle.

References

1. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3291–3300 (2018)
2. Liu, X., Shi, K., Wang, Z., Chen, J.: Exploit camera raw data for video super-resolution via hidden markov model inference. IEEE Transactions on Image Processing **30**, 2127–2140 (2021)
3. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3360–3369 (2020)
4. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
5. Yue, H., Cao, C., Liao, L., Chu, R., Yang, J.: Supervised raw video denoising with a benchmark dataset on dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2301–2310 (2020)