Benchmark Dataset and Effective Inter-Frame Alignment for Real-World Video Super-Resolution Supplementary Material

Ruohao Wang[1], Xiaohui Liu[1], Zhilu Zhang[1], Xiaohe Wu[1](✉), Chun-Mei Feng[2], Lei Zhang[3], Wangmeng Zuo[1]
[1]Harbin Institute of Technology, China
[2]Institute of High Performance Computing, A*STAR, Singapore
[3]The Hong Kong Polytechnic University, China
{rhwangHIT, cszlzhang}@outlook.com, {lxh720199, xhwu.cpsl.hit, strawberry.feng0304}@gmail.com,
cslzhang@comp.polyu.edu.hk, wmzuo@hit.edu.cn

Table A. Structure configuration of the encoder. The kernel size of 'Conv' layers is 3×3.

#	Layer name(s)
0	Conv $(3, 64)$, ReLU
1	Conv $(64, 64)$, ReLU
2	Conv $(64, 128)$, ReLU
3	Conv $(128, 128)$, ReLU
4	Conv $(128, 256)$, ReLU

A. Content

The content of this supplementary material involves:

- Network structures of encoder, reconstructor and upsampler in Sec. B.
- Effect of pre-trained optical flow network in Sec. C.
- Effect of training loss in Sec. D.
- More visual comparisons in Sec. E.

B. Network Structures of Encoder, Reconstructor and Upsampler

To estimate the effectiveness of the proposed method comprehensively, we select bidirectional [1] and second-order grid [2] as our temporal propagation scheme, named EAVSR and EAVSR+, respectively. Table A shows the detailed architectures of the encoder in EAVSR and EAVSR+. Table B shows the detailed architectures of the reconstructor and upsampler in EAVSR. 'Propagated Feature' in Table B refers to the well-aligned features propagated from neighboring frames, and RCAB denotes the residual channel attention block [12]. The reconstructor and upsampler architectures of EAVSR+ follow BasicVSR++ [2].

Table B. Structure configuration of the combination of reconstructor and upsampler. The kernel size of 'Conv' layers is 3×3 and the kernel size of 'Conv1×1' layer is 1×1. The negative slope of LeakyReLU is 0.1.

#	Layer name(s)
0	Concat [LR, Propagated Feature]
1	Conv $(67, 64)$, LeakyReLU
2	RCAB × 30
3	Conv 1×1 $(64, 64)$, LeakyReLU
4	Conv $(64, 256)$, PixelShuffle, LeakyReLU
5	Conv $(64, 64)$, LeakyReLU
6	Conv $(64, 3)$
7	BilinearUpsample $(\mathrm{LR}, 2)$
8	ElementwiseAdd $(\#6, \#7)$

C. Effect of Pre-Trained Optical Flow Network

As described in Sec. 4.2 of the main text, we follow BasicVSR [1] and BasicVSR++ [2], using a light-weight pre-trained optical flow network SPyNet [7] to calculate the basic offset between neighboring frames. In this section, we replace SPyNet with PWC-Net [8] to verify the effect of the proposed ResflowNet and DeformNet when taking a better and more robust pre-trained optical flow network.

Table C shows the experiment results on RealVSR dataset [10]. The first and the fourth rows demonstrate that despite deploying a better optical flow network, the

Figure A. Visual comparison between different data post-processing strategies during training. The first ✓ or × means using color correction or not, and the second ✓ or × means using spatial position alignment or not. Please zoom in for details.

Figure B. Visual comparison on RealVSR dataset [10]. Our methods EAVSR+ and EAVSR can better recover the window contours, especially the window in the upper left corner. Please zoom in for details.

Table C. Quantitative comparison with the different pre-trained optical flow networks on RealVSR [10] dataset.

Optical Flow Network	ResflowNet	DeformNet	PSNR	SSIM
SPyNet [7]	×	×	24.13	0.7854
SPyNet [7]	✓	×	24.24	0.7902
SPyNet [7]	✓	✓	24.41	0.7953
PWC-Net [8]	×	×	24.19	0.7847
PWC-Net [8]	✓	×	24.26	0.7906
PWC-Net [8]	✓	✓	24.44	0.7939

final performance still has a limited improvement on real-world data. Taking our ResflowNet to compensate for the error caused by the pre-trained flow network, the second and fifth rows show that the performance promotes significantly. Moreover, the third and the last lines indicate the validity of our DeformNet as well.

D. Effect of Training Loss

In this section, we show some visual results in Fig A when using different data post-processing strategies during training, which is mentioned in Sec. 6.2 of the main text. When neither color correction nor spatial position alignment is applied, the results cannot restore the correct details. When we use the PWC-Net [8] to mitigate the spatial misalignment between LR and HR, the result can generate more details. Moreover, if we only apply the guided image filtering [4] to correct the color of HR, the result can keep the brightness consistent with LR but lead to blurry. Utilizing both color correction and spatial position alignment can recover more textures and remain color consistent with the LR image.

E. Visual Comparison

In this section, we provide more qualitative comparisons between our methods and other state-of-the-art algorithms on RealVSR [10] dataset (see Fig. B, Fig. C and Fig. F), and MVSR4× dataset (see Fig. D, Fig. E and Fig. G), respectively. Specifically, we show results of methods trained only with ℓ_1 loss in Figs B∼E, and methods trained with additional adversarial loss in Figs F∼G, respectively.

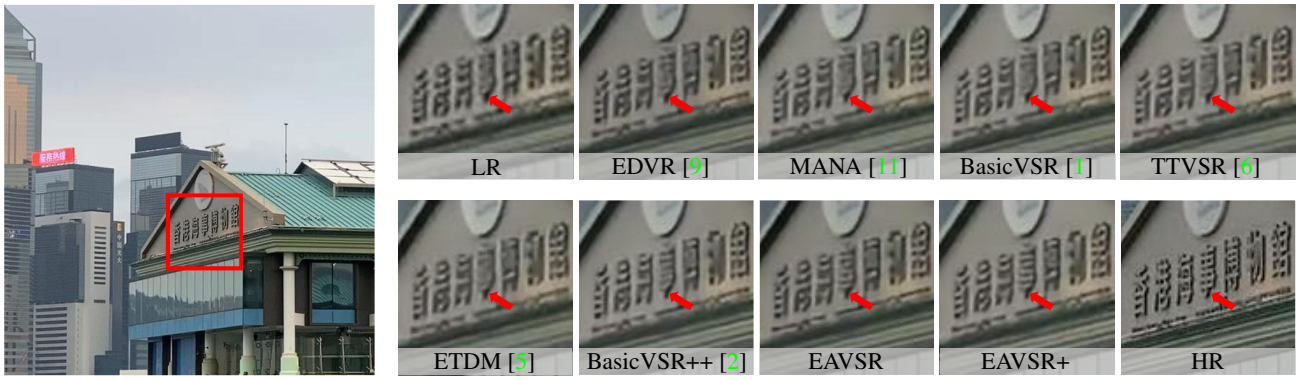

Figure C. Visual comparison on RealVSR dataset [10]. Our methods EAVSR+ and EAVSR can recover the characters better. Please zoom in for details.

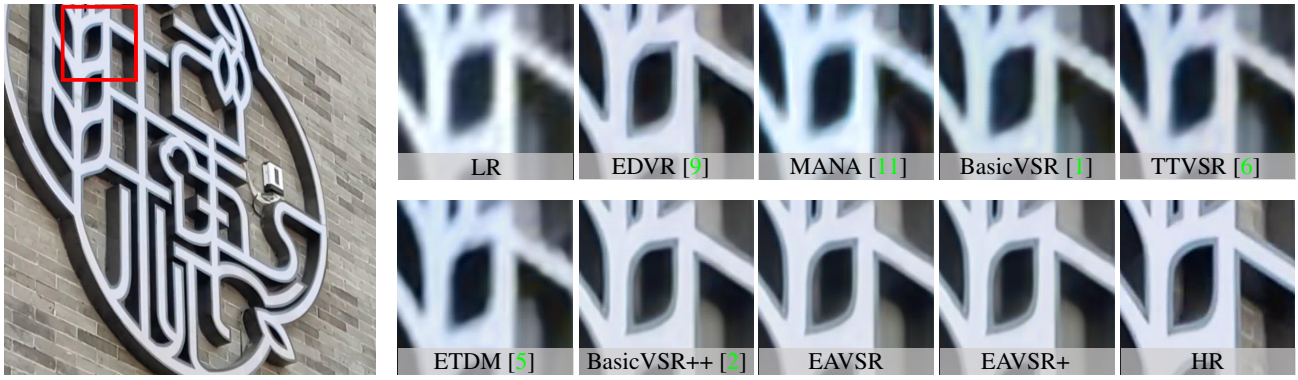

Figure D. Visual comparison on MVSR4× dataset. Our methods EAVSR+ and EAVSR can generate clearer contours with fewer artifacts. Please zoom in for details.

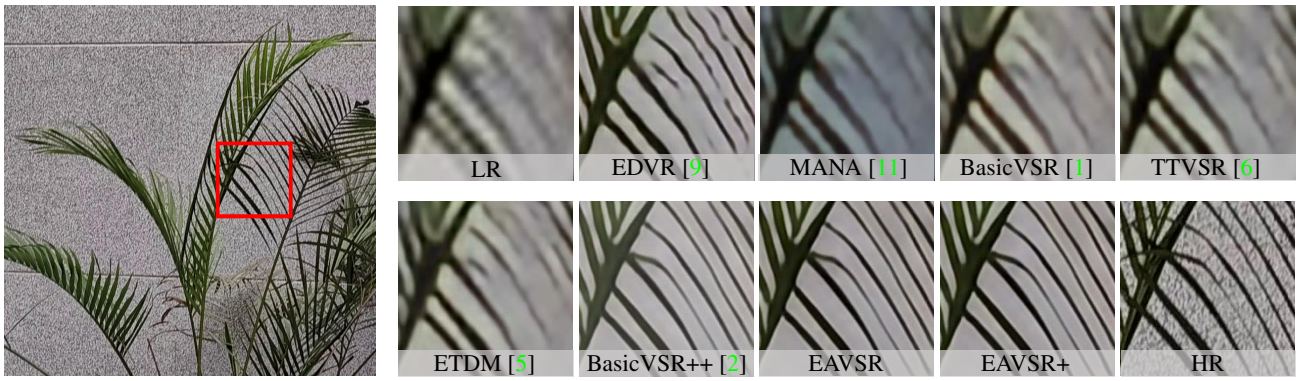

Figure E. Visual comparison on MVSR4× dataset. Our methods EAVSR and EAVSR+ can restore sharper branches. Please zoom in for details.

Figure F. Visual comparison on RealVSR [10] dataset between methods trained with adversarial loss. Our results from EAVSRGAN+ have more details and are more photo-realistic. Please zoom in for more details.

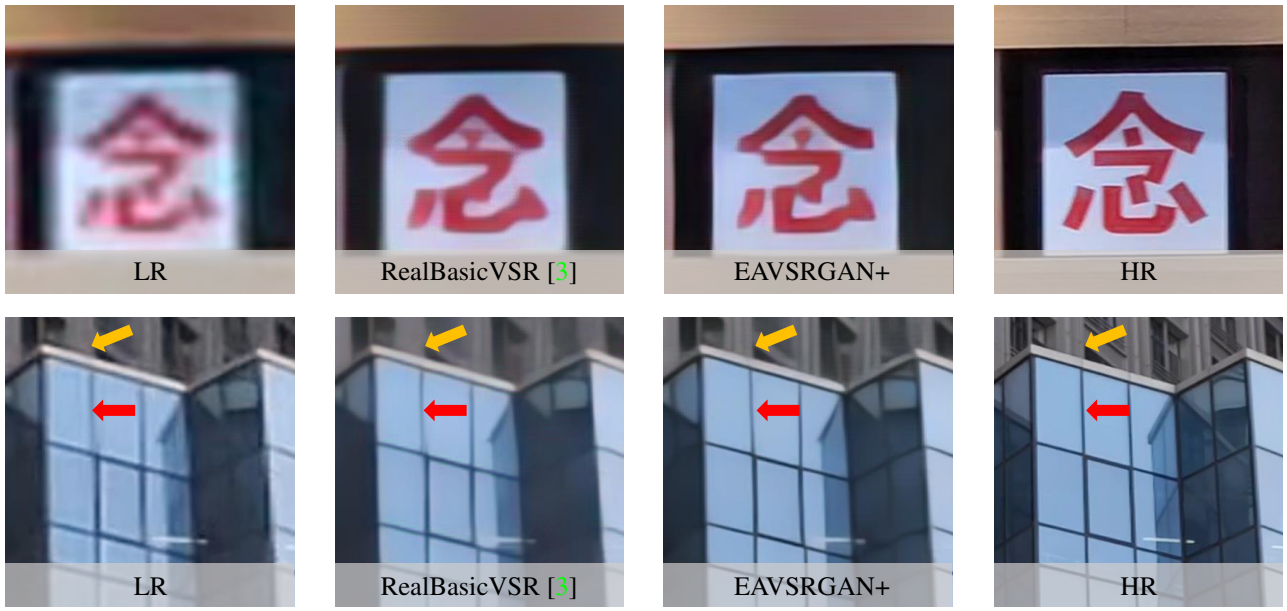

Figure G. Visual comparison on MVSR4× dataset between methods trained with adversarial loss. Our results from EAVSRGAN+ have clearer edges and are more photo-realistic. Please zoom in for more details.

References

[1] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 1, 2, 3

[2] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *arXiv preprint arXiv:2104.13371*, 2021. 1, 2, 3

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 4

[4] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012. 2

[5] Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, and Yu-Wing Tai. Look back and forth: Video super-resolution with explicit temporal difference modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17411–17420, June 2022. 2, 3

[6] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5687–5696, June 2022. 2, 3

[7] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 1, 2

[8] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 1, 2

[9] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3

[10] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4781–4790, 2021. 1, 2, 3, 4

[11] Jiyang Yu, Jingen Liu, Liefeng Bo, and Tao Mei. Memory-augmented non-local attention for video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022. 2, 3

[12] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 1