Event-based Video Super-Resolution via State Space Models
—— Supplementary Material ——

Zeyu Xiao Xinchao Wang✉
National University of Singapore

Overview

This supplementary document is organized as follows:
Section 1 provides more qualitative and quantitative comparisons.
Section 2 provides more detailed ablation studies.
Section 3 offers a discussion of the proposed MamEVSR.

1. Qualitative Comparisons

Figure 1 presents visual comparisons for $\times 4$ VSR on the CED dataset. MamEVSR consistently provides sharper and clearer images compared to other methods, preserving fine details and textures. While other baseline methods improve upon the LR input, they struggle to match the level of detail and clarity achieved by MamEVSR.

Table 1 list the LPIPS results and the computational costs of different methods.

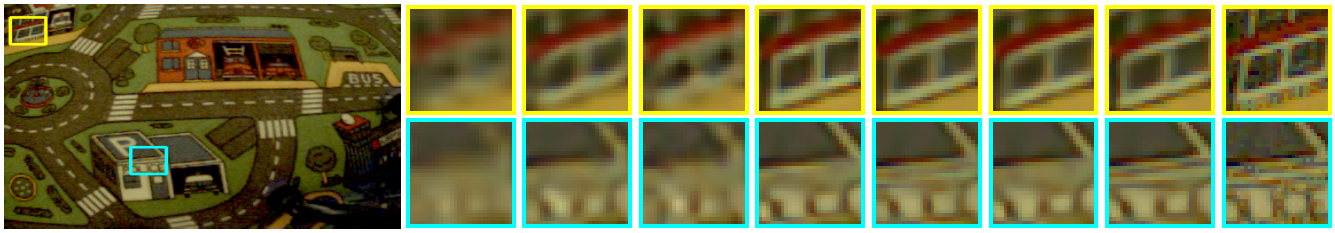

Figure 1. Visual comparisons for $\times 4$ VSR on CED. From left to right in sequence are patches cropped from LR, EDVR, BasicVSR, EBVSR, EGVSR, EvTexture, MamEVSR, and the ground truth image.

Table 1. Quantitative comparison in terms of LPIPS and computational costs for the $\times 4$ event-based VSR task on the REDS4 dataset.

Method	EDVR	BasicVSR	IconVSR	BasicVSR++	VRT
LPIPS	0.2091	0.2018	0.1946	0.1786	0.1864
Time (ms)	378	63	70	77	243
FPS (1/s)	2.6	15.9	14.3	13.0	4.1
Method	EGVSR	EBVSR	EvTexture	EvTexture+	MamEVSR
LPIPS	0.3024	0.1996	0.1684	0.1642	0.1639
Time (ms)	193	92	136	139	127
FPS (1/s)	5.2	10.9	7.4	7.2	7.9

Table 2. Ablation study of different components on CED.

Method		CED $\times 2$		CED $\times 4$	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑
Ours	All residual blocks	39.99	0.9798	33.53	0.9082
	Full model	41.14	0.9831	34.03	0.9189
Core modules	(a) w/o iMamba	40.17	0.9803	33.60	0.9109
	(b) w/o cMamba	40.26	0.9805	33.63	0.9110
	(c) w/o backward cell	40.07	0.9789	33.48	0.9076
	(d) w/o forward cell	40.08	0.9789	33.40	0.9065
iMamba	(e) Residual blocks	40.62	0.9816	33.68	0.9111
	(f) Deform. conv.	40.72	0.9819	33.77	0.9127
	(g) Flow warping	40.84	0.9821	33.80	0.9129
	(h) Concat & attention	40.80	0.9820	33.77	0.9126
	(i) w/o Int. Reorg.	40.99	0.9826	33.88	0.9140
	(j) w/o Channal att.	41.02	0.9829	33.94	0.9145
cMamba	(k) Residual blocks	40.69	0.9818	33.71	0.9114
	(l) EBVSR-BCS module [1]	40.77	0.9819	33.80	0.9139
	(m) Cross-modal att. [2]	40.98	0.9826	33.88	0.9150
	(n) Concat & iMamba	40.80	0.9820	33.77	0.9129
	(o) w/o Cross SSM	40.87	0.9825	33.90	0.9170
Recon.	(p) w/o iMamba	41.04	0.9829	34.00	0.9179
	(q) All iMamba	40.89	0.9826	33.94	0.9168

2. Ablation Study

In this section, we conduct experiments on the CED dataset to demonstrate the effectiveness of the proposed MamEVSR. Due to space limitations, we cannot provide a detailed description of the experimental setup and analysis of the ablation studies in the main text. Here, we present a comprehensive discussion of these aspects. Results are shown in Table 2. Notably, when modifying or removing modules, we replace them with residual blocks to ensure consistent parameter counts for fair comparisons.

2.1. Effectiveness of the Core Components in MamEVSR

We introduce the following variants to demonstrate the effectiveness of the proposed core components. (1) MamEVSR - All residual blocks: we replace all modules with residual blocks. This configuration serves as a baseline for evaluating the effectiveness of our proposed MamEVSR. (2) MamEVSR w/o iMamba: we remove the iMamba blocks from the MamEVSR for this variant. (3) MamEVSR w/o cMamba: we remove the cMamba blocks from the MamEVSR for this variant. (4) MamEVSR w/o backward cell: we remove the backward cells from the MamEVSR for this variant. (5) MamEVSR w/o forward cell: we remove the forward cells from the MamEVSR for this variant. Our full model demonstrates superior performance, achieving a PSNR of 41.14 dB and an SSIM of 0.9831 for CED $\times 2$, and a PSNR of 34.03 dB and an SSIM of 0.9189 for CED $\times 4$. Ablation studies reveal that each core module significantly impacts the final outcome; for instance, omitting the iMamba or cMamba components notably reduces the PSNR and SSIM scores. Specifically, excluding the iMamba module (method (a)) results in a PSNR reduction of about 0.97 dB and an SSIM decrease of 0.005 for CED $\times 2$ relative to the full model. Similarly, removing the cMamba module (method (b)) causes a PSNR drop of around 0.88 dB and an SSIM decrement of 0.0026 for CED $\times 2$. Notably, the absence of the forward cell (method (d)) exhibits the least impact on performance, with minimal reductions in PSNR and SSIM for both scale factors. Conversely, eliminating the backward cell (method (c)) leads to a more substantial decline, especially for CED $\times 4$, where the PSNR falls by 0.59 dB and the SSIM by 0.011 compared to the full model. These findings underscore the critical role of each component in attaining high-quality super-resolved images.

2.2. A Close Look At the iMamba Block

The iMamba block is designed for efficient feature fusion and propagation across bi-directional frames. To analyze its effectiveness, we evaluate several variants. (1) iMamba - Residual blocks: we replace the iMamba block with residual blocks

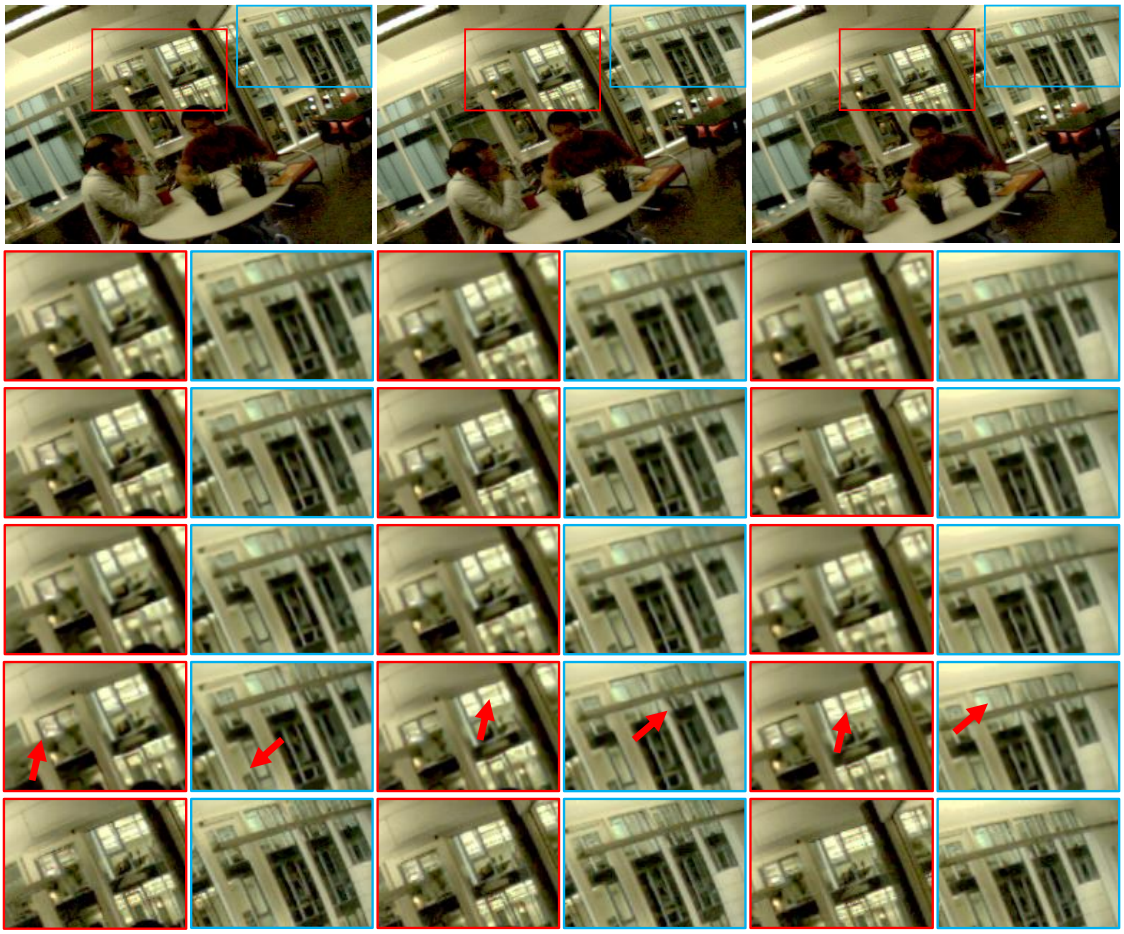

Figure 2. Visual results of different variants of the iMamba block. From top to bottom are ground truth image, and patches cropped from iMamba - Deform. conv, iMamba - Flow warping, iMamba w/o Int. Reorg., iMamba, and the ground truth image.

for frame alignment and fusion. (2) iMamba - Deform. conv.: we utilize deformable convolution for frame alignment, similar to EDVR [3]. (3) iMamba - Flow warping: we employ optical flow estimation combined with warping for frame alignment. (4) iMamba - Concat & attention: we concatenate features from two frames along the feature dimension, followed by attention-based fusion [3]. (5) iMamba - w/o Int. Reorg.: we remove the interleaved reorganization in the iMamba block, replacing it with a naive scanning approach. (6) iMamba - w/o Channel att.: we remove channel attention from the iMamba block, reducing its capacity for enhanced global perception. Our full model achieves a baseline performance of 41.14 dB PSNR and 0.9831 SSIM for CED $\times 2$, and 34.03 dB PSNR and 0.9189 SSIM for CED $\times 4$. Ablation study (e) replaces the iMamba block with residual blocks, resulting in 40.62 dB PSNR and 0.9816 SSIM ($\times 2$), and 33.68 dB PSNR and 0.9111 SSIM ($\times 4$), showing that residual blocks contribute positively but are less impactful overall. Study (f) excludes deformable convolutions, yielding 40.72 dB PSNR and 0.9819 SSIM ($\times 2$), and 33.77 dB PSNR and 0.9127 SSIM ($\times 4$), indicating their importance but limited effectiveness for the CED dataset. Removing flow warping in study (g) achieves 40.84 dB PSNR and 0.9821 SSIM ($\times 2$), and 33.80 dB PSNR and 0.9129 SSIM ($\times 4$), highlighting its utility but showing it is less effective than the iMamba block. Study (h) evaluates concatenation and attention, yielding similar results to flow warping with 40.80 dB PSNR and 0.9820 SSIM ($\times 2$), and 33.80 dB PSNR and 0.9129 SSIM ($\times 4$). Excluding interleaved reorganization in study (i) results in 40.99 dB PSNR and 0.9826 SSIM ($\times 2$), and 33.77 dB PSNR and 0.9126 SSIM ($\times 4$), demonstrating its benefits. Finally, study (j) removes channel attention, achieving 41.02 dB PSNR and 0.9829 SSIM ($\times 2$), and 33.94 dB PSNR and 0.9145 SSIM ($\times 4$), showing it significantly enhances global information utilization for improved texture recovery.

We show different variants of the iMamba block in Figure 2. The results highlight that neither deformable convolution nor flow-based warping achieves satisfactory performance. This limitation arises from inherent issues: deformable convolutions, despite leveraging offsets, struggle to capture global contextual information crucial for accurate reconstruction. On

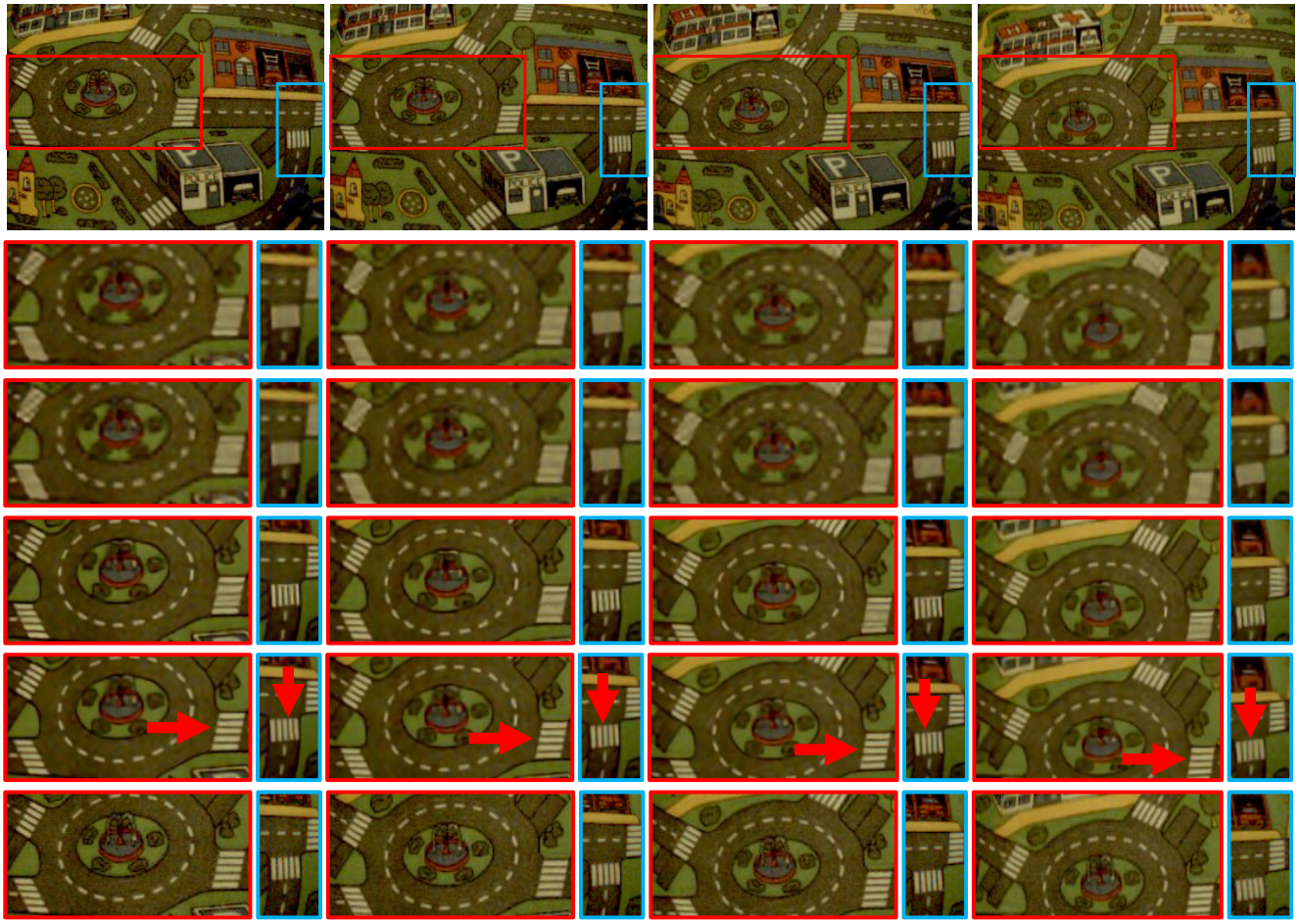

Figure 3. Visual results of different variants of the iMamba block. From top to bottom are ground truth image, and patches cropped from iMamba - EBVSR-BCS module, iMamba - Cross-modal att., iMamba w/o Cross SSM, iMamba, and the ground truth image.

the other hand, flow estimation faces significant challenges in low-resolution and texture-deficient regions, leading to imprecise motion alignment and constrained performance. Our interleaved reorganization mechanism, by contrast, maximizes information interaction and fusion between frames, allowing for more effective utilization of temporal dependencies. This approach outperforms the naive scanning strategy by facilitating richer feature integration and capturing global temporal-spatial correlations, resulting in superior reconstruction quality.

2.3. A Close Look At the cMamba Block

The cMamba block is designed for efficient cross-modal fusion. To analyze its effectiveness, we evaluate several variants. (1) cMamba - Residual blocks: we replace the cMamba block with residual blocks for cross-modal fusion. (2) cMamba - EBVSR-BCS module: we utilize the bidirectional cross-modal synthesis model in [1] for cross-modal fusion. (3) cMamba - Cross-modal att.: we employ the cross-modal attention operation for fusion. (4) cMamba - Concat & iMamba: we concatenate features from two frames along the feature dimension, followed by the iMamba blocks. (5) cMamba - w/o Cross SSM: we remove the cross SSM in cMamba. Removing the residual blocks (k) results in a noticeable decrease in both PSNR and SSIM for both scaling factors. The EBVSR-BCS module (l) and concat & imamba (n) also show a significant impact when removed, with notable decreases in performance metrics. Cross-modal SSM (m) has a positive effect on the performance, with improvements in both PSNR and SSIM compared to the baseline. This is the core of cMamba.

We show different variants of the cMamba block in Figure 3. Although the EBVSR-BCS module and cross-modal attention mechanisms have been demonstrated to be effective in event-based VSR and event-based motion deblurring tasks, they show inferior performance when compared to our proposed cMamba block. When the cMamba block is replaced with

either of these modules, the reconstructed images exhibit significant artifacts, such as noticeable aliasing and misalignment, particularly evident in areas like sidewalks. Even after removing the cross-modal attention mechanism, the performance gap between these alternatives and the cMamba block remains substantial. This underscores the cMamba block's superior capability in managing intricate spatiotemporal dependencies, leading to more accurate and artifact-free reconstructions.

2.4. Effectiveness of the reconstructor in MamEVSR

The configuration without the iMamba module (p) achieves marginally higher PSNR values for both scaling factors compared to the configuration with all iMamba modules (q). However, the difference is very small, indicating that the presence of the iMamba module does not significantly affect the PSNR metric. In terms of SSIM, the configuration without the iMamba module (p) also shows a slight improvement over the configuration with all iMamba modules (q), again with a negligible difference. Both configurations perform similarly well across both scaling factors, suggesting that the iMamba module may not play a crucial role in determining the overall performance of the model in terms of PSNR and SSIM. Overall, while the iMamba module appears to contribute positively to the model's performance, its removal does not lead to a significant degradation in the quality of the reconstructed images, as measured by PSNR and SSIM.

3. Discussion

We acknowledge the limitations of our work and provide a detailed discussion as follows: (1) Novelty and Design Considerations: As the first work to introduce the Mamba framework into event-based VSR, our key contribution lies in demonstrating its effectiveness in this domain. However, applying Mamba to event-based VSR is far from straightforward. The framework required specially tailored designs and improvements to ensure MamEVSR achieves state-of-the-art performance. Through comprehensive ablation studies, we validate the advantages and design motivations of the iMamba and cMamba blocks, which are central to the success of our approach. Specifically, iMamba reduces hidden state dependence by fusing local and global temporal information from the previous frame, minimizing error propagation. Additionally, residual learning refines the current frame, reducing accumulated errors. Empirical validation, such as the stable performance on the 100-frame REDS4 dataset, demonstrates robustness against long-term dependencies. (2) Dataset Limitations: While we evaluated our model on both the synthetic REDS and real-world CED datasets, achieving state-of-the-art results across the board, our experiments were constrained by limited GPU resources. This restricted our ability to conduct additional experiments on larger datasets. Furthermore, we chose not to include Vimeo90k due to its inherent limitations for event-based VSR: the dataset has low resolution, fewer frames (7 per sequence), and poorly simulated events, making it suboptimal for evaluating event-based methods. In future work, we plan to expand our experiments to include more comprehensive datasets as computational resources become available. (3) Future Directions: Beyond event-based VSR, the proposed Mamba framework holds significant potential for application in other event-based tasks. We aim to extend MamEVSR to tackle challenges such as event-based super-resolution, denoising, and HDR reconstruction, further validating the versatility and robustness of the framework. Moreover, integrating additional temporal and spatial priors could further enhance its performance across diverse scenarios.

References

[1] Dachun Kai, Yueyi Zhang, and Xiaoyan Sun. Video super-resolution via event-driven temporal alignment. In *ICIP*, 2023.

[2] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *ECCV*, 2022.

[3] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019.