Event-based Video Super-Resolution via State Space Models

Zeyu Xiao Xinchao Wang[*]
National University of Singapore
zeyuxiao@nus.edu.sg, xinchao@nus.edu.sg

Abstract

Exploiting temporal correlations is crucial for video super-resolution (VSR). Recent approaches enhance this by incorporating event cameras. In this paper, we introduce MamEVSR, a Mamba-based network for event-based VSR that leverages the selective state space model, Mamba. MamEVSR stands out by offering global receptive field coverage with linear computational complexity, thus addressing the limitations of convolutional neural networks and Transformers. The key components of MamEVSR include: (1) The interleaved Mamba (iMamba) block, which interleaves tokens from adjacent frames and applies multi-directional selective state space modeling, enabling efficient feature fusion and propagation across bi-directional frames while maintaining linear complexity. (2) The cross-modality Mamba (cMamba) block facilitates further interaction and aggregation between event information and the output from the iMamba block. The cMamba block can leverage complementary spatio-temporal information from both modalities and allows MamEVSR to capture finer motion details. Experimental results show that the proposed MamEVSR achieves superior performance on various datasets quantitatively and qualitatively.

1. Introduction

Video super-resolution (VSR) [18, 25–28] is a fundamental task in computer vision aimed at reconstructing high-resolution (HR) videos from low-resolution (LR) inputs. Given its broad applications in areas such as video surveillance [1, 39, 91], high-definition television [2–4, 12–15, 20, 45–47, 57, 79, 83], and satellite imagery [16, 54, 76, 77], VSR has attracted significant interest. The main challenge lies in effectively exploiting temporal information. Advanced VSR techniques often leverage temporal information through sliding window methods [29, 34, 42, 48, 53, 70, 73–75, 88], recurrent architectures [8–10, 36, 40, 61, 86], and Transformer-based designs [49, 50, 59, 63, 65, 66], all demonstrating significant achievements. Recently, researchers have integrated event cameras into VSR to boost performance by harnessing their high temporal resolution and dynamic range capabilities.

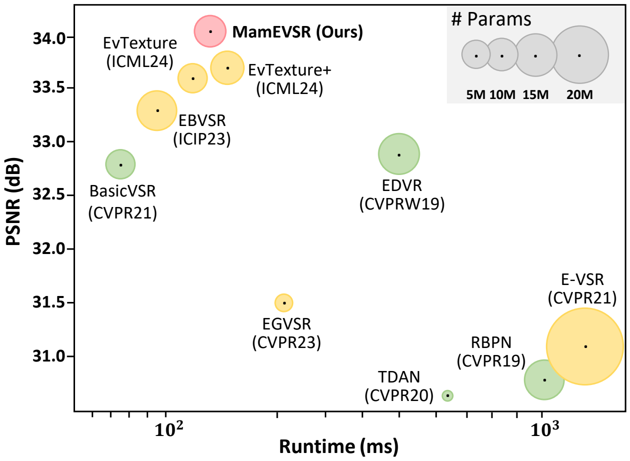

Figure 1. Inference time and performance comparison. MamEVSR outperforms advanced methods with high efficiency. Comparisons are performed on the CED dataset. The green circles represent RGB-based VSR methods, while the yellow circles represent event-based VSR methods.

Event cameras are bio-inspired sensors that asynchronously detect intensity changes at the microsecond level on a per-pixel basis, generating millions of events per second while maintaining robustness in high-dynamic-range lighting. Leveraging these advantages, recent event-based VSR methods [30, 33, 52, 81, 82] have outperformed frame-only approaches by enhancing flow estimation, temporal alignment, and cross-modal fusion using event streams. However, current methods, primarily based on CNNs and attention models, face limitations, including poor adaptability to dynamic inputs, insufficient receptive fields for capturing inter-frame correlations at high resolutions, and high computational costs.

On the other hand, natural language processing (NLP) has recently witnessed the emergence of structured state space models (SSMs) [22]. Theoretically, SSMs combine the benefits of recurrent neural networks (RNNs) and

[*] Corresponding author

CNNs, leveraging the global receptive field characteristic of RNNs and the computational efficiency of CNNs. One particularly notable SSM is the selective state space model, also known as Mamba [21], which has garnered significant attention within the vision community. Mamba's key feature is its ability to make SSM parameters time-variant (*i.e.*, data-dependent), enabling it to effectively select relevant context within sequences–a crucial factor for enhancing model performance. Inspired by Mamba's capabilities, we explore its potential in addressing the task of Event-based VSR, and we propose the *first* Mamba-based method for this task, named MamEVSR.

MamEVSR leverages a bidirectional recurrent pipeline to recurrently align and propagate temporally correlated features and integrate cross-modal information from the event stream. Its key components include: (1) The interleaved Mamba (iMamba) block: To facilitate efficient feature fusion and propagation across bi-directional frames while maintaining linear complexity, we introduce the iMamba block. Since the original SSMs model processes a single sequence, we merge tokens from two adjacent frames into one sequence for effective inter-frame modeling. By interleaving tokens from both frames and conducting multi-directional SSMs, we enable interactions between adjacent tokens from different propagated frames during sequence modeling. This design ensures that intermediate tokens in the sequence are from their spatio-temporal neighborhood. Stacking multiple iMamba blocks enhances MamEVSR's ability to handle complex inter-frame information exchange and frame alignment. (2) The cross-modality Mamba (cMamba) block: The cMamba block employs a cross-multiplication mechanism to enhance the interaction and aggregation between event information and the output from the iMamba block. This mechanism allows for the effective integration of event data with the temporal-fused features, thereby enriching the model's ability to leverage complementary cross-modal information and improve the overall reconstruction quality. By combining these two blocks, MamEVSR effectively exploits the strengths of both Mamba and event data to achieve superior performance in event-based VSR tasks. Experimental results on benchmark datasets demonstrate the effectiveness of our proposed MamEVSR (see an example in Figure 1).

Our contributions can be summarized as follows: (1) To the best of our knowledge, this marks the first successful application of state space models, specifically Mamba, in event-based VSR. (2) We propose the iMamba block and the cMamba block as core components in MamEVSR for event-based VSR. The former enables efficient feature fusion and propagation across bi-directional frames while maintaining linear complexity, the latter leverages complementary spatio-temporal information to capture finer motion details. (3) Extensive evaluations on benchmark datasets demonstrate MamEVSR's superior performance, establishing a new benchmark for future explorations of Mamba's potential within event-based VSR.

2. Related Work

Video super-resolution. Existing VSR methods are designed to enhance the quality of LR frames by leveraging temporal information through sliding windows and recurrent structures. Sliding-window-based methods align adjacent LR frames with a reference frame to estimate an HR output. Early work focused on explicit optical flow estimation for frame alignment [6, 53, 69, 80, 84]. More recent methods have shifted to implicit alignment using dynamic filters [31], deformable convolutions [70, 75], and attention modules [28, 43]. Recent attention-based methods achieve new state-of-the-art performance [7, 37, 38, 43, 49, 53, 75, 78, 80]. However, these methods face challenges in capturing long-distance temporal features. Recurrent-based methods use RNNs to exploit temporal information across multiple frames. Sajjadi *et al.* [61] introduce a recurrent framework with optical flow for alignment. BasicVSR [9] and BasicVSR++[10] utilize bi-directional hidden states and advanced grid propagation techniques. PSRT [63] builds on BasicVSR++ with multi-frame self-attention for feature processing. In this paper, we focus on event-based VSR and introduce MamEVSR.

Event-based VSR. Event cameras have the unique capability to measure intensity changes at each pixel independently with microsecond accuracy, making them valuable for VSR. Jing *et al.* [30] propose a two-stage method that uses events to interpolate low-resolution (LR) video, generating a high-frequency video for subsequent high-resolution (HR) frame reconstruction. Lu *et al.* [52] present a joint framework learning implicit neural representations from both RGB frames and events, enabling arbitrary-scale VSR. Kai *et al.* [32] introduce EvTexture, an event-driven texture enhancement network dedicated to restoring textures in VSR through the incorporation of high-frequency event signals. Xiao *et al.* [82] take inspiration from the parameter-efficient tuning, introducing an event adapter for event-based VSR. In this paper, we introduce MamEVSR, the first Mamba-based method for event-based VSR.

State space models. SSMs [17, 23, 44, 55, 64, 87, 90] are emerging as strong alternatives to Transformers [71] in natural language processing. S4 [23] is introduced for efficient, linear-complexity sequence modeling, followed by S5 [64], which improves parallelization, and GSS [55], which incorporates gated mechanisms. Mamba (S6) [21] stands out with its data-dependent parameters and hardware efficiency, outperforming Transformers in long-sequence tasks. In the computer vision domain, Vim [92] permutes 2D images into sequences for global modeling using bidirectional SSMs. Vmamba [51] extends this to four directions with a hier-

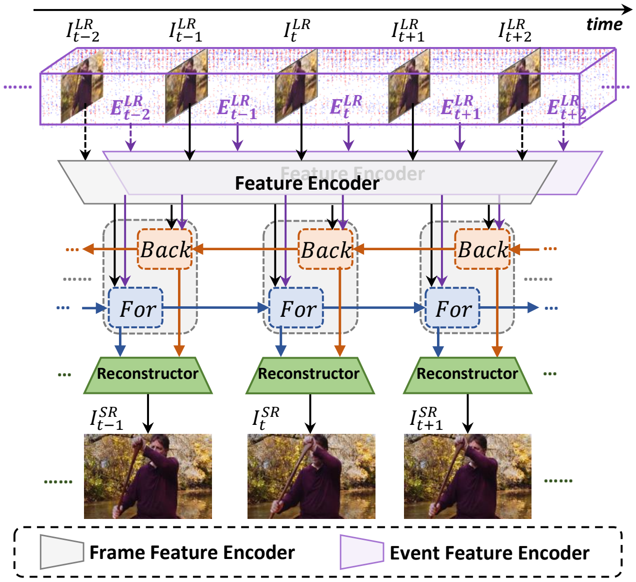

Figure 2. Overview of the proposed MamEVSR. MamEVSR follows a typical bi-directional propagation scheme [9], consisting of bi-directional recurrent cells $For(\cdot)$ and $Back(\cdot)$. For clarity, we have omitted the residual connection from the bicubic upsampling of the input, which is added to the final output, in the figure.

archical design. VideoMamba [41] applies S6 to spatio-temporal video sequences, while MambaIR [24] is the first to use S6 for image restoration, surpassing Transformer-based methods. In this work, we explore the potential of Mamba in event-based VSR.

3. Method

3.1. Revisiting State Space Models

State space models (SSM) [17, 23, 55, 64] represent a class of sequence-to-sequence modeling systems characterized by constant dynamics over time, a property also known as linear time-invariant. With linear complexity, SSM can effectively capture the inherent dynamics through an implicit mapping to latent states, which can be defined as

$$y(t) = Ch(t) + Dx(t), \dot{h}(t) = Ah(t) + Bx(t). \tag{1}$$

Here, $x(t) \in \mathbb{R}, h(t) \in \mathbb{R}^N$, and $y(t) \in \mathbb{R}$ denotes the input, hidden state, and the output, respectively. N is the state size and $\dot{h}(t)$ refers to the time derivative of $h(t)$. Additionally, $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, $C \in \mathbb{R}^{1\times N}$, and $D \in \mathbb{R}$ are the system matrices. To process discrete sequences like image and text, SSMs adopt Zero-Order Hold discretization [22] to map the input sequence $\{x_1, x_2, ..., x_K\}$ to the output sequence $\{y_1, y_2, ..., y_K\}$. Specifically, suppose $\Delta \in \mathbb{R}^D$ is the pre-defined timescale parameter to map continuous parameters A, B into a discrete space, the discretization process can be formulated as

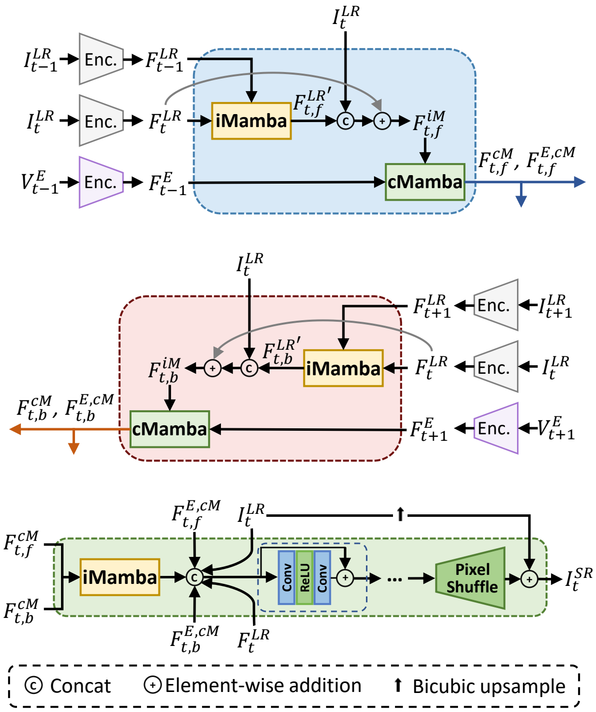

Figure 3. Overview of the forward recurrent cell $For(\cdot)$, the backward recurrent cell $Back(\cdot)$, and the reconstructor $R(\cdot)$.

$$\begin{aligned} &\overline{A} = \exp(\Delta A), \overline{B} = (\Delta A)^{-1}(\exp(A) - I)\Delta B, \overline{C} = C, \\ &y_k = \overline{C}h_k + \overline{D}x_k, h_k = \overline{A}h_{k-1} + \overline{B}x_k. \end{aligned} \tag{2}$$

Here, all the matrices keep the same dimension as the operation iterates. Notably, $\overline{D}$, serving as a residual connection, is often discarded in the equation

$$y_k = \overline{C}h_k. \tag{3}$$

Besides, following Mamba [21], the matrix $\overline{B}$ can be approximated by the first-order Taylor series

$$\overline{B} = (\exp(A) - I)A^{-1}B \approx (\Delta A)(\Delta A)^{-1}\Delta B = \Delta B. \tag{4}$$

The proposed iMamba and cMamba blocks are built upon Mamba and are specifically designed to address the challenges of event-based VSR tasks.

3.2. Overview

Let $\mathcal{I}^{LR} = \{\dots, I_{t-1}^{LR}, I_t^{LR}, I_{t+1}^{LR}, \dots\}$ ($I_t^{LR} \in \mathbb{R}^{H\times W\times 3}$) be a sequence of LR input frames, $\boldsymbol{E}^{f,b}$ be forward and backward event streams, the goal of MamEVSR is to generate $\mathcal{I}^{SR} = \{\dots, I_{t-1}^{SR}, I_t^{SR}, I_{t+1}^{SR}, \dots\}$ ($I_t^{SR} \in \mathbb{R}^{sH\times sW\times 3}$), which should be close to the ground-truth sequence $\mathcal{I}^{GT} = \{\dots, I_{t-1}^{GT}, I_t^{GT}, I_{t+1}^{GT}, \dots\}$ ($I_t^{GT} \in \mathbb{R}^{sH\times sW\times 3}$). T, H, and W are the frame number, height, and width, respectively. s is the upscaling factor.

Because the event streams are not convenient for observation and processing by convolutional neural networks, we convert forward and backward event streams as voxel grids $\boldsymbol{V}^{f,b} \in \mathbb{R}^{(T-1)\times H\times W\times Bin}$ via the temporal bilinear interpolation scheme

$$V(i) = \sum_k p_k \max\left(0, 1-\left|(i-1)-\frac{t_k-t_0}{t_{N_e}-t_0}\,(Bin-1)\right|\right), \quad (5)$$

where $i \in \{1, \cdots, Bin\}$ represents the i-th time bin. In our experiments, consistent with previous studies [32, 33], we also set $Bin = 5$. Furthermore, to mitigate the impact of hot pixels, we follow the study [32, 33] and normalize the voxel grid V as

$$\hat{V}(i) = \min((V(i), \eta))/\eta, \quad (6)$$

where η is the 98-th percentile value in the non-zero values of V. In this way, we obtain the normalized voxel grid $\hat{V} \in \mathbb{R}^{H\times W\times Bin}$, which contains rich high-frequency textural information. We represent $\boldsymbol{V}^{f,b}$ using this voxel grid representation for further processing.

Figure 2 shows an overview of the proposed MamEVSR. MamEVSR employs bidirectional recurrent cells $For(\cdot)$ and $Back(\cdot)$ akin to the scheme proposed in [9]. However, it introduces novel elements, such as extra inputs and specialized modules to harness event streams, which sets it apart from prior approaches. The event voxels and the LR frames are first converted into the feature domain using the feature encoders (f^V_{En} and f^I_{En}, consisting of N_1 residual blocks), following which they are fed to the bidirectional recurrent cells. These cells consist of the iMamba block and the cMamba block. This step aims to utilize temporal correspondence from adjacent frames and aggregate cross-modal information from the event stream. As shown in Figure 3, using timestamp t as an example, the process can be expressed as

$$F^{cM}_{t,f}, F^{E,cM}_{t,f} = For(I^{LR}_{t-1}, I^{LR}_t, F^f_t), \quad (7)$$

$$F^{cM}_{t,b}, F^{E,cM}_{t,b} = Back(I^{LR}_{t+1}, I^{LR}_t, F^b_t). \quad (8)$$

Specifically, using the forward recurrent cell as an example, at time t, F^{LR}_{t-1} and F^{LR}_t are fed to the iMamba block for efficient feature fusion and propagation across adjacent frames, resulting in $F^{LR'}_{t,f}$. Next, $F^{LR'}_{t,f}$ and I^{LR}_t are concatenated along the channel dimension, followed by the residual of F^{LR}_t to enhance the representation capacity. The resulting features $F^{iM}_{t,f}$, along with F^f_t, are fed to the cMamba block, which facilitates further interaction and aggregation in a cross-modality manner. The outputs of the cMamba block, $F^{cM}_{t,f}$ and $F^{E,cM}_{t,f}$, are used for further reconstruction. The backward cell operates similarly.

To generate the super-resolved output I^{SR}_t at timestamp t, features $F^{cM}_{t,f}$, $F^{cM}_{t,b}$, $F^{E,cM}_{t,f}$, $F^{E,cM}_{t,b}$, along with I^{LR}_t and F^{LR}_t, are fed to the reconstructor, which consists of an iMamba block, N_2 residual blocks, and a pixel-shuffling operation. The process can be expressed as

$$I^{SR}_t = R(F^{cM}_{t,f}, F^{cM}_{t,b}, F^{E,cM}_{t,f}, F^{E,cM}_{t,b}, I^{LR}_t, F^{LR}_t). \quad (9)$$

3.3. Interleaved Mamba Block

Effective inter-frame modeling is essential for VSR [9, 75, 86]. Conventional methods [75] rely on convolutional layers and attention mechanisms to capture temporal dependencies, offering fast inference but struggling with limited receptive fields. More advanced approaches, such as SemanticLens [68], introduce semantic-aware alignment strategies to enhance feature correspondence, but at the cost of increased computational complexity. To achieve a better balance between efficiency and effectiveness, we explore Mamba [21], an SSM, for inter-frame modeling. By leveraging its long-range dependency modeling and selective state updates, Mamba enables robust temporal feature fusion while maintaining computational efficiency, making it well-suited for VSR.

As shown in Figure 4(a), we introduce the iMamba block, which follows a Transformer-like structure (*i.e.*, Norm → Attention → Norm → MLP) [89], with two key modifications: (1) We replace the attention mechanism with a custom block that scans feature maps from two input frames. Thanks to the interleaved token reorganization design, this enables global inter-frame modeling with linear complexity. (2) Inspired by MambaIR [24] and MLFSR [19], which address limitations in locality and inter-channel interaction, we substitute the original MLP with a channel-attention block to enhance feature fusion and propagation. Here, Norm indicate the normalization operation and MLP means the multilayer perceptron.

For inter-frame modeling, given two adjacent low-resolution frame features, F^{LR}_{t-1} and F^{LR}_t (in the forward direction), both are first fed into the patch embedding module. This is followed by a series of operations, including layer normalization, linear projection, depth-wise convolution, and token reorganization, resulting in the reorganized feature $F^{LR}_{Re,f}$. The selective SSM-based Mamba block is then applied to independently model each direction. After processing, the sequences are rearranged and merged via token partitioning, another round of layer normalization, and linear operations. Subsequently, a channel attention block with a residual connection is applied to enhance feature propagation and stability. The output of the iMamba block is denoted as $F^{LR'}_{t,f}$ for the forward direction. Similarly, for the backward direction, $F^{LR'}_{t,b}$ is obtained following the same operations.

The interleaved reorganization within the iMamba block offers significant advantages for VSR. By ensuring that aggregated tokens come from both spatial and temporal neighborhoods, this design enhances local modeling, which is essential for capturing fine details between adjacent frames. Minimizing the distance between spatially and temporally adjacent tokens preserves the local context and reduces noise, which is particularly beneficial for maintaining tem-

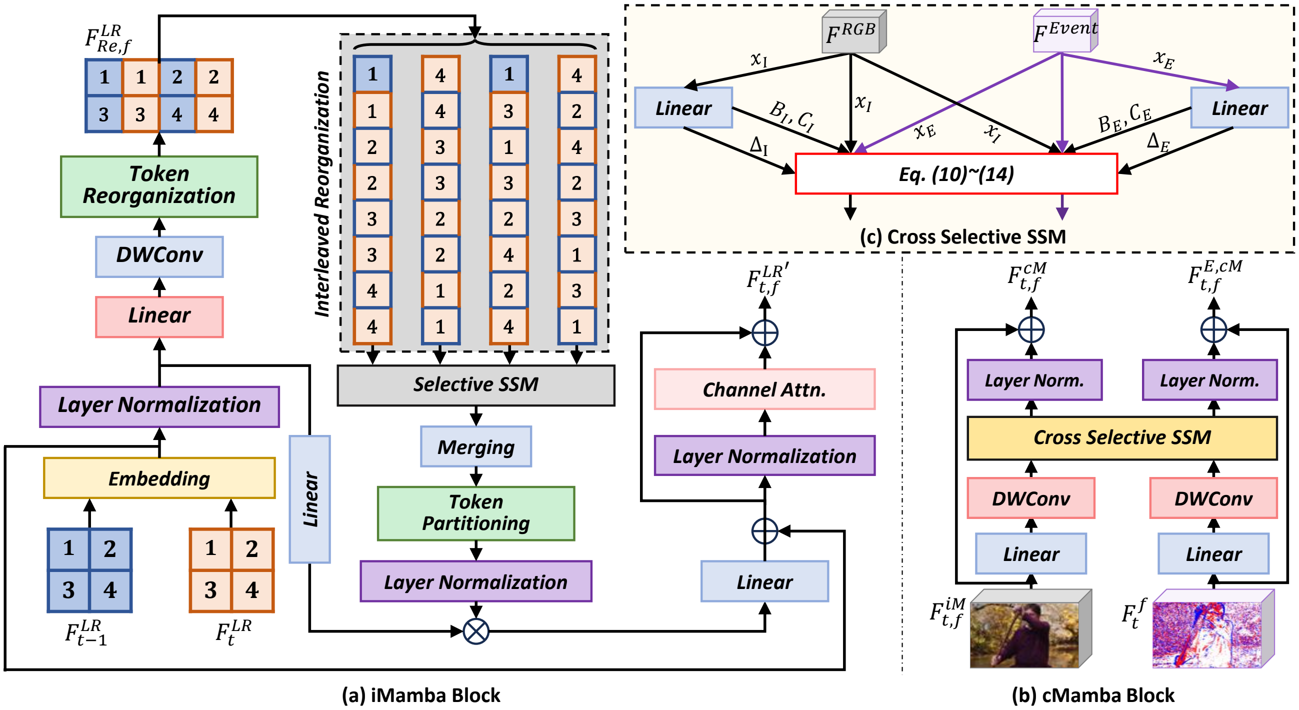

Figure 4. The structure of the proposed iMamba block and the cMamba block. The blocks offer specialized mechanisms to handle inter-frame modeling and cross-modal interactions effectively. See texts for more details.

poral coherence across frames. This improved context modeling directly enhances the quality of fine detail reconstruction, leading to superior video restoration results.

3.4. Cross-Modality Mamba Block

The fusion of event data and RGB information is crucial for event-based VSR. Event cameras capture high-temporal-resolution motion cues that complement the rich spatial details in RGB frames. By integrating this sparse, high-frequency motion data with RGB content, we can significantly improve motion estimation, edge preservation, and image quality, especially in challenging conditions like low light or fast motion. This fusion enhances temporal consistency and fine detail reconstruction, leading to superior VSR performance. We, therefore, design the cMamba block for cross-modality fusion.

Taking the forward direction as an example, as shown in Figure 4(b), the input features from the RGB frame output of the iMamba block, $F_{t,f}^{iM}$, and the event feature, F_t^f, are first processed by linear layers and depth-wise convolutions, generating F^{RGB} and F^{Event} before being fed to the cross selective SSM.

Following the Mamba selection mechanism described in Section 3.1, system matrices B, C, and Δ are generated to endow the model with context-aware capabilities, with linear projection layers generating these matrices. According to Eq. 2, matrix C decodes information from the hidden state h_k to produce the output y_k. Inspired by the cross-attention mechanism [11], we enable cross-modal interaction between the RGB and event modalities through information exchange between multiple selective scan modules. In particular, the process can be represented as

$$\overline{A}_I = \exp(\Delta_\mathrm{I} A_I),\ \overline{A}_E = \exp(\Delta_\mathrm{E} A_E), \tag{10}$$

$$\overline{B}_I = \Delta_{\mathrm{rgb}} B_I,\ \overline{B}_E = \Delta_{\mathrm{x}} B_E, \tag{11}$$

$$h_I^t = \overline{A}_I h_I^{t-1} + \overline{B}_I x_I^t,\ h_E^t = \overline{A}_E h_E^{t-1} + \overline{B}_E x_E^t, \tag{12}$$

$$y_I^t = C_\mathrm{x} h_I^t + D_I x_I^t,\ y_E^t = C_{\mathrm{rgb}} h_E^t + D_E x_E^t, \tag{13}$$

$$y_I = [y_I^1, y_I^2, \ldots, y_I^l],\ y_E = [y_E^1, y_E^2, \ldots, y_E^l]. \tag{14}$$

Here, $x_{I/E}^t$ represents the input at time step t, and $y_{I/E}$ denotes the selective scan output. C_E and C_I are the cross-modal matrices used for recovering the outputs at each time step from the hidden states.

The outputs from the two modality branches—RGB and event—are individually fed into the layer normalization operation, followed by a residual connection. This results in the final cross-modality enhanced feature, which effectively integrates information from both branches, improving feature representation for tasks like VSR. The above process can be denoted as

$$\hat{F_{t,f}^{iM}}, \hat{F_t^f} = \mathrm{CSSM}(\mathrm{D}(\mathrm{Linear}(F_{t,f}^{iM})), \mathrm{D}(\mathrm{Linear}(F_t^f))) \tag{15}$$

$$F_{t,f}^{cM} = \mathrm{LN}(\hat{F_{t,f}^{iM}}) + F_{t,f}^{iM}, \tag{16}$$

$$F_{t,f}^{E,cM} = \mathrm{LN}(\hat{F_t^f}) + F_t^f. \tag{17}$$

Table 1. Quantitative comparison in terms of PSNR and SSIM for the ×2 event-based VSR task on the CED dataset. The best and the second best results are highlighted in **bold** and underlined, respectively. Results are from [32] and [82].

Type	Method	CED											
		Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6	Scene 7	Scene 8	Scene 9	Scene 10	Scene 11	Average
RGB-based VSR	DUF	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	31.09 0.9183
	SOF-VSR	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	31.84 0.9226
	TDAN	35.83 0.9540	32.12 0.9339	31.57 0.9466	35.73 0.9566	35.42 0.9536	37.75 0.9440	28.91 0.9062	32.54 0.9006	35.55 0.9541	30.67 0.9323	35.09 0.9561	33.74 0.9398
	RBPN	40.07 0.9868	34.15 0.9739	33.83 0.9739	39.56 0.9869	39.44 0.9859	40.33 0.9782	30.36 0.9648	34.91 0.9502	40.05 0.9878	31.51 0.9551	39.03 0.9862	36.66 0.9754
	BasicVSR	39.35 0.9784	39.81 0.9766	39.73 0.9832	39.60 0.9789	39.45 0.9778	42.71 0.9815	39.15 0.9748	36.97 0.9672	39.35 0.9776	38.45 0.9732	39.41 0.9799	39.57 0.9778
Event-based VSR	E-VSR	41.08 0.9891	34.77 0.9775	34.44 0.9773	40.49 0.9891	40.32 0.9880	40.80 0.9801	40.80 0.9801	35.16 0.9536	41.00 0.9978	31.79 0.9586	39.97 0.9884	37.32 0.9783
	EGVSR	38.78 0.9794	38.68 0.9750	38.67 0.9815	39.06 0.9798	38.93 0.9792	41.96 0.9831	38.03 0.9755	36.14 0.9635	38.84 0.9787	37.68 0.9726	38.86 0.9810	38.69 0.9771
	EBVSR	39.95 0.9811	40.23 0.9780	40.31 0.9849	40.22 0.9816	40.06 0.9805	43.08 0.9830	39.97 0.9754	37.24 0.9689	39.95 0.9802	38.88 0.9757	40.07 0.9804	40.14 0.9801
	EvTexture	40.39 0.9824	40.54 0.9789	40.75 0.9859	40.66 0.9829	40.45 0.9819	43.27 0.9834	40.53 0.9800	37.57 0.9705	40.35 0.9815	39.27 0.9769	40.54 0.9837	40.52 0.9813
	EvTexture+	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	40.57 0.9815
	MamEVSR	41.17 0.9848	40.65 0.9791	41.10 0.9866	41.46 0.9853	41.31 0.9845	43.27 0.9826	41.70 0.9844	37.76 0.9719	41.16 0.9839	39.37 0.9775	41.50 0.9863	**41.14** **0.9831**

The features $F_{t,f}^{cM}$ and $F_{t,f}^{E,cM}$, obtained after cross-modal fusion, are fed into the reconstructor (see Figure 3) for the final image reconstruction. These fused features effectively capture both the high-frequency event data and spatial richness from the RGB frames, enhancing the quality of the reconstructed output by leveraging complementary information from both modalities.

4. Experiments

4.1. Experimental Settings

Datasets. For training, we utilize the REDS [56] and CED [62] datasets. Specifically, the REDS4 subset (clips 000, 011, 015, 020) from the REDS training set is employed as our test set, evaluating performance in the RGB channels. Consistent with prior work [30, 32, 33], we simulate events using the ESIM [60] tool for these clips. These simulated events are subsequently transformed into voxel grids according to Eqs. (5) and (6). Voxels are downsampled through the Bicubic interpolation operation, aligning with the frame downsampling method [32]. Following previous event-based VSR studies [33, 52], we use the CED [62] dataset for training and evaluating on real-world scenes[1]. The dataset is captured with a DAVIS346 [5] event camera, which outputs temporally synchronized events and frames at a resolution of 346 × 260. When calculating the metrics (PSNR and SSIM), we exclude boundary 8 pixels and evaluate in the RGB channel.

Implementation details. We use 15 frames as input during training, with a mini-batch size of 6 and an input frame resolution of 64 × 64. We apply data augmentation techniques to the training data, including horizontal flips and random rotations of 90°, 180°, and 270°. MamEVSR is trained for 300K iterations using the Adam optimizer with a Cosine Annealing learning rate scheduler. Network architecture parameters are set to $N_1 = 2$ and $N_2 = 45$. Supervision is provided by the Charbonnier penalty loss [35]:

$$\mathcal{L} = \sqrt{\|I^{SR} - I^{GT}\|^2 + \varepsilon^2}, \tag{18}$$

where ε is set to 1×10^{-3} in our experiments. I^{SR} denotes the results generated by MamEVSR, and I^{GT} represents the ground-truth frames. We omit the subscript t for simplicity. The initial learning rate for MamEVSR is 2×10^{-4}. Training is conducted on 2 NVIDIA RTX 3090 GPUs.

4.2. Quantitative and Qualitative Comparisons

We compare the proposed MamEVSR with a wide range of potential methods that could be used to address event-based VSR, aiming to explore as many diverse and rich approaches as possible. (1) RGB-based VSR methods: DUF [31], SOF-VSR [72], RBPN [25], EDVR [75], BasicVSR [9], VRT [49], and TTVSR [50]. In particular, we exclude the event stream and solely feed the LR video frames into these VSR networks for reconstruction, resulting in the final reconstructed video output. (2) Event-based VSR methods: we compare our MamEVSR with E-VSR [30], EGVSR [52], EBVSR [33], EvTexture, and EvTexture+ [32] to provide a thorough evaluation.

Quantitative results. As illustrated in Table 1, Table 2 and Table 3, MamEVSR consistently outperforms other event-

[1] The 11 scenes are as follows: People_Dynamic_Wave, Indoors_Foosball_2, Simple_Wires_2, People_Dynamic_Dancing, People_Dynamic_Jumping, Simple_Fruit_Fast, Outdoor_Jumping_Infrared_2, Simple_Carpet_Fast, People_Dynamic_Armroll, Indoors_Kitchen_2, and People_Dynamic_Sitting [82].

Table 2. Quantitative comparison in terms of PSNR and SSIM for the ×4 event-based VSR task on the REDS4 dataset. The best and second-best results are highlighted in **bold** and underlined, respectively. Results are from [32].

REDS4	RGB-based VSR					Event-based VSR				
	DUF	EDVR	BasicVSR	TTVSR	VRT	EGVSR	EBVSR	EvTexture	EvTexture+	MamEVSR
000	27.30/0.7937	28.01/0.8250	28.40/0.8434	28.82/0.8566	28.85/0.8553	25.16/0.7066	28.44/0.8446	30.72/0.9082	–/–	30.87/0.9103
011	28.38/0.8056	32.17/0.8864	32.47/0.8979	33.47/0.9100	33.49/0.9072	26.56/0.7722	32.55/0.8987	33.72/0.9145	–/–	33.87/0.9166
015	31.55/0.8846	34.06/0.9206	34.18/0.9224	35.01/0.9325	35.26/0.9332	29.83/0.8526	34.22/0.9235	35.06/0.9314	–/–	35.21/0.9335
020	27.30/0.8164	30.09/0.8881	30.63/0.9000	31.17/0.9094	31.16/0.9078	25.94/0.7846	30.67/0.9009	31.65/0.9154	–/–	31.80/0.9175
Average	28.63/0.8251	31.09/0.8800	31.42/0.8909	32.12/0.9021	32.19/0.9006	26.87/0.7790	31.47/0.8919	32.79/0.9174	32.93/**0.9195**	**32.94/0.9195**
#Params	5.8	20.6	6.3	6.8	35.6	2.6	12.2	8.9	10.1	9.6

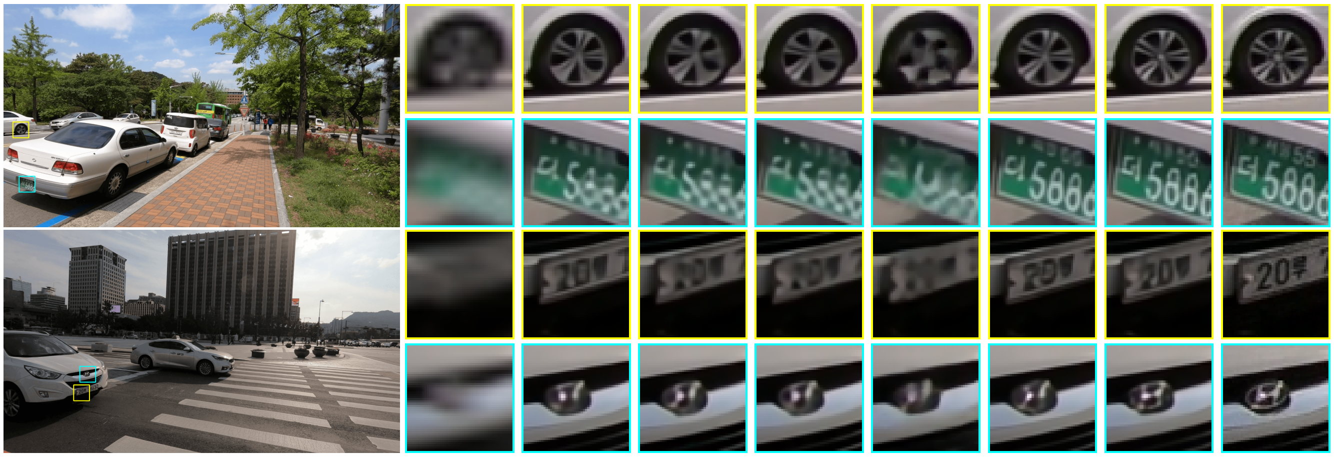

Figure 5. Visual comparisons for ×4 VSR on REDS4. From left to right in sequence are patches cropped from LR, TDAN, BasicVSR, EBVSR, EGVSR, EvTexture, MamEVSR, and the ground truth image.

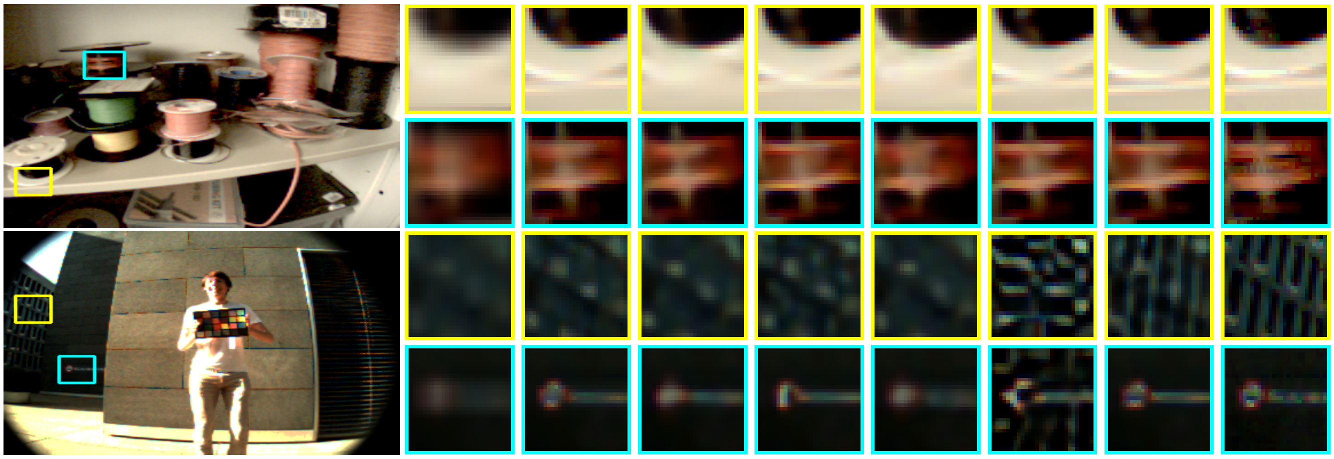

Figure 6. Visual comparisons for ×4 VSR on CED. From left to right in sequence are patches cropped from LR, EDVR, BasicVSR, EBVSR, EGVSR, EvTexture, MamEVSR, and the ground truth image.

Table 3. Quantitative comparison in terms of PSNR and SSIM for the ×2 event-based VSR task on the CED dataset. The best and the second best results are highlighted in **bold** and underlined.

Method	PSNR (↑)	SSIM (↑)	Method	PSNR (↑)	SSIM (↑)
SOF-VSR	31.84	0.9226	EGVSR	38.69	0.9771
TDAN	33.74	0.9398	EBVSR	40.14	0.9801
RBPN	36.66	0.9754	EvTexture	40.52	0.9813
BasicVSR	39.57	0.9778	EvTexture+	40.57	0.9815
E-VSR	37.32	0.9783	MamEVSR	**41.14**	**0.9831**

based VSR methods such as E-VSR, EGVSR, EBVSR, EvTexture, and EvTexture+ across various datasets like CED and REDS4. For instance, on the CED dataset for the ×2 event-based VSR task, MamEVSR achieves an average PSNR of 33.74 dB and an SSIM of 0.9831, surpassing the next best method, EvTexture+, which scores 33.40 dB and 0.9827 respectively. Compared to RGB-based VSR methods like DUF, TDAN, RBPN, and BasicVSR, MamEVSR demonstrates superior performance. On the CED dataset for the ×2 event-based VSR task, MamEVSR's average PSNR of 33.74 dB and SSIM of 0.9831 far exceed those of the top-performing RGB-based method, BasicVSR, which achieves

Table 4. Ablation study of different components on CED.

Method		CED ×2		CED ×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 [33]	40.77	0.9819	33.80	0.9139
	(m) Cross-modal att. [67]	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

31.69 dB and 0.9226 respectively. Across all datasets and magnification factors, MamEVSR consistently delivers superior performance, showcasing its effectiveness in enhancing event-based video super-resolution tasks compared to both event-based and RGB-based VSR methods.

Computational cost results. We calculate the number of parameters and runtime on the CED dataset. Results are shown in Table 2 and Figure 1. MamEVSR is positioned in the upper-left quadrant, reflecting an optimal trade-off between high performance and low computational cost. It achieves high-quality reconstructions (PSNR) with efficient computation and a modest parameter count, showcasing effective resource utilization.

Qualitative results. We present visual comparison results on REDS4 and CED in Figure 5 and Figure 6, respectively. The LR patch appears blurry and lacks detail as we move through the SR patches; sharpness and clarity noticeably increase. MamEVSR produces results that closely resemble the ground truth, effectively restoring fine details and textures. For example, MamEVSR more accurately reconstructs wheel textures and license plate details, closely matching the ground truth in Figure 5.

4.3. Ablation Study

We conduct experiments on CED in terms of PSNR and SSIM. Results are in Table 4.

Effectiveness of core components. Removing the iMamba block decreases PSNR and SSIM for both CED ×2 and CED ×4 tasks. Omitting the cMamba block has a smaller impact on performance. Disabling either backward or forward cells reduces PSNR and SSIM, especially for CED ×4. These results confirm the importance of the iMamba and cMamba blocks, as well as the bidirectional cell structure, in achieving optimal performance.

Effectiveness of the iMamba block. Flow warping performs better than concatenation and attention. Removing internal reorganization or attention mechanisms within the iMamba block causes slight drops in PSNR and SSIM. The iMamba block plays a critical role in improving MamEVSR's performance.

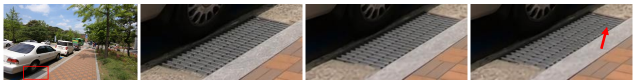

Figure 7. A failure case. From left to right in sequence are patches cropped from the ground truth image, EvTexture, and MamEVSR.

Effectiveness of the cMamba block. Incorporating the EBVSR-CBCS module [33] marginally boosts PSNR and SSIM. Cross-modal attention [67] yields comparable performance gains. Concatenating and using iMamba shows a slight decrease in metrics. Eliminating the cross-state space modeling mechanism within the cMamba block leads to minimal reductions in PSNR and SSIM.

Effectiveness of the reconstructor. Removing the iMamba block from the reconstructor slightly improves PSNR and SSIM for both CED ×2 and CED ×4 tasks. However, replacing all iMamba blocks with residual blocks leads to a notable drop in performance, particularly for CED ×4.

4.4. Limitations and Discussions

While MamEVSR demonstrates strong results in event-based VSR, some challenges remain. As shown in Figure 7, MamEVSR struggles to recover particularly fine textures due to significant information loss in RGB data and low-resolution, poorly defined event data. Potential solutions include incorporating VGG loss to generate more realistic textures, as it effectively captures fine details in restoration tasks. Alternatively, using a longer temporal sequence of both event and RGB data could provide richer, continuous information, enhancing texture recovery. Additionally, we plan to design lighter-weight modules for event-based VSR and explore extending MamEVSR to other video restoration tasks. Investigating the effectiveness of other novel architectures [58, 85] for this task would also be insightful.

5. Conclusion

In this paper, we propose MamEVSR for event-based VSR. MamEVSR leverages the selective state space model to offer global receptive field coverage with linear complexity. Key components include the iMamba block for efficient bidirectional feature fusion and the cMamba block for integrating event information and capturing finer motion details. Experiments show that MamEVSR outperforms existing methods on various datasets, achieving superior quantitative and qualitative results.

Acknowledgment. This project is supported by the National Research Foundation, Singapore, under its Medium Sized Center for Advanced Robotics Technology Innovation.

References

[1] Andreas Aakerberg, Kamal Nasrollahi, and Thomas B Moeslund. Real-world super-resolution of face-images from surveillance cameras. *IET Image Processing*, 16(2):442–452, 2022.

[2] Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *CVPR*, 2025.

[3] Haowen Bai, Zixiang Zhao, Jiangshe Zhang, Baisong Jiang, Lilun Deng, Yukun Cui, Shuang Xu, and Chunxia Zhang. Deep unfolding multi-modal image fusion network via attribution analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[4] Haowen Bai, Zixiang Zhao, Jiangshe Zhang, Yichen Wu, Lilun Deng, Yukun Cui, Baisong Jiang, and Shuang Xu. Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning. *International Journal of Computer Vision*, pages 1–21, 2024.

[5] Christian Brandli, Lorenz Muller, and Tobi Delbruck. Real-time, high-speed video decompression using a frame-and event-based davis sensor. In *ISCAS*, pages 686–689, 2014.

[6] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, 2017.

[7] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[8] Jiezhang Cao, Jingyun Liang, Kai Zhang, Wenguan Wang, Qin Wang, Yulun Zhang, Hao Tang, and Luc Van Gool. Towards interpretable video super-resolution via alternating optimization. In *ECCV*, 2022.

[9] 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 *CVPR*, 2021.

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

[11] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, 2021.

[12] Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *CVPR*, 2023.

[13] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. *NeurIPS*, 2025.

[14] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Haoze Sun, Xueyi Zou, Zhensong Zhang, Youliang Yan, and Lei Zhu. Low-res leads the way: Improving generalization for super-resolution by self-supervised learning. In *CVPR*, 2024.

[15] Haoyu Chen, Jingjing Ren, Jinjin Gu, Hongtao Wu, Xuequan Lu, Haoming Cai, and Lei Zhu. Snow removal in video: A new dataset and a novel method. in 2023 ieee. In *ICCV*.

[16] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*, 2020.

[17] Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. Hungry hungry hippos: Towards language modeling with state space models. In *ICLR*, 2022.

[18] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *ICCVW*, 2019.

[19] Ruisheng Gao, Zeyu Xiao, and Zhiwei Xiong. Mamba-based light field super-resolution with efficient subspace scanning. In *ACCV*, 2024.

[20] Tomio Goto, Takafumi Fukuoka, Fumiya Nagashima, Satoshi Hirano, and Masaru Sakurai. Super-resolution system for 4k-hdtv. In *ICPR*, 2014.

[21] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[22] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[23] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *ICLR*, 2021.

[24] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024.

[25] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, 2019.

[26] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. *NeurIPS*, 28, 2015.

[27] Yan Huang, Wei Wang, and Liang Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):1015–1028, 2017.

[28] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *ECCV*, 2020.

[29] 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 *CVPR*, 2022.

[30] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Turning frequency to resolution: Video super-resolution via event cameras. In *CVPR*, 2021.

[31] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, 2018.

[32] Dachun Kai, Jiayao Lu, Yueyi Zhang, and Xiaoyan Sun. Evtexture: Event-driven texture enhancement for video super-resolution. *arXiv preprint arXiv:2406.13457*, 2024.

[33] Dachun Kai, Yueyi Zhang, and Xiaoyan Sun. Video super-

resolution via event-driven temporal alignment. In *ICIP*, 2023.

[34] Soo Ye Kim, Jeongyeon Lim, Taeyoung Na, and Munchurl Kim. 3DSRnet: Video super-resolution using 3d convolutional neural networks. *arXiv preprint arXiv:1812.09079*, 2018.

[35] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017.

[36] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *CVPR*, 2022.

[37] Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, and Luoqi Liu. Dropkey for vision transformer. In *CVPR*, 2023.

[38] Bonan Li, Zicheng Zhang, Xuecheng Nie, Congying Han, Yinhan Hu, and Tiande Guo. Styo: Stylize your face in only one-shot. *arXiv preprint arXiv:2303.03231*, 2023.

[39] Bonan Li, Zicheng Zhang, Xingyi Yang, and Xinchao Wang. Focus on neighbors and know the whole: Towards consistent dense multiview text-to-image generator for 3d creation. *arXiv preprint arXiv:2408.13149*, 2024.

[40] Fei Li, Linfeng Zhang, Zikun Liu, Juan Lei, and Zhenbo Li. Multi-frequency representation enhancement with privilege information for video super-resolution. In *ICCV*, 2023.

[41] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024.

[42] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *CVPR*, 2019.

[43] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. MuCAN: Multi-correspondence aggregation network for video super-resolution. In *ECCV*, 2020.

[44] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *NeurIPS*, 37, 2025.

[45] Zhuoyuan Li, Jiacheng Li, Yao Li, Li Li, Dong Liu, and Feng Wu. In-loop filtering via trained look-up tables. In *VCIP*, 2024.

[46] Zhuoyuan Li, Junqi Liao, Chuanbo Tang, Haotian Zhang, Yuqi Li, Yifan Bian, Xihua Sheng, Xinmin Feng, Yao Li, Changsheng Gao, et al. Ustc-td: A test dataset and benchmark for image and video coding in 2020s. *arXiv preprint arXiv:2409.08481*, 2024.

[47] Zhuoyuan Li, Zikun Yuan, Li Li, Dong Liu, Xiaohu Tang, and Feng Wu. Object segmentation-assisted inter prediction for versatile video coding. *IEEE Transactions on Broadcasting*, 2024.

[48] Wenyi Lian and Wenjing Lian. Sliding window recurrent network for efficient video super-resolution. In *ECCV*, 2022.

[49] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022.

[50] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *CVPR*, 2022.

[51] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[52] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In *CVPR*, 2023.

[53] Zhihe Lu, Zeyu Xiao, Jiawang Bai, Zhiwei Xiong, and Xinchao Wang. Can sam boost video super-resolution? *arXiv preprint arXiv:2305.06524*, 2023.

[54] Yimin Luo, Liguo Zhou, Shu Wang, and Zhongyuan Wang. Video satellite imagery super resolution via convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2398–2402, 2017.

[55] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947*, 2022.

[56] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*, 2019.

[57] Tianbo Pan, Zidong Cao, and Lin Wang. Srfnet: Monocular depth estimation with fine-grained structure via spatial reliability-oriented fusion of frames and events. In *ICRA*, 2024.

[58] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

[59] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *ECCV*, 2022.

[60] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *CoRL*, 2018.

[61] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018.

[62] Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Ced: Color event camera dataset. In *CVPRW*, 2019.

[63] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *NeurlPS*, 2022.

[64] Jimmy TH Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *ICLR*, 2022.

[65] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Compact transformer tracker with correlative masked modeling. In *AAAI*, 2023.

[66] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *CVPR*, 2022.

[67] 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.

[68] Qi Tang, Yao Zhao, Meiqin Liu, Jian Jin, and Chao Yao. Semantic lens: Instance-centric semantic alignment for video

super-resolution. In *AAAI*, 2024.

[69] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017.

[70] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020.

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[72] Longguang Wang, Yulan Guo, Zaiping Lin, Xinpu Deng, and Wei An. Learning for video super-resolution through HR optical flow estimation. In *ACCV*, 2018.

[73] Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu, and Jianke Zhu. Lidar2map: In defense of lidar-based semantic map construction using online camera distillation. In *CVPR*, 2023.

[74] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *CVPR*, 2024.

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

[76] Yi Xiao, Xin Su, Qiangqiang Yuan, Denghong Liu, Huanfeng Shen, and Liangpei Zhang. Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2021.

[77] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Xianyu Jin, Jiang He, Liangpei Zhang, and Chia-wen Lin. Local-global temporal difference learning for satellite video super-resolution. *arXiv preprint arXiv:2304.04421*, 2023.

[78] Zeyu Xiao, Jiawang Bai, Zhihe Lu, and Zhiwei Xiong. A dive into sam prior in image restoration. *arXiv preprint arXiv:2305.13620*, 2023.

[79] Zeyu Xiao, Zhen Cheng, and Zhiwei Xiong. Space-time super-resolution for light field videos. *IEEE Transactions on Image Processing*, 32:4785–4799, 2023.

[80] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. Space-time distillation for video super-resolution. In *CVPR*, pages 2113–2122, 2021.

[81] Zeyu Xiao, Dachun Kai, Yueyi Zhang, Xiaoyan Sun, and Zhiwei Xiong. Asymmetric event-guided video super-resolution. In *ACMMM*, pages 2409–2418, 2024.

[82] Zeyu Xiao, Dachun Kai, Yueyi Zhang, Zheng-Jun Zha, Xiaoyan Sun, and Zhiwei Xiong. Event-adapted video super-resolution. In *ECCV*, 2024.

[83] Zeyu Xiao, Yutong Liu, Ruisheng Gao, and Zhiwei Xiong. Cutmib: Boosting light field super-resolution via multi-view image blending. In *CVPR*, 2023.

[84] Zeyu Xiao and Zhiwei Xiong. Incorporating degradation estimation in light field spatial super-resolution. *Computer Vision and Image Understanding*, page 104295, 2025.

[85] Xingyi Yang and Xinchao Wang. Kolmogorov-arnold transformer. *ICLR*, 2025.

[86] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Tao Lu, Xin Tian, and Jiayi Ma. Omniscient video super-resolution. In *ICCV*, pages 4429–4438, 2021.

[87] Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? *CVPR*, 2025.

[88] Huanjing Yue, Zhiming Zhang, and Jingyu Yang. Real-rawvsr: Real-world raw video super-resolution with a benchmark dataset. In *ECCV*, 2022.

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

[90] Guozhen Zhang, Chuxnu Liu, Yutao Cui, Xiaotong Zhao, Kai Ma, and Limin Wang. Vfimamba: Video frame interpolation with state space models. *NeurIPS*, 37, 2024.

[91] Liangpei Zhang, Hongyan Zhang, Huanfeng Shen, and Pingxiang Li. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859, 2010.

[92] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.