MambaSeg: Harnessing Mamba for Accurate and Efficient Image-Event Semantic Segmentation

Fuqiang Gu[1,2], Yuanke Li[2], Xianlei Long[1*], Kangping Ji[2], Chao Chen[1], Qingyi Gu[3], Zhenliang Ni[3*]

[1]College of Computer Science, Chongqing University
[2]National Elite Institute of Engineering, Chongqing University
[3]Institute of Automation, Chinese Academy of Sciences
gufq@cqu.edu.cn, lyk@cqu.edu.cn, xianlei.long@cqu.edu.cn, jkp@stu.cqu.edu.cn, cschaochen@cqu.edu.cn, qingyi.gu@ia.ac.cn nizhenliang@outlook.com

Abstract

Semantic segmentation is a fundamental task in computer vision with wide-ranging applications, including autonomous driving and robotics. While RGB-based methods have achieved strong performance with CNNs and Transformers, their effectiveness degrades under fast motion, low-light, or high dynamic range conditions due to limitations of frame cameras. Event cameras offer complementary advantages such as high temporal resolution and low latency, yet lack color and texture, making them insufficient on their own. To address this, recent research has explored multimodal fusion of RGB and event data; however, many existing approaches are computationally expensive and focus primarily on spatial fusion, neglecting the temporal dynamics inherent in event streams. In this work, we propose MambaSeg, a novel dual-branch semantic segmentation framework that employs parallel Mamba encoders to efficiently model RGB images and event streams. To reduce cross-modal ambiguity, we introduce the Dual-Dimensional Interaction Module (DDIM), comprising a Cross-Spatial Interaction Module (CSIM) and a Cross-Temporal Interaction Module (CTIM), which jointly perform fine-grained fusion along both spatial and temporal dimensions. This design improves cross-modal alignment, reduces ambiguity, and leverages the complementary properties of each modality. Extensive experiments on the DDD17 and DSEC datasets demonstrate that MambaSeg achieves state-of-the-art segmentation performance while significantly reducing computational cost, showcasing its promise for efficient, scalable, and robust multimodal perception.

Code — https://github.com/CQU-UISC/MambaSeg

Introduction

Semantic segmentation is a fundamental task in computer vision with broad applications in autonomous driving, robotics, and scene understanding (Lateef and Ruichek 2019). Driven by advances in convolutional neural networks (CNNs) and Transformers, RGB image-based methods have achieved significant progress (Badrinarayanan, Kendall, and Cipolla 2017; Xie et al. 2021; Ni et al. 2024; Ma, Ni, and Chen 2024a). However, their reliance on conventional image sensors poses inherent limitations under adverse conditions such as high-speed motion, low illumination, and high dynamic range scenes, where motion blur and latency severely degrade performance (Rebecq et al. 2019).

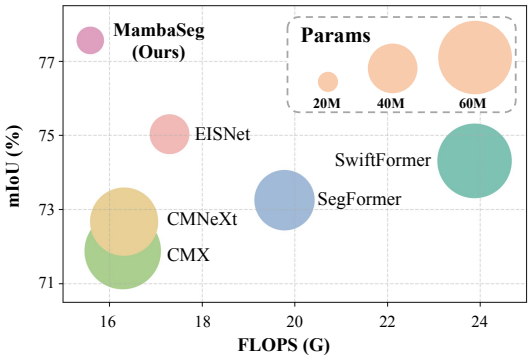

Figure 1: Comparison of event-image fusion segmentation methods in terms of mIoU (%), Parameters (M), and multiply-accumulate operations (G).

Event cameras offer a compelling alternative by asynchronously capturing per-pixel intensity changes with microsecond latency, high temporal resolution, and wide dynamic range (Gallego et al. 2020). These properties make them particularly well-suited for dynamic and low-light environments (Gehrig and Scaramuzza 2024). Yet, event streams lack color and fine-grained texture information, limiting their standalone utility in dense prediction tasks such as semantic segmentation (Sun et al. 2022; Jia et al. 2023). To harness the complementary strengths of both modalities, recent research has focused on RGB-event fusion techniques that combine the spatial richness of images with the temporal precision of events.

State-of-the-art multimodal methods typically adopt Transformer-based backbones for joint feature modeling (Zhang et al. 2023a,b). These models leverage cross-attention mechanisms for modality interaction and often incorporate domain adaptation to transfer semantics from labeled image data to sparsely labeled event streams (Sun

*Corresponding author: *Xianlei Long and Zhenliang Ni.*

et al. 2022; Xie, Gao, and Guo 2024). While effective, such approaches are computationally demanding due to the quadratic complexity of self-attention and the high dimensionality of combined inputs. Moreover, most existing fusion strategies emphasize spatial-level alignment while underutilizing the unique temporal dynamics of event data, which can result in suboptimal cross-modal synergy and semantic inconsistency.

To overcome these challenges, we draw inspiration from Mamba (Gu and Dao 2023), a recently proposed state space model that supports input-conditioned, long-range sequence modeling with linear computational complexity. Mamba has demonstrated strong performance and scalability in a variety of vision tasks, including classification (Liu et al. 2024; Ma, Ni, and Chen 2024b), detection (Wang et al. 2024), and segmentation (Ruan, Li, and Xiang 2024). Motivated by these advantages, we present MambaSeg, a novel dual-branch semantic segmentation framework that leverages parallel Mamba encoders to model RGB images and event streams independently, enabling efficient and scalable multimodal representation learning.

To reduce cross-modal ambiguity, we introduce the Dual-Dimensional Interaction Module (DDIM) to enhance the information exchange between different modalities, which fuses image and event features along both spatial and temporal dimensions. DDIM is composed of two key components: the Cross-Spatial Interaction Module (CSIM), which aligns spatial semantics by integrating dense texture features from images with structural edge cues from events; and the Cross-Temporal Interaction Module (CTIM), which exploits Mamba's global modeling capacity to refine temporal dependencies via attention-guided fusion. This design ensures fine-grained, modality-aware feature integration while minimizing computational overhead. We validate MambaSeg on two public benchmarks, DDD17 and DSEC. Our method achieves 77.56% mIoU on DDD17 and 75.11% mIoU on DSEC, setting new SOTA performance while maintaining high efficiency. These results highlight the strength of our MambaSeg for multimodal segmentation.

Our main contributions are summarized as follows:

- We propose MambaSeg, a dual-branch semantic segmentation framework based on parallel Mamba encoders that model image and event modalities efficiently with long-range dependency modeling and linear complexity.
- We design the DDIM module, which includes CSIM and CTIM for structured spatial-temporal fusion. This module jointly leverages Mamba and attention to enhance cross-modal feature alignment and reduce ambiguity.
- We conduct extensive experiments on the DDD17 and DSEC datasets, where MambaSeg achieves SOTA performance while significantly reducing computational cost compared to Transformer-based baselines.

Related Work

Semantic Segmentation

RGB-based semantic segmentation has advanced significantly with CNNs and Transformers. Encoder-decoder architectures like SegFormer (Xie et al. 2021) and SegNeXt (Guo et al. 2022) capture both spatial details and high-level semantics, achieving SOTA results. However, their dependence on conventional sensors limits robustness in fast motion, low-light, and high dynamic range scenarios, where motion blur and latency degrade performance.

To address these issues, event cameras have attracted interest for their high temporal resolution, low latency, and wide dynamic range. EV-SegNet (Alonso and Murillo 2019) pioneered event-based segmentation. Due to scarce labeled event data, methods like ESS (Sun et al. 2022) and CMESS (Xie, Gao, and Guo 2024) employ unsupervised domain adaptation to transfer semantic knowledge from images. Recent Transformer-based models such as EvSegFormer (Jia et al. 2023) integrate motion priors into attention to exploit event dynamics. Spiking neural networks (SNNs), including Spike-BRGNet (Long et al. 2024) and SLTNet (Long et al. 2025), have also been used to model temporal information efficiently.

RGB-event fusion leverages complementary spatial and temporal cues for robust perception. Hybrid frameworks like HALSIE (Das Biswas et al. 2024) and SpikingEDN (Zhang et al. 2024) combine ANNs and SNNs to capture static textures and dynamic edges. RGB-X methods such as CMX (Zhang et al. 2023a) and CMNeXt (Zhang et al. 2023b) employ multi-scale alignment and cross-attention for spatial fusion. EISNet (Xie et al. 2024) further introduces gated attention and progressive recalibration for adaptive feature alignment.

However, many methods rely on computationally expensive Transformer attention and suffer from cross-modal ambiguity. To address this, we propose a cross-modal interaction module that enhances spatial-temporal complementarity, improving fusion quality and segmentation performance.

State Space Models

State Space Models (SSMs), introduced in S4 (Gu, Goel, and Ré 2021), exhibit strong global modeling and long-range dependency handling, surpassing CNNs and Transformers in sequence tasks. Subsequent work (Smith, Warrington, and Linderman 2022) enhanced efficiency via parallelizable linear-complexity scans. Mamba (Gu and Dao 2023) advances this with an input-dependent selective scanning mechanism for dynamic sequence processing. Vision Mamba (Liu et al. 2024) adapts it for visual tasks through multi-directional state propagation, achieving strong image classification results. The framework has since succeeded in medical segmentation (Guo et al. 2024) and point cloud analysis (Wang et al. 2024), demonstrating its versatility across vision applications.

However, the potential of Mamba in multimodal fusion, particularly in integrating dense RGB images with sparse event data, remains largely untapped. To address this, we propose a novel Mamba-Attention architecture designed for cross-modal segmentation, combining Mamba's efficient sequence modeling with structured spatial-temporal fusion for improved alignment and representation across modalities.

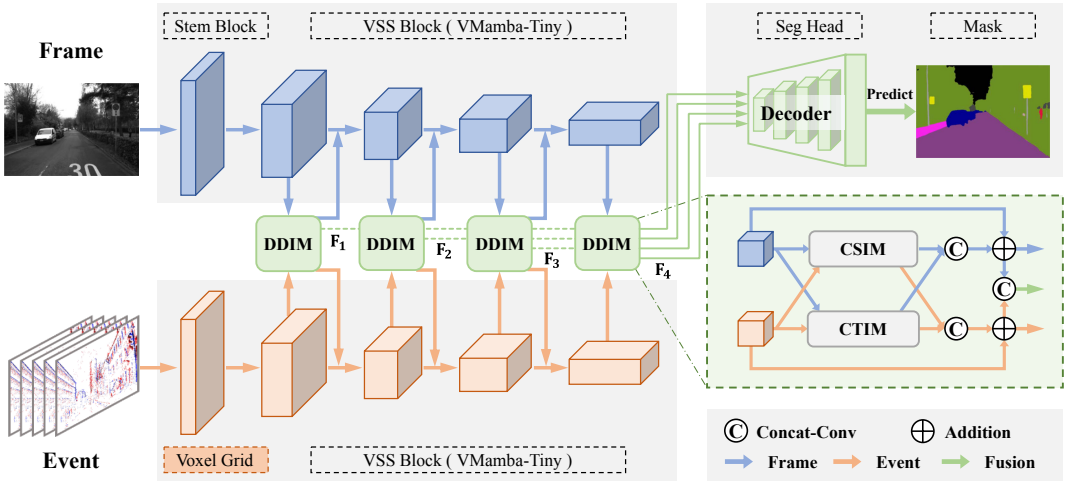

Figure 2: Overview of the MambaSeg framework. MambaSeg consists of a dual-branch Mamba encoder and a Dual-Dimensional Interaction Module (DDIM), which includes CSIM and CTIM. Voxelized event and image streams are independently processed by multi-scale VSS Blocks. DDIM performs spatial-temporal fusion at each scale, and the fused features are iteratively fed into the encoders to enhance cross-modal consistency, followed by a decoder for semantic segmentation.

Proposed Method: MambaSeg

Overview of the MambaSeg

The proposed MambaSeg framework, depicted in Fig. 2, consists of the dual-branch Mamba encoder and the DDIM. The DDIM comprises two integral components: the CSIM and the CTIM, designed to enhance multimodal feature fusion across spatial and temporal dimensions.

MambaSeg takes as input both image frames and asynchronous events. The image frames are represented as $I \in \mathbb{R}^{C\times H\times W}$ as input. Meanwhile, the raw asynchronous event stream is transformed into a structured voxel grid to preserve spatio-temporal information. Each event is represented as a tuple $e_i = (x_i, y_i, t_i, p_i)$, where (x_i, y_i) denotes the spatial location, t_i the timestamp, and $p_i \in \{-1, +1\}$ the polarity. Given a fixed time window, we divide the time span into T discrete bins and accumulate the events into a voxel grid $E \in \mathbb{R}^{T\times H\times W}$. The voxel intensity at each spatio-temporal coordinate is computed as:

$$E(t, x, y) = \sum_{j=1}^{N} \delta(x_j = x, y_j = y) \cdot \delta(t_j \in B_t) \cdot p_j, \quad (1)$$

where B_t denotes the time interval of the t-th temporal bin, and $\delta(\cdot)$ is the Kronecker Delta function. The resulting voxel grid encodes the spatial and temporal distribution of events and serves as the input to the event branch of our model.

The DDIM module is central to our framework. We use four-scale Visual State Space (VSS) Blocks (Liu et al. 2024) to encode image and event modalities independently. At each scale, DDIM facilitates cross-modal interaction and fusion. Its CSIM component leverages image spatial semantics and event edge cues for fine-grained spatial fusion, while CTIM operates along the temporal axis, aligning dynamic events with static visuals to enhance temporal modeling. After each stage, fused features are fed to subsequent encoders, progressively improving cross-modal consistency. Finally, the fused features across all scales are passed through a decoder to generate the final segmentation predictions.

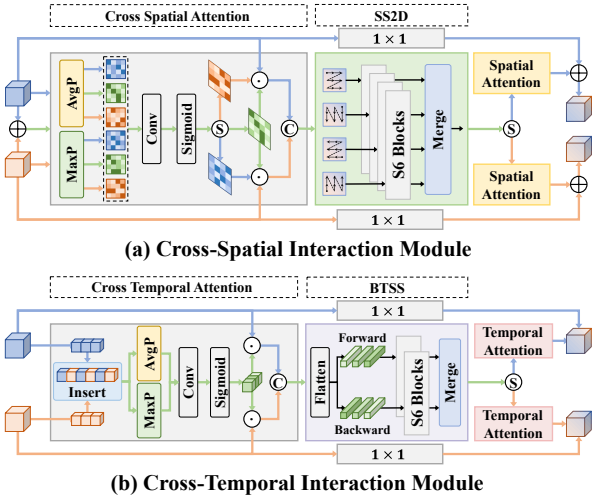

Figure 3: The detail of (a) CSIM and (b) CTIM.

Cross Spatial Interaction Module

To fully exploit the edge cues from event data and the texture-rich information from images while reducing cross-modal ambiguity, we introduce the CSIM, as illustrated in Fig. 3 (a), which achieves efficient cross-modal interaction and fusion through three key components: cross-modal spatial attention, spatial refinement via SS2D, and modality-aware residual updates. Leveraging this design, it effectively integrates complementary features.

Cross-Modal Spatial Attention. To ensure dimensional compatibility for cross-modal fusion, we align the number of image channels with the number of event time steps at each feature extraction stage. Consequently, the event and image features at stage i are represented as $E_i \in \mathbb{R}^{T\times H\times W}$ and $I_i \in \mathbb{R}^{T\times H\times W}$, respectively. We first compute a shallow fusion feature via element-wise addition $F_i^S = E_i + I_i$. To capture complementary spatial cues across modalities, we apply both average and max pooling on each of the features E_i, I_i, and F_i, yielding a total of six spatial feature maps:

$$\begin{aligned} X_i = \big[&\mathrm{AvgPool}(E_i),\ \mathrm{MaxPool}(E_i),\\ &\mathrm{AvgPool}(I_i),\ \mathrm{MaxPool}(I_i),\\ &\mathrm{AvgPool}(F_i),\ \mathrm{MaxPool}(F_i)\big] \in \mathbb{R}^{6\times H\times W}. \end{aligned} \quad (2)$$

These spatial maps are concatenated along the channel dimension and passed through two convolutional layers and a sigmoid activation to generate the spatial attention weight:

$$W^S = \sigma(\mathrm{Conv}_2(\mathrm{ReLU}(\mathrm{Conv}_1(X_i)))) \in \mathbb{R}^{3\times H\times W}, \quad (3)$$

where σ is the sigmoid function. We divide W^S into three spatial attention maps, denoted as W_E^S, W_I^S, and W_F^S, corresponding to E_i, I_i, and F_i, respectively. Cross-modal interaction is then achieved by applying these attention maps. Specifically, event-guided attention sharpens geometric edges while image-guided attention enriches texture, facilitating deep interaction between the modalities:

$$\begin{aligned} E_c^S &= E_i \odot W_I^S \odot W_F^S,\\ I_c^S &= I_i \odot W_E^S \odot W_F,\\ F_c^S &= \mathrm{Concat}(E_c^S, I_c^S), \end{aligned} \quad (4)$$

where $\odot$ denotes element-wise multiplication and Concat represents channel-wise concatenation.

Spatial Refinement via SS2D. To further enhance the spatially fused features, we feed F_c^S into the SS2D module:

$$F_s^S = \mathrm{SS2D}(F_c^S). \quad (5)$$

Within SS2D, the feature map is unfolded into four directional sequences based on predefined patch configurations. Each sequence is independently processed by a dedicated S6 Block to capture long-range dependencies along its respective orientation. Resulting outputs are subsequently fused and reshaped into a 2D spatial feature map. This directional sequence modeling enables SS2D to capture diverse contextual cues across multiple spatial orientations effectively.

Modality-Aware Residual Update. The fused feature F_s^S is split back into separate image and event modality features: $\{E_s^S, I_s^S\} = \mathrm{Split}(F_s^S)$. We then apply a spatial attention module $\mathrm{SA}(\cdot)$ to each branch, followed by residual connections. This design allows the updated features to retain modality-specific characteristics while integrating contextually enriched information:

$$\begin{aligned} E_{i+1}^S &= E_i + E_s^S \odot \mathrm{SA}(E_s^S),\\ I_{i+1}^S &= I_i + I_s^S \odot \mathrm{SA}(I_s^S), \end{aligned} \quad (6)$$

where E_{i+1}^S and I_{i+1}^S denote the refined event and image features, which are propagated to the next stage of the network.

Cross Temporal Interaction Module

To exploit the inherent temporal dynamics of event data, we introduce CTIM, as shown in Fig. 3(b). This module enables temporal cross-modality alignment through three components: cross-modal temporal attention, bi-directional temporal selective scanning, and modality-aware residual update. By aligning event features with corresponding image features, CTIM mitigates temporal inconsistencies and enhances complementarity between modalities.

Cross-Modal Temporal Attention. Given the event and image features at stage i, denoted as $E_i \in \mathbb{R}^{T\times H\times W}$ and $I_i \in \mathbb{R}^{T\times H\times W}$, we construct a temporal fusion feature $F_i^T \in \mathbb{R}^{2T\times H\times W}$ by temporally interleaving event and image features:

$$F_i^T = \mathrm{Insert}(E_i, I_i), \quad (7)$$

where event features are inserted between adjacent image channels along the temporal dimension. This design preserves the temporal sequence while explicitly modeling interleaved dependencies between modalities.

We then apply global max pooling and average pooling to the temporally interleaved feature F_i^T to extract temporal response descriptors:

$$F_{\max}^T = \mathrm{MaxPool}(F_i^T), \quad F_{\mathrm{avg}}^T = \mathrm{AvgPool}(F_i^T). \quad (8)$$

These pooled features are projected into temporal attention weights through two consecutive 1×1 convolution layers followed by a sigmoid activation:

$$W_F^T = \sigma\Big(\mathrm{Conv}(F_{\max}^T) + \mathrm{Conv}(F_{\mathrm{avg}}^T)\Big) \in \mathbb{R}^{T\times 1\times 1}. \quad (9)$$

The resulting attention weights W_F^T are broadcast and applied to the original modality inputs, producing temporally modulated representations:

$$E_c^T = E_i \odot W_F^T, \qquad I_c^T = I_i \odot W_F^T. \quad (10)$$

This attention mechanism emphasizes motion-sensitive features in the event stream and suppresses redundant or ambiguous static cues in the image stream, thereby promoting robust and discriminative temporal fusion across modalities.

Bi-Directional Temporal Selective Scan. To further model long-range dependencies, we concatenate the attended features as:

$$F_c^T = \mathrm{Concat}(E_c^T, I_c^T) \in \mathbb{R}^{2T\times H\times W}, \quad (11)$$

and then flatten the spatial to form a temporal sequence $F_{\mathrm{flat}}^T \in \mathbb{R}^{2T\times HW}$. This sequence is processed in both forward and reverse directions using separate S6 blocks:

$$F_{\mathrm{fwd}}^T = \mathrm{S6}(F_{\mathrm{flat}}^T), \quad F_{\mathrm{bwd}}^T = \mathrm{S6}(\mathrm{Reverse}(F_{\mathrm{flat}}^T)). \quad (12)$$

The outputs are summed and reshaped back to a 3D form to produce the temporally enriched feature map:

$$F_b^T = \mathrm{Reshape}(F_{\mathrm{fwd}}^T + F_{\mathrm{bwd}}^T). \quad (13)$$

This bidirectional design enables the aggregation of temporal context from both past and future frames, resulting in more coherent and contextually aligned representations.

Modality-Aware Residual Update. Similar to CSIM, we split the fused features back into modality-specific branches, i.e., $\{E_b^T, I_b^T\} = \text{Split}(F_b^T)$. Each branch is refined using a temporal attention module $\text{TA}(\cdot)$, followed by a residual connection with the original input:

$$E_{i+1}^T = E_i + E_b^T \odot \text{TA}(E_b^T),$$
$$I_{i+1}^T = I_i + I_b^T \odot \text{TA}(I_b^T), \quad (14)$$

where E_{i+1}^T, I_{i+1}^T are the updated event and image features passed to the next stage. This update preserves the semantic identity of each modality while integrating complementary temporal cues, contributing to a progressively aligned and semantically enriched cross-modal representation.

Experiments and Results

Datasets and Evaluation Metrics

Experiments are conducted on two widely used event-based semantic segmentation datasets: DDD17 and DSEC-Semantic (DSEC). The DDD17 dataset (Binas et al. 2017; Alonso and Murillo 2019) contains paired event data and grayscale images captured from driving scenes using DAVIS sensors at a resolution of 346×260. Semantic annotations are generated using a pretrained segmentation model on synchronized images, covering six categories. The dataset comprises 15,950 training pairs and 3,890 testing pairs.

The DSEC dataset (Sun et al. 2022) includes driving sequences with event streams and high-resolution RGB images (440×640), annotated with 11 fine-grained semantic categories. We follow the official data split, which consists of 8,082 training frames across eight sequences and 2,809 testing frames across three sequences, as well as the original preprocessing pipeline.

We evaluate segmentation performance using mean Intersection over Union (mIoU) and pixel accuracy. Model complexity is reported in terms of the number of trainable parameters and multiply-accumulate operations (MACs).

Implementation Details

We employ the VMamba-T model (Liu et al. 2024), pretrained on ImageNet-1K, as the encoder for both the event and image branches. Each branch utilizes a four-stage encoder. For the segmentation head, we adopt the MLP decoder architecture from SegFormer (Xie et al. 2021).

All models are implemented in PyTorch and trained on a single NVIDIA RTX-4090D GPU using the AdamW optimizer and cross-entropy loss. Training is conducted for 60 epochs on both datasets. On DDD17, we use an initial learning rate (*lr*) of 2e-4 with a batch size (*bs*) of 12; on DSEC, the *lr* is set to 6e-5, with a *bs* of 4. Following prior work (Xie et al. 2024), we construct voxel grids by segmenting the event stream into 10 intervals. For DDD17, segmentation is based on fixed 50 ms intervals, while for DSEC, each segment contains 100,000 events. To ensure fair comparison, we apply standard data augmentations, including random cropping, horizontal flipping, and random resizing.

Results and Analysis

We benchmark MambaSeg against SOTA segmentation methods, categorizing by input modality into: image-only, event-only, and event-image fusion. Image-only methods include SegFormer (Xie et al. 2021) and SegNeXt (Guo et al. 2022). Event-only methods comprise EV-SegNet (Alonso and Murillo 2019) and ESS (Sun et al. 2022). Event-image fusion methods include EDCNet (Zhang, Yang, and Stiefelhagen 2021), HALSIE (Das Biswas et al. 2024),HybridSeg (Li et al. 2025), CMX (Zhang et al. 2023a), CMNeXt (Zhang et al. 2023b), SE-Adapter (Yao et al. 2024), and EISNet (Xie et al. 2024).

Quantitative Evaluation. Table 1 illustrates that MambaSeg outperforms SOTA methods on the DDD17 and DSEC benchmarks. On the DDD17 dataset, MambaSeg achieves a mIoU of 77.56% and an accuracy of 96.33%, surpassing the previous best method, EISNet, by 2.53% in mIoU. Similarly, on the DSEC dataset, MambaSeg achieves the highest mIoU of 75.10% and accuracy of 95.71%, exceeding EISNet by 2.03% in mIoU. These results underscore MambaSeg's superior performance in event-image fusion semantic segmentation. The enhanced performance is attributed to three key innovations: (1) Parallel Mamba encoders with global receptive fields effectively capture rich feature representations from both modalities. (2) The CSIM facilitates fine-grained spatial fusion by leveraging complementary image textures and event edges, improving robustness and spatial consistency. (3) The CTIM enhances temporal coherence in event streams through cross-modal temporal fusion, mitigating cross-modal ambiguity.

Beyond segmentation accuracy, we also evaluate computational efficiency on the DDD17 dataset (Table 2). Compared to Transformer-based fusion methods such as CMX and EISNet, MambaSeg achieves the best mIoU (77.56%) with significantly fewer parameters (25.44M) and moderate MACs (15.59G). Relative to CNN-based methods, MambaSeg delivers substantially higher accuracy with comparable or lower computational cost. These results highlight the efficiency of the Mamba architecture and demonstrate that MambaSeg achieves an excellent balance between performance and efficiency.

Qualitative Evaluation. As illustrated in Fig. 4, we present qualitative segmentation results on the DDD17 and DSEC datasets, comparing MambaSeg with ESS (event-only), SegFormer (image-only), and the SOTA fusion method EISNet. Due to the inherent sparsity of event data, event-only methods often fail to recover complete semantic regions. In contrast, image-only methods such as SegFormer struggle with small object segmentation (e.g., pedestrians) under challenging lighting or cluttered backgrounds, as they are less sensitive to dynamic changes.

Compared to EISNet, MambaSeg produces more accurate segmentation of small and complex objects, such as pedestrians and traffic signs. This highlights the effectiveness of our dual-dimensional fusion strategy, which combines the temporal dynamics of event data with the rich texture information of images along both spatial and temporal axes.

Method	Publication	Modality	Backbone	Representation	DDD17		DSEC	
					mIoU (%)	Acc. (%)	mIoU (%)	Acc. (%)
SegNeXt	NeurIPS'21	Image	CNN	-	71.46	95.97	71.55	94.89
SegFormer	NeurIPS'22	Image	Transformer	-	71.05	95.73	71.99	94.97
EV-SegNet	CVPR'19	Event	CNN	6-Channel	54.81	89.76	51.76	88.61
ESS	ECCV'22	Event	CNN	Voxel Grid	61.37	91.08	51.57	89.25
ESS	ECCV'22	Image-Event	CNN	Voxel Grid	60.43	90.37	53.29	89.37
EDCNet	TITS'22	Image-Event	CNN	Voxel Grid	61.99	93.80	56.75	92.39
HALSIE	WACV'24	Image-Event	CNN+SNN	Voxel Grid	60.66	92.50	52.43	89.01
Hybrid-Seg	AAAI'25	Image-Event	CNN+SNN	Voxel Grid	67.31	95.07	66.57	94.27
CMX	TITS'23	Image-Event	Transformer	Voxel Grid	71.88	95.64	72.42	95.07
CMNeXt	CVPR'23	Image-Event	Transformer	Voxel Grid	72.67	95.74	72.54	95.10
SE-Adapter	ICRA'24	Image-Event	Transformer	MSP	69.06	95.32	69.77	93.58
EISNet	TMM'24	Image-Event	Transformer	AET	75.03	96.04	73.07	95.12
MambaSeg	Ours	Image-Event	Mamba	Voxel Grid	**77.56**	**96.33**	**75.10**	**95.71**

Table 1: Comparison with state-of-the-art semantic segmentation methods on DDD17 and DSEC datasets.

Method	Backbone	Params (M)	MACs (G)	mIoU (%)
Ev-SegNet	CNN	29.09	73.62	54.81
EDCNet	CNN	**23.06**	**6.14**	61.99
SegFormer	Transformer	51.54	19.78	73.25
SwiftFromer	Transformer	64.61	23.88	74.31
CMX	Transformer	66.56	16.29	71.88
CMNeXt	Transformer	58.68	16.32	72.67
EISNet	Transformer	34.39	17.30	75.03
MambaSeg (Ours)	Mamba	25.44	15.59	**77.56**

Table 2: Model complexity on DDD17 dataset.

Ablation Study

To assess the contribution of each component, we conduct ablation studies on the DDD17 dataset, focusing on the proposed CSIM, CTIM, and their respective sub-components.

Comparison of Cross-Modal Fusion Methods. We assess the effectiveness of DDIM by comparing it with representative fusion strategies under a unified setup. All methods adopt the same VMamba-T encoder and are applied after feature encoding, with the rest of the architecture unchanged. We use element-wise addition as the baseline and compare with FFM (Zhang et al. 2023a), MRFM (Xie et al. 2024), CSF (Li et al. 2025), and our DDIM. As shown in Table 3, DDIM outperforms all competitors, achieving 77.56% mIoU and 96.33% pixel accuracy on DDD17. This improvement stems from DDIM's dual-axis fusion, which effectively aligns spatial and temporal features to enhance cross-modal complementarity. In contrast, element-wise addition lacks any interaction modeling (74.38% mIoU), while FFM, MRFM, and CSF offer limited fusion capabilities, lagging behind DDIM by 1.12%, 1.37%, and 0.91% mIoU, respectively. By aligning modalities along both spatial and temporal dimensions, DDIM enables more effective feature integration and leads to superior segmentation performance.

Fusion Method	mIoU (%)	Acc. (%)
Baseline	74.38	95.96
FFM (Zhang et al. 2023a)	76.44	96.06
MRFM (Xie et al. 2024)	76.19	95.97
CSF (Li et al. 2025)	76.65	96.22
DDIM (Ours)	**77.56**	**96.33**

Table 3: Comparison of different cross-modal fusion methods on DDD17 dataset.

CSIM	CTIM	mIoU (%)	Acc. (%)
✗	✗	74.38	95.96
✗	✓	76.20	96.06
✓	✗	76.32	96.25
✓	✓	**77.56**	**96.33**

Table 4: Effectiveness of CSIM and CTIM on DDD17. Removing either CSIM or CTIM leads to a noticeable drop in performance, confirming their complementary roles in enhancing segmentation accuracy.

Effect of CSIM and CTIM. We first evaluate the individual and combined impact of CSIM and CTIM. As shown in Table 4, removing either module results in a performance drop, while combining both yields the highest mIoU of 77.56% and accuracy of 96.33%. These results underscore the complementary strengths of spatial and temporal fusion in improving segmentation performance.

Ablation on CSIM Components. We further break down CSIM into three key components: Cross Spatial Attention (CSA), 2D Selective Scan (SS2D), and Spatial Attention (SA). As shown in Table 5, each component contributes to the overall performance, with the full CSIM achieving the highest mIoU and accuracy. Notably, the combination of CSA and SA yields substantial improvements, highlighting their synergistic effect in capturing both fine-grained spatial structure and broader contextual semantics.

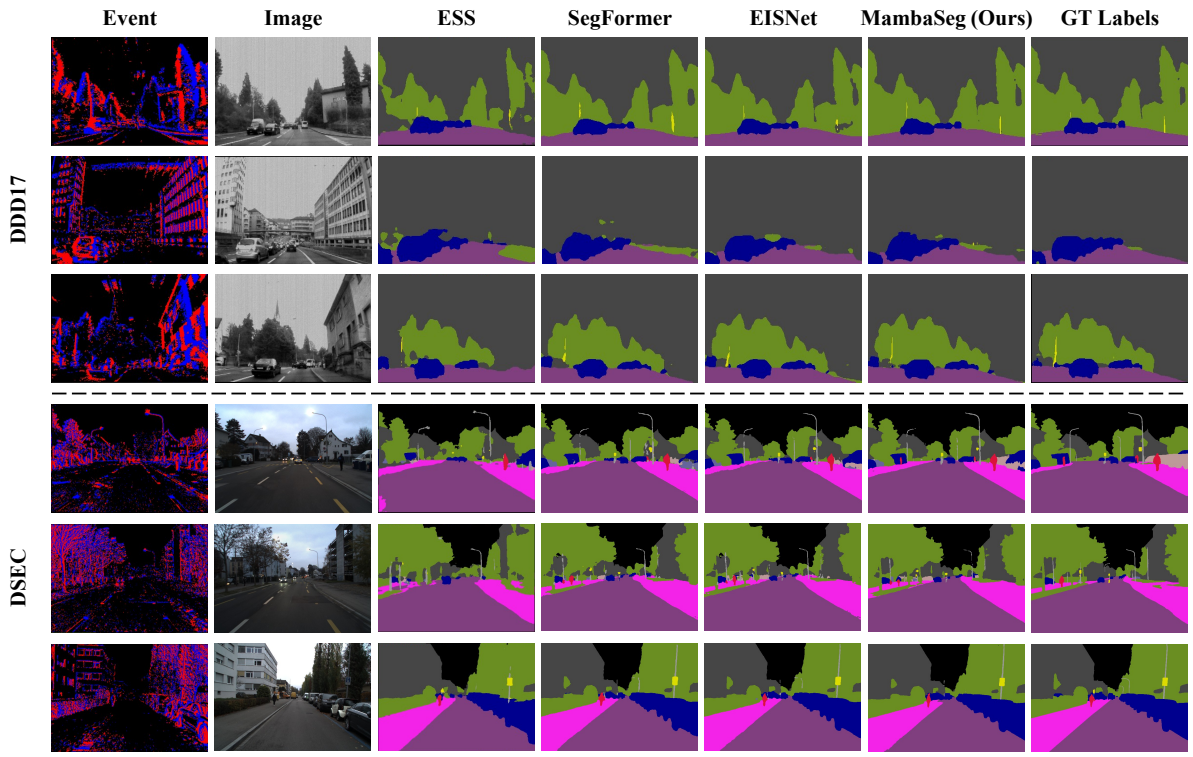

Figure 4: Qualitative comparison of different segmentation methods on DDD17 and DSEC datasets.

Variation	CSA	SS2D	SA	mIoU (%)	Acc. (%)
CSA + SS2D	✓	✓	✗	76.47	96.24
SS2D + SA	✗	✓	✓	76.59	96.26
CSA + SA	✓	✗	✓	76.71	96.24
Full CSIM	✓	✓	✓	**77.56**	**96.33**

Table 5: Component-wise ablation of CSIM. All components contribute positively to performance, with the complete CSIM achieving the highest accuracy and mIoU.

Variation	CTA	BTSS	TA	mIoU (%)	Acc. (%)
CTA + BTSS	✓	✓	✗	76.49	96.15
BTSS + TA	✗	✓	✓	76.66	96.17
CTA + TA	✓	✗	✓	76.53	96.14
Full CTIM	✓	✓	✓	**77.56**	**96.33**

Table 6: Component-wise ablation of CTIM. All temporal fusion components contribute to performance gains, with the full CTIM achieving the best results.

Ablation on CTIM Components. We then conduct an ablation study on CTIM, which comprises Cross Temporal Attention (CTA), Bi-Directional Temporal Selective Scan (BTSS), and Temporal Attention (TA). As shown in Table 6, each component contributes meaningfully to performance. The complete CTIM configuration achieves the best results, demonstrating that jointly modeling temporal dynamics and alignment enhances the effectiveness of temporal fusion.

Conclusion

In this work, we proposed MambaSeg, a novel dual-branch framework for multimodal semantic segmentation, built upon the efficient and scalable Mamba architecture. To address the high computational demands of Transformer-based models and the difficulty of fusing RGB images with sparse event data, we introduced the DDIM, which incorporates both CSIM and CTIM components. DDIM enables fine-grained, complementary fusion by jointly aligning spatial details and temporal dynamics across modalities. Extensive evaluations on the DDD17 and DSEC datasets show that MambaSeg achieves SOTA performance while significantly reducing model complexity, offering an excellent trade-off between segmentation accuracy and efficiency.

Our future work aims to deploy MambaSeg on real-world robotic platforms to validate its practical effectiveness in resource-constrained, dynamic environments.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China (No. 42474027, 62403085, 42174050, 62322601), China Postdoctoral Science Foundation (No. 2023M740402), and Fundamental Research Funds for the Central Universities (No. 2024IAIS-QN017, 2025CDJZDGF001).

References

Alonso, I.; and Murillo, A. C. 2019. EV-SegNet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12): 2481–2495.

Binas, J.; Neil, D.; Liu, S.-C.; and Delbruck, T. 2017. DDD17: End-to-end DAVIS driving dataset. *arXiv preprint arXiv:1711.01458*.

Das Biswas, S.; Kosta, A.; Liyanagedera, C.; Apolinario, M.; and Roy, K. 2024. HALSIE: Hybrid approach to learning segmentation by simultaneously exploiting image and event modalities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5964–5974.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 154–180.

Gehrig, D.; and Scaramuzza, D. 2024. Low-latency automotive vision with event cameras. *Nature*, 629(8014): 1034–1040.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; and Xia, S.-T. 2024. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, 222–241. Springer.

Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.; Cheng, M.-M.; and Hu, S.-M. 2022. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in neural information processing systems*, 35: 1140–1156.

Jia, Z.; You, K.; He, W.; Tian, Y.; Feng, Y.; Wang, Y.; Jia, X.; Lou, Y.; Zhang, J.; Li, G.; et al. 2023. Event-based semantic segmentation with posterior attention. *IEEE Transactions on Image Processing*, 32: 1829–1842.

Lateef, F.; and Ruichek, Y. 2019. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338: 321–348.

Li, H.; Peng, Y.; Yuan, J.; Wu, P.; Wang, J.; Zhang, Y.; and Sun, X. 2025. Efficient Event-Based Semantic Segmentation via Exploiting Frame-Event Fusion: A Hybrid Neural Network Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18296–18304.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Long, X.; Zhu, X.; Guo, F.; Chen, C.; Zhu, X.; Gu, F.; Yuan, S.; and Zhang, C. 2024. Spike-BRGNet: Efficient and accurate event-based semantic segmentation with boundary region-guided spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*.

Long, X.; Zhu, X.; Guo, F.; Zhang, W.; Gu, Q.; Chen, C.; and Gu, F. 2025. SLTNet: Efficient Event-based Semantic Segmentation with Spike-driven Lightweight Transformer-based Networks. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4331–4338.

Ma, X.; Ni, Z.; and Chen, X. 2024a. Ssa-seg: Semantic and spatial adaptive pixel-level classifier for semantic segmentation. *arXiv preprint arXiv:2405.06525*.

Ma, X.; Ni, Z.; and Chen, X. 2024b. Tinyvim: Frequency decoupling for tiny hybrid vision mamba. *arXiv preprint arXiv:2411.17473*.

Ni, Z.; Chen, X.; Zhai, Y.; Tang, Y.; and Wang, Y. 2024. Context-guided spatial feature reconstruction for efficient semantic segmentation. In *European Conference on Computer Vision*, 239–255. Springer.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6): 1964–1980.

Ruan, J.; Li, J.; and Xiang, S. 2024. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*.

Smith, J. T.; Warrington, A.; and Linderman, S. W. 2022. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*.

Sun, Z.; Messikommer, N.; Gehrig, D.; and Scaramuzza, D. 2022. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, 341–357. Springer.

Wang, Z.; Li, C.; Xu, H.; and Zhu, X. 2024. Mamba YOLO: SSMs-Based YOLO For Object Detection. arXiv:2406.05835.

Xie, B.; Deng, Y.; Shao, Z.; and Li, Y. 2024. Eisnet: A multi-modal fusion network for semantic segmentation with events and images. *IEEE Transactions on Multimedia*, 26: 8639–8650.

Xie, C.; Gao, W.; and Guo, R. 2024. Cross-modal learning for event-based semantic segmentation via attention soft alignment. *IEEE Robotics and Automation Letters*, 9(3): 2359–2366.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Yao, B.; Deng, Y.; Liu, Y.; Chen, H.; Li, Y.; and Yang, Z. 2024. Sam-event-adapter: Adapting segment anything model for event-rgb semantic segmentation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 9093–9100. IEEE.

Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; and Stiefelhagen, R. 2023a. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12): 14679–14694.

Zhang, J.; Liu, R.; Shi, H.; Yang, K.; Reiß, S.; Peng, K.; Fu, H.; Wang, K.; and Stiefelhagen, R. 2023b. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1136–1147.

Zhang, J.; Yang, K.; and Stiefelhagen, R. 2021. Exploring event-driven dynamic context for accident scene segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 23(3): 2606–2622.

Zhang, R.; Leng, L.; Che, K.; Zhang, H.; Cheng, J.; Guo, Q.; Liao, J.; and Cheng, R. 2024. Accurate and efficient event-based semantic segmentation using adaptive spiking encoder–decoder network. *IEEE Transactions on Neural Networks and Learning Systems*.