CONTINUOUS-TIME VOXEL-FREE EVENT–RGB SEMANTIC SEGMENTATION WITH FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic segmentation for autonomous driving and robotics remains challenging under fast motion, low illumination, and high dynamic range, where conventional RGB sensing becomes unreliable. Event cameras provide complementary measurements with high temporal resolution, yet most existing RGB-event segmentation methods first quantize asynchronous events into voxel grids, discarding fine-grained temporal structure and potentially introducing cross-modal misalignment. We present a voxel-free RGB-event segmentation framework that preserves event timing throughout the representation and fusion process. At its core is CT-EvSSM, a continuous-time event encoder that operates directly on raw events using per-event temporal embeddings and a selective state-space model whose discretization is conditioned on the true inter-event intervals. To integrate event features with RGB representations, we adapt a pretrained DINOv2-L/14 backbone into a multi-scale feature pyramid and introduce a dual-branch fusion mechanism that combines geometric cross-modal alignment with frequency-domain feature fusion and temporal state-space modeling. This design enables event information to be aligned and integrated with RGB features while retaining its asynchronous temporal structure. Extensive experiments on two widely used datasets, DSEC and DDD17, demonstrate the effectiveness of the proposed method compared to several state-of-the-art approaches.

1 INTRODUCTION

Semantic segmentation from event cameras promises low-latency, high-dynamic-range scene understanding for robotics and autonomous driving, but a single event stream alone under-constrains dense per-pixel prediction, motivating RGB–event fusion architectures. Event cameras asynchronously measure per-pixel brightness changes and output a continuous stream of events (x, y, t, p), where (x, y) are pixel coordinates, t is a timestamp and p is polarity. Event cameras provide microsecond temporal resolution, extremely low latency, and virtually no motion blur, preserving motion cues that conventional frame-based RGB cameras discard between exposures – precisely the conditions (fast motion, low light, high dynamic range) under which RGB-only segmentation degrades.

Despite encouraging progress, existing RGB–event semantic segmentation methods share a common design choice: before feature extraction, the asynchronous event stream is first converted into a synchronized, frame-like representation such as a voxel-grid histogram (Zhu et al., 2019), and only then encoded—often by a hierarchical backbone—and fused with RGB features at each stage of a shared decoder (Zhang et al., 2023a;b; Xie et al., 2024; Gu et al., 2026). This preprocessing step enables the reuse of conventional convolutional or state-space architectures, but it fundamentally changes the nature of the input: events falling within the same temporal bin become indistinguishable, and the bin width becomes an additional hyperparameter that trades temporal precision for computational convenience. In dense RGB–event prediction pipelines, the continuous temporal structure of the event stream is therefore discarded before the network starts learning.

The central limitation of existing RGB–event segmentation pipelines lies in how events are represented before learning begins. Although event cameras provide measurements at microsecond resolution, most methods first compress the asynchronous stream into a fixed number of temporal bins and then process the resulting voxel grid with a conventional vision backbone. Once this conversion is performed, the exact ordering and inter-event timing within each bin are irreversibly lost.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Consequently, downstream encoders and fusion modules can only reason over a temporally quantized surrogate of the original event stream, rather than the continuous event dynamics recorded by the sensor. The key question we therefore ask is: can RGB–event semantic segmentation operate on raw asynchronous events without sacrificing the efficiency and multi-scale structure required for dense prediction?

Addressing this question introduces two practical challenges. First, raw events form a long, irregularly sampled sequence, whereas semantic segmentation requires dense spatial feature maps at multiple resolutions. We therefore introduce CT-EvSSM, a continuous-time state-space encoder that processes events in temporal order and conditions its state evolution on the true inter-event intervals. Spatial discretization is deferred until temporal modeling is complete, after which the learned event features are projected onto a 2D grid only once. This preserves continuous-time event information while retaining compatibility with hierarchical dense prediction.

Second, once the event branch is no longer tied to a conventional voxel-grid backbone, the RGB and fusion components must be redesigned to make effective use of the resulting representation. We adapt a pretrained DINOv2-L/14 backbone into a multi-scale feature pyramid for the RGB stream and explicitly correct residual spatial misalignment between RGB and event features before fusion. Our fusion module combines geometric alignment with spectral and state-space interactions, enabling the two modalities to exchange complementary spatial and temporal information across scales.

Together, these designs form a voxel-free RGB–event segmentation framework in which event timing is preserved until after temporal feature extraction rather than discarded during preprocessing. We evaluate the framework on DSEC-Semantic and DDD17 and conduct controlled ablations on event representation, RGB backbone adaptation, fusion, and computational efficiency.

Our main contributions are threefold: (1) We formulate RGB–event semantic segmentation without temporal voxelization and introduce CT-EvSSM, which directly models raw asynchronous events using their true inter-event timing before any spatial rasterization. (2) We adapt a pretrained DINOv2 vision transformer to multi-scale RGB–event segmentation through a lightweight hierarchical feature interface. (3) We introduce a cross-modal fusion design that explicitly aligns RGB and event features before multi-scale interaction, and validate the full framework on DSEC-Semantic and DDD17.

2 RELATED WORKS

The dominant strategy for feeding an asynchronous event stream into a standard CNN or ViT backbone is rasterization into a fixed-size tensor—most commonly a voxel grid (Zhu et al., 2019), but also time surfaces, event histograms, and multi-channel encodings—at the cost of temporal binning: events within a bin become indistinguishable, and bin width becomes a hyperparameter trading temporal precision for computational convenience. Point-based approaches instead process raw events as a sequence, point set, or evolving graph (Ren et al., 2024; Schaefer et al., 2022), preserving native temporal resolution without quantization, and EventMamba (Ge et al., 2025) applies a state-space model to point-cloud event representations. These methods have been developed largely for classification and regression, however, where the absence of a dense multi-scale feature requirement makes their computational profile more favourable than it is for segmentation. CT-EvSSM occupies the space between these regimes: it consumes every raw event once with its true per-event timing, as point-based methods do, but defers spatial discretization to a single scatter-to-grid operation after temporal modeling is complete, so the output remains compatible with the hierarchical, per-stage fusion that dense prediction requires.

EV-SegNet (Alonso & Murillo, 2019) established event-only segmentation on DDD17, and subsequent work improved accuracy through synthetic event generation, domain adaptation, and event-specific attention. RGB–event fusion has since proven more robust, exploiting the complementarity of the two modalities: RGB provides dense semantic appearance but degrades under motion blur and poor illumination, while events capture motion boundaries at high temporal precision but lack dense appearance. EISNet (Xie et al., 2024) introduces asynchronous event–image interaction modules that progressively exchange information between branches; MambaSeg (Gu et al., 2026) fuses per-stage VMamba (Liu et al., 2024) features through a Dual-Dimensional Interaction Module; and

more general multimodal frameworks such as CMX (Zhang et al., 2023a) and CMNeXt (Zhang et al., 2023b) apply feature rectification and cross-modal attention across RGB, depth, thermal, and event inputs. Regardless of fusion sophistication, all operate on a voxelized event representation whose temporal information has already been compressed before multimodal learning begins, bounding fusion quality by the input representation. Most closely related to our spatial fusion branch is Evita (Peng et al., 2026), which embeds deformable cross-modal alignment and amplitude/phase spectral fusion inside a unified backbone aligned RGB–event pairs. We adapt these two mechanisms, denoted CMGA and SFF, into our hierarchical fusion block.

Mamba (Gu & Dao, 2023) introduced input-dependent state-space scans as a linear-time alternative to self-attention, with its practical speed owing to a fused CUDA kernel that keeps the scan's intermediate state in on-chip memory; VMamba (Liu et al., 2024) adapts this to a hierarchical vision backbone via cross-scan. We reuse the same kernel family for our event-branch recurrence, with the discretization step driven by the measured inter-event gap rather than by a learned but time-agnostic step. Separately, DINOv2 (Oquab et al., 2023) shows that self-supervised ViT features transfer strongly to dense prediction, typically through a lightweight head on frozen features or through adapters that supply the multi-scale structure a flat ViT lacks (Chen et al., 2022; Cheng et al., 2022). Our adapter follows this line; we additionally evaluate the fine-tuned setting, which is less commonly reported in this literature and which we find outperforms freezing once the downstream features are cross-modal.

3 Problem Analysis

Task Formulation Given a synchronized RGB frame $I \in \mathbb{R}^{3\times H\times W}$ and an event stream $\mathcal{E}$ captured over the corresponding exposure interval, RGB–Event semantic segmentation seeks a mapping $f_\theta(I, \mathcal{E}) \mapsto \hat{Y} \in \{1, \dots, C\}^{H\times W}$ assigning one of C classes to every pixel. The task is typically framed as a dual-branch encoder–decoder problem: an RGB encoder Φ_{rgb} and an event encoder Φ_{evt} each produce multi-scale feature maps $\{F^{(i)}_{\text{rgb}}\}_{i=1}^{L}$ and $\{F^{(i)}_{\text{evt}}\}_{i=1}^{L}$ at L decreasing resolutions, which are fused stage-wise and decoded into $\hat{Y}$. Our contributions concern Φ_{evt} (Section 4.1), Φ_{rgb}, and the stage-wise fusion operator (both Section 4.2), in which a spatial pathway built from CMGA and SFF operates alongside a state-space temporal pathway at each resolution, their outputs merged residually before decoding.

3.1 Event Streams and Voxel-Grid Rasterization

An event camera outputs an asynchronous stream $\mathcal{E} = \{e_k\}_{k=1}^{N}$ with $e_k = (x_k, y_k, t_k, p_k)$, so both N and the inter-event intervals $\delta_k = t_k - t_{k-1}$ vary substantially within and across sequences. To make this consumable by a standard backbone, prior work rasterizes $\mathcal{E}$ into a fixed-size tensor; the dominant choice, the voxel grid (Zhu et al., 2019), partitions the exposure window $[t_0, t_N]$ into B bins and accumulates $V \in \mathbb{R}^{B\times H\times W}$ via

$$V[b, y, x] = \sum_{k:\,(x_k,y_k)=(x,y)} p_k \cdot \max\left(0,\ 1 - \left| b - \frac{t_k - t_0}{t_N - t_0}(B-1) \right| \right), \tag{1}$$

so each event contributes to its two temporally nearest bins by linear interpolation, with B fixed a priori (e.g., B=10). Any two events sharing (x, y, b) are then indistinguishable to the encoder, however far apart t_k and $t_{k'}$ actually were. CT-EvSSM (Section 4.1) inverts this rasterize-then-encode pipeline: it consumes $\{e_k\}$ directly in event order and produces a spatial tensor only once temporal modeling is complete.

3.2 Selective State-Space Models

State-space models (SSMs) map a sequence $u_1, \dots, u_N$ to an output through a linear recurrence over a hidden state $h_k \in \mathbb{R}^d$:

$$h_k = \bar{A}_k h_{k-1} + \bar{B}_k u_k, \qquad y_k = C_k h_k + D \odot u_k, \tag{2}$$

where $(\bar{A}_k, \bar{B}_k)$ discretize a continuous-time system (A, B) with step size Δ_k. Under a zero-order hold $\bar{A}_k = \exp(\Delta_k A)$ and $\bar{B}_k = (\Delta_k A)^{-1}(\exp(\Delta_k A) - I)\,\Delta_k B$ (Gu et al., 2021), though in

practice the first-order simplification $\bar{B}_k = \Delta_k B_k$ is used (Gu & Dao, 2023); C_k is applied directly without discretization and D is a learned skip connection.

Mamba (Gu & Dao, 2023) makes this recurrence *selective* by predicting Δ_k, B_k and C_k from u_k itself, giving content-aware state transitions at linear cost. In vision backbones such as VMamba (Liu et al., 2024), u_k is obtained by flattening a feature map along fixed directional scan orders, and Δ_k is therefore predicted from token content alone, agnostic to any physical notion of time—image tokens carry no timestamp.

CT-EvSSM (Section 4.1) repurposes this recurrence for a sequence that does have a physically meaningful time axis. Our single modification is to Eq. 2's step size: we retain the content term and add an explicit dependence on the measured inter-event gap $\delta_k = \max(t_k - t_{k-1}, 0)$, so that discretization tracks the sensor's actual sampling density rather than a bin width or a content-only heuristic. Eq. 2 itself, the zero-order-hold discretization, and the stabilized parameterization of A are unchanged from the standard formulation.

4 METHOD

CT-Net (Continuous-time Net), shown in Figure 1, comprises three components. A DINOv2-L/14 RGB backbone with a four-level pyramid adapter produces the image-side feature hierarchy. A continuous-time event branch, CT-EvSSM, processes raw events directly without voxelization and produces the matching event-side hierarchy. At each of four pyramid scales, a stage-wise fusion module then combines them along two parallel pathways: a spatial pathway in which CMGA geometrically realigns event features to the RGB grid before SFF fuses the two spectrally, and a temporal pathway built on a state-space recurrence. The two outputs are merged residually and the refined pyramid is decoded into the segmentation map. We first describe Setup and Notation, then the event branch, the RGB branch and fusion, and the training objective.

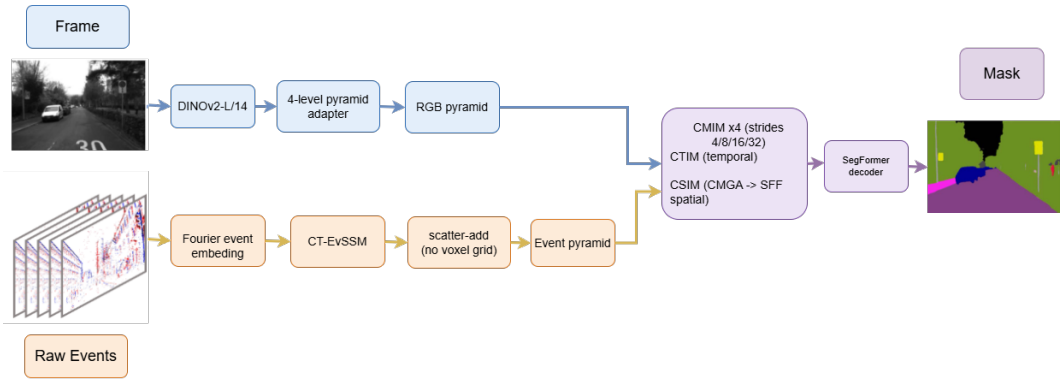


Figure 1: Main architecture of our CT-Net.

The event budget N is fixed per dataset (100,000 for DSEC-Semantic, 4,096 for DDD17) by padding to the cap and carrying a validity mask $m_i \in \{0, 1\}$ for samples with fewer real events. Both branches emit a four-level pyramid $\{F_s^{\text{img}}, F_s^{\text{ev}}\}_{s=1}^{4}$ at channel widths $C_s \in \{96, 192, 384, 768\}$ and strides $\{4, 8, 16, 32\}$, fused stage-by-stage (Section 4.2) and decoded into $\hat{Y}$ by a SegFormer head (Xie et al., 2021). The design constraint distinguishing CT-Net from voxel-based approaches is that temporal binning is never introduced, neither in preprocessing nor inside the event encoder: spatial discretization occurs exactly once, after temporal modeling is complete (Section 4.1).

4.1 CONTINUOUS-TIME EVENT BRANCH: CT-EVSSM

Event Embedding A raw event carries only a coordinate, a timestamp, and a polarity, so the encoder requires a spatial prior before sequence modeling. We encode each event's normalized coordinates $\tilde{x}_i = x_i/W$, $\tilde{y}_i = y_i/H$ with sinusoidal Fourier features over $K{=}8$ bands,

$$\gamma(\tilde{x}_i, \tilde{y}_i) = \left[\sin(2^k \pi \tilde{x}_i),\ \cos(2^k \pi \tilde{x}_i),\ \sin(2^k \pi \tilde{y}_i),\ \cos(2^k \pi \tilde{y}_i)\right]_{k=0}^{K-1} \in \mathbb{R}^{4K}, \tag{3}$$

and concatenate the result with polarity p_i and normalized timestamp t_i. A two-layer MLP with GELU activation projects this 34 dimensional vector to the event embedding $u_i \in \mathbb{R}^{d_{\text{ev}}}$, d_{ev}=64.

Continuous-time Selective Recurrence The time-sorted embeddings $U = (u_1, \ldots, u_N)$ are processed by a selective state-space recurrence driven by the physical time elapsed between consecutive events, $\delta_i = \max(t_i - t_{i-1}, 0)$ with $\delta_1 = 0$, where the clamp guards against non-monotonic timestamps. The hidden state evolves as

$$h_i = \exp(\Delta_i A) \odot h_{i-1} + \Delta_i B_i \odot u_i, \qquad z_i = C_i^\top h_i + D \odot u_i, \tag{4}$$

with $h_i \in \mathbb{R}^{d_{\text{ev}} \times n}$, $n = d_{\text{state}}$, and D a learned skip connection. The projections are input-dependent, $[B_i \,\|\, C_i] = W_{BC}\, u_i$, and the state matrix is parameterized as $A = -\exp(A_{\log})$ so that the dynamics decay rather than diverge over an N-event sequence.

Our modification to the standard recurrence is confined to the discretization step, which we learn jointly from the event's content and its measured time gap:

$$\Delta_i = \text{softplus}(W_\Delta u_i + W_t \delta_i)\,, \tag{5}$$

with W_t initialized to zero. At initialization Δ_i therefore depends on content alone, exactly recovering a standard time-agnostic selective scan; training then learns how strongly the physical gap should modulate the state update without destabilizing early optimization. We apply the same zero-initialization discipline to the gating mechanism introduced in Section 4.2. Because Δ_i is set per event rather than per bin, two events at the same pixel separated by different intervals produce different state updates, a distinction no fixed binning scheme can represent.

Algorithm 1: CT-EvSSM forward pass

Input: Time-sorted embeddings $U \in \mathbb{R}^{B \times N \times d_{\text{ev}}}$, timestamps $t \in \mathbb{R}^{B \times N}$
Output: Output features $Z \in \mathbb{R}^{B \times N \times d_{\text{ev}}}$

$\delta_{:,1} \leftarrow 0$; $\delta_{:,i} \leftarrow \max(t_{:,i} - t_{:,i-1}, 0)$ for $i = 2, \ldots, N$;
$[B \,\|\, C] \leftarrow \texttt{x_proj}(U)$;
$\Delta \leftarrow \text{softplus}(\texttt{dt_proj}(U) + \texttt{time}(\delta))$;
$A \leftarrow -\exp(A_{\log})$; $h \leftarrow 0$;
for $i = 1$ **to** N **do**
  $h \leftarrow \exp(\Delta_i A) \odot h + \Delta_i B_i \odot U_i$;
  $Z_i \leftarrow C_i^\top h + D \odot U_i$;
return Z;

Algorithm 1 states the recurrence in its clearest form, but a literal sequential implementation is slower than a single dense convolution over a voxel grid. We therefore execute the identical recurrence through a fused CUDA kernel that keeps the intermediate state in on-chip memory. Since the kernel initializes from a zero state, we preserve exact state-carrying across chunks with a closed-form correction that exploits the recurrence's affine structure in h, computed with cumulative products and a single einsum rather than a loop.

Voxel-free Spatial Scatter After the scan each event carries a contextual feature z_i but no spatial structure. We scatter these onto a 2D grid exactly once, using a masked scatter-mean over the events falling in each cell:

$$G[y_s, x_s] = \frac{\sum_{i \in \mathcal{S}(x_s, y_s)} z_i\, m_i}{\sum_{i \in \mathcal{S}(x_s, y_s)} m_i + \epsilon}, \quad \mathcal{S}(x_s, y_s) = \Big\{ i : \big(\lfloor x_i W_s / W \rfloor, \lfloor y_i H_s / H \rfloor\big) = (x_s, y_s) \Big\} \tag{6}$$

where m_i is the validity mask of Section 5 and ϵ guards empty cells. This is a single scatter-add kernel producing $G \in \mathbb{R}^{d_{\text{ev}} \times H_s \times W_s}$, which a pretrained VMamba encoder (Liu et al., 2024) then processes into the event pyramid. Because only the input representation has changed, the event branch retains the pretrained low-level features the voxel-grid baseline relies on.

4.2 RGB BRANCH AND EVENT–RGB FUSION

The RGB branch uses DINOv2-L/14 (Oquab et al., 2023), a 24-block Vision Transformer pretrained self-supervised at a scale no segmentation dataset approaches. Unlike VMamba it produces no

hierarchical pyramid: every block operates at the same patch resolution with feature dimension 1024. We therefore insert a lightweight adapter that taps hidden states at blocks $\{6, 12, 18, 24\}$ and maps them, in order, to strides $\{4, 8, 16, 32\}$. At each tapped block it drops the class token, reshapes the patch tokens to a $1024 \times h \times w$ map, projects to the target width C_s with a 1×1 convolution, bilinearly resizes, and refines with GroupNorm, GELU, and a 3×3 convolution. Tapping four depths rather than only the final representation is what makes a flat, isotropic ViT usable in multi-scale fusion: earlier blocks retain the spatial detail the finest levels require, later blocks the semantics the coarsest levels need. Because DINOv2 is a single-pass ViT with no per-stage recurrence for a fusion correction to re-enter, all four levels are computed once per forward call and reused as each stage's RGB input. Grayscale DDD17 frames are channel-replicated before DINOv2's own normalization, inside the adapter, and the backbone is fully fine-tuned at a learning rate $10\times$ lower than the rest of the network, which Section 6 shows outperforms freezing it.

Stage-wise Fusion At each level k the RGB feature I_k and event feature E_k share a spatial size and channel width and are fused by a dual-pathway block operating in parallel on the same normalized inputs. The temporal pathway applies a cross-modal channel-attention gate followed by a selective state-space operation along the channel–time axis. The spatial pathway, adapted from Peng et al. (2026), proceeds in two stages that answer two distinct questions: *where* the modalities disagree, and *what* each should contribute. CMGA settles the first, SFF the second, and the two run in sequence because spectral fusion is only meaningful once the streams are spatially registered.

The need for the first stage arises because sensor parallax, asynchronous trigger delay, and the quantization of Eq. 6 leave the event grid slightly misregistered against the RGB grid—a discrepancy that position-wise gating reweights but never corrects. CMGA therefore warps the event features onto the RGB grid before any fusion occurs. A context encoder consumes both modalities and their difference, predicting offset and modulation logits for the nine taps of a 3×3 deformable kernel,

$$(f_x, f_y, f_m) = \mathrm{Conv}\big(\mathrm{Conv}\big([\, I_k;\, E_k;\, I_k - E_k \,]\big)\big), \quad \Delta r = \tau \tanh\big([f_x; f_y]\big),\ m = \sigma(f_m) \tag{7}$$

where $\tanh$ bounds the displacement to $\pm\tau$ pixels, encoding the prior that the misalignment is small. The final convolution is zero-initialized, so $\Delta r = 0$ at step 0 and CMGA begins as the identity warp—the same discipline applied to W_t in Eq. 5. A modulated deformable convolution (Zhu et al., 2020) then resamples the event features, giving $\tilde{E}_k^{\mathrm{al}}$.

With the streams now aligned, SFF decides what each contributes. Absolute intensity and temporal contrast occupy different parts of the spectrum, so adding them in the spatial domain mixes structure with high-frequency change energy. SFF instead separates the two in the Fourier domain, retaining the RGB phase, which carries structure and edges, and admitting the event amplitude, which carries texture and motion energy, through a learned gate. With $\mathcal{F}^{\mathrm{img}} = \mathrm{rFFT2}(I_k)$, $\mathcal{F}^{\mathrm{ev}} = \mathrm{rFFT2}(\tilde{E}_k^{\mathrm{al}})$, amplitudes $A^{\mathrm{img}}, A^{\mathrm{ev}}$ and phase ϕ^{img},

$$g = \sigma\big(\mathrm{MLP}(\mathrm{avgpool}(A^{\mathrm{img}}))\big),\ \hat{A} = A^{\mathrm{img}} + g \odot A^{\mathrm{ev}},\ I_k^{\mathrm{fused}} = I_k + \mathrm{Conv}_{1\times1}\big(\mathrm{irFFT2}(\hat{A}\, e^{\,i\phi^{\mathrm{img}}})\big) \tag{8}$$

Computing g from the RGB amplitude alone lets the network decide how much event energy to admit based on what RGB already resolves: where RGB structure is weak, more event amplitude passes through.

CMGA and SFF together produce the spatial pathway's output, which must now be recombined with the temporal pathway that has been running alongside it. The fused RGB features I_k^{fused} and aligned event features $\tilde{E}_k^{\mathrm{al}}$ are first refined by a 2D state-space operation and a spatial attention gate. The spatial and temporal outputs are then concatenated, projected back to the stage width, and added residually,

$$\tilde{I}_k = I_k + \mathrm{Conv}_{1\times1}\big(\big[\,\text{temporal};\ \text{spatial}\,\big]\big), \tag{9}$$

and symmetrically for the event branch, so that geometric realignment and temporal aggregation refine the same features without either overwriting the other.

4.3 Training objective and complexity

The network is trained with a hybrid loss combining weighted cross-entropy with the Lovász-softmax surrogate,

$$\mathcal{L} = \mathcal{L}_{\mathrm{WCE}} + \lambda\, \mathcal{L}_{\mathrm{Lovasz}}, \qquad \lambda = 0.5, \tag{10}$$

where the first term upweights under-represented classes and the second directly optimizes a surrogate of the evaluation metric. The event branch is $\mathcal{O}(N)$ in the number of events and the fusion and decoding stages are independent of N, so cost scales linearly with the event budget rather than with a bin count.

5 EXPERIMENTS AND RESULTS

We evaluate on two public RGB–Event benchmarks, following the baseline protocol of Gu et al. (2026) throughout (Table 1). DSEC-Semantic (Sun et al., 2022) augments the DSEC driving dataset (Gehrig et al., 2021) with 11-class labels over 8 training and 3 test sequences at 440×640, from which we build fixed-count $100{,}000$-event segments. DDD17-Seg (Alonso & Murillo, 2019) augments the DDD17 dataset (Binas et al., 2017) with 6-class labels over 7 grayscale sequences at 200×346, from which we build a fixed 50 ms window per label, capped and padded to $4{,}096$ events.

Table 1: Dataset configurations and training settings.

Configuration	**DSEC-Semantic**	**DDD17-Seg**
Classes	11	6
Image size	440×640	200×346
Image channels	3 (RGB)	1 (grayscale)
Train / test samples	4,021 / 1,397	15,950 / 3,890
Event window	count-based	50 ms, 4,096-cap
Events / segment	100,000	$\leq$ 4,096
Batch size	4	12
Base LR	$3{\times}10^{-5}$	$2{\times}10^{-4}$
Epochs	60	60

Within each dataset, every configuration shares the settings in Table 1 along with the optimizer (AdamW), schedule, and warmup; only the component under study differs. All runs use a single RTX A6000 (48 GB), and we report mean Intersection-over-Union (mIoU) and pixel accuracy on the held-out test split.

5.1 RESULTS AND ANALYSIS

Quantitative Evaluation. Table 2 compares CT-Net against event-only, image-only, and fusion baselines. CT-Net reaches 79.43% mIoU and 96.50% accuracy on DSEC-Semantic, 4.33 points above the best voxel-based fusion method, MambaSeg Gu et al. (2026), and 78.34% mIoU on DDD17, improving on MambaSeg Gu et al. (2026) by 0.78 and EISNet Xie et al. (2024) by 3.31. It uses no cross-modal pretraining, initializing the RGB branch from ImageNet-only DINOv2 weights.

Three design decisions account for this. CT-EvSSM drives its discretization step from the measured inter-event gap, retaining the sub-bin ordering that voxelization discards. The pyramid adapter lets a self-supervised ViT serve as the RGB encoder in a per-stage fusion pipeline, and fine-tuning it at a $10\times$-reduced learning rate recovers capacity the frozen setting leaves unused once features are cross-modal. Finally, CMGA corrects residual misregistration between the two feature grids before SFF combines them spectrally, preserving RGB phase while admitting event amplitude as motion energy.

Qualitative Evaluation. Figure 2 compares CT-Net with ESS (Sun et al., 2022) (event-only), SegFormer (Xie et al., 2021) (image-only), and the fusion methods EISNet (Xie et al., 2024) and MambaSeg (Gu et al., 2026). Event sparsity leaves event-only methods unable to recover complete semantic regions, while image-only methods lose small objects such as pedestrians under difficult lighting or cluttered backgrounds, being insensitive to dynamic change. CT-Net segments small and complex structures—pedestrians, poles, traffic signage—more accurately than either fusion baseline, reflecting the combined effect of the DINOv2 backbone, the voxel-free CT-EvSSM event branch, and the CMGA+SFF spatial fusion.

Table 2: Comparison with state-of-the-art methods on the DSEC-Semantic and DDD17 datasets.

Method	Modality	Backbone	Representation	DSEC		DDD17	
				mIoU (%)	Acc. (%)	mIoU (%)	Acc. (%)
SegNeXt	Image	CNN	–	71.55	94.89	71.46	95.97
SegFormer	Image	Transformer	–	71.99	94.97	71.05	95.73
EV-SegNet	Event	CNN	6-Channel	51.76	88.61	54.81	89.76
ESS	Event	CNN	Voxel Grid	51.57	89.25	61.37	91.08
ESS	Image-Event	CNN	Voxel Grid	53.29	89.37	60.43	90.37
EDCNet	Image-Event	CNN	Voxel Grid	56.75	92.39	61.99	93.80
HALSIE	Image-Event	CNN+SNN	Voxel Grid	52.43	89.01	60.66	92.50
Hybrid-Seg	Image-Event	CNN+SNN	Voxel Grid	66.57	94.27	67.31	95.07
CMX	Image-Event	Transformer	Voxel Grid	72.42	95.07	71.88	95.64
CMNeXt	Image-Event	Transformer	Voxel Grid	72.54	95.10	72.67	95.74
SE-Adapter	Image-Event	Transformer	MSP	69.77	93.58	69.06	95.32
EISNet	Image-Event	Transformer	AET	73.07	95.12	75.03	96.04
MambaSeg	Image-Event	Mamba	Voxel Grid	75.10	95.71	77.56	**96.33**
CT-Net	Image-Event	DINOv2	CT-EvSSM	**79.43**	**96.50**	**78.34**	96.26

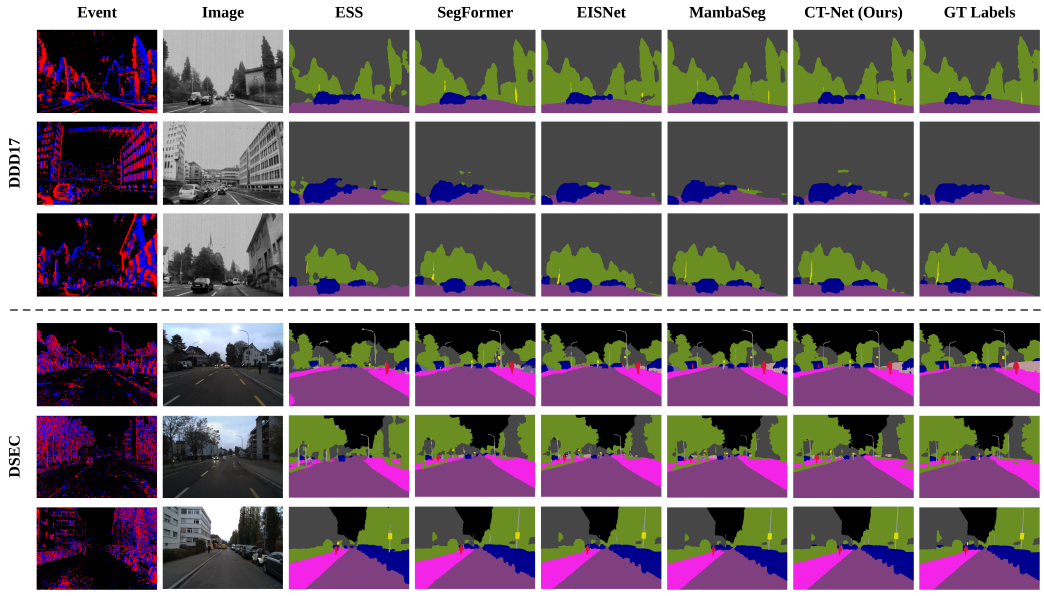


Figure 2: Qualitative comparison on DSEC-Semantic and DDD17.

6 Ablation Study

All configurations share the architecture, optimizer, schedule, augmentation, and splits of Table 1; only the component under study is modified. Table 3 reports best-epoch test mIoU and pixel accuracy.

Table 3: Ablation ladder, in terms of mIoU and pixel accuracy. Row 1 is the published baseline, quoted from its own paper.

#	Configuration	DSEC-Semantic		DDD17-Seg	
		mIoU	Acc.	mIoU	Acc.
1	MambaSeg (published)	75.10	95.71	77.56	**96.33**
2	Frozen DINOv2 + voxel	76.50	95.81	77.32	96.29
3	Fine-tuned DINOv2 + CT-EvSSM	78.79	96.36	77.69	96.14
4	Fine-tuned DINOv2 + CT-EvSSM + CMGA+SFF	**79.43**	**96.50**	**78.34**	96.26

Foundation backbone and pyramid adapter. Replacing the hierarchical RGB encoder with a frozen DINOv2-L/14 behind the four-level adapter (row 2) already exceeds the published baseline by $+1.40$ mIoU on DSEC-Semantic while training only the adapter, fusion, and decoder. Fine-tuning the backbone at a $10\times$-reduced learning rate (row 3) adds a further $+2.29$, and lifts the plateau mean by $+2.71$. This runs counter to the frozen-adapter convention usual for self-supervised ViTs, indicating that cross-modal fused features depart far enough from the RGB pretraining distribution that frozen features leave usable capacity unexploited.

Voxel-free event representation. A matched comparison under the frozen backbone, holding every other component fixed, shows the voxel-free branch on par with voxelization (76.48 vs. 76.50 mIoU, 95.88 vs. 95.81 accuracy on DSEC-Semantic), while removing the bin-count hyperparameter entirely and preserving event ordering through the encoder. Test-time input ablations on the trained model confirm the branch is used: scrambling event coordinates costs 0.41 (DSEC) and 0.65 (DDD17) mIoU, and masking all events costs 0.52 and 1.64.

Spatial fusion. Rows 3–4 are the cleanest isolation in the ladder: both backbones, the temporal pathway, and every hyperparameter are byte-identical, and only the spatial module differs. CMGA+SFF adds $+0.64$ mIoU on DSEC and $+0.65$ on DDD17.

7 Conclusion

We presented CT-Net, a voxel-free framework for RGB–event semantic segmentation that preserves the native temporal structure of asynchronous events instead of quantizing them into predefined temporal bins before feature extraction. To realize this formulation, we introduce a continuous-time event encoder, CT-EvSSM, which models raw event streams using their measured inter-event intervals and postpones spatial discretization until temporal modeling is complete. Combined with a multi-scale RGB representation and cross-modal fusion, the resulting framework enables continuous-time event information to be incorporated into hierarchical dense prediction. Experiments on DSEC-Semantic and DDD17 show consistent improvements in mIoU over prior RGB–event fusion methods, while controlled ablations further demonstrate the contributions of the event representation and fusion design.

More broadly, these results suggest that dense event-based perception can directly model asynchronous event timing without relying on temporal voxelization.

AI Use Statement

In this work, we used generative AI tools to assist with the refinement of methodological descriptions, mathematical notation, and the implementation of the proposed method. We have not used generative AI tools to independently formulate the core research hypothesis, generate or determine the experimental results, or make final methodological decisions, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for literature exploration, manuscript editing, improving language and readability, and assisting with code-related debugging and implementation. All AI-assisted work was reviewed and verified by the authors. AI-assisted code was checked and tested for correctness by the authors, and AI-assisted technical suggestions were independently evaluated before being incorporated into the work. The authors

take responsibility for the final content of this work, including all text, claims, experimental results, methodological decisions, and artifacts produced with the aid of generative AI.

REFERENCES

Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1624–1633. IEEE, 2019.

Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Chengjie Ge, Xueyang Fu, Peng He, Kunyu Wang, Chengzhi Cao, and Zheng-Jun Zha. Event-mamba: Enhancing spatio-temporal locality with state space models for event-based video reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3104–3112, 2025.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021.

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

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

Fuqiang Gu, Yuanke Li, Xianlei Long, Kangping Ji, Chao Chen, Qingyi Gu, and Zhenliang Ni. Mambaseg: harnessing mamba for accurate and efficient image-event semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4302–4310, 2026.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Chenxu Peng, Dicheng Liu, Bo-Wen Yin, Yimian Dai, Xialei Liu, Ming-Ming Cheng, Xiang Li, et al. Weaving light and time: Unified harmonic-geometric representation learning for dense rgb-event parsing. *arXiv preprint arXiv:2607.09143*, 2026.

Hongwei Ren, Yue Zhou, Jiadong Zhu, Haotian Fu, Yulong Huang, Xiaopeng Lin, Yuetong Fang, Fei Ma, Hao Yu, and Bojun Cheng. Rethinking efficient and effective point-based networks for event camera classification and regression: Eventmamba. *arXiv preprint arXiv:2405.06116*, 2024.

Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12361–12371. IEEE, 2022.

Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pp. 341–357. Springer, 2022.

Bochen Xie, Yongjian Deng, Zhanpeng Shao, and Youfu Li. Eisnet: A multi-modal fusion network for semantic segmentation with events and images. *IEEE Transactions on Multimedia*, 26:8639–8650, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023a.

Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1136–1147. IEEE, 2023b.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 989–997. IEEE, 2019.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.