Following the Flow: Advection-Consistent Modeling for Event-based Small Object Detection

Wen Guo[1], Fulong Cai[1], and Wuzhou Quan[2]*

[1] School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai 264005, China
{wguo,2024410008}@sdtbu.edu.cn
[2] Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
q.wuzhou@gmail.com

Abstract. Event cameras enable high-frequency visual perception with microsecond latency, offering advantages for dynamic scenes. However, event-based small object detection remains challenging due to sparse asynchronous measurements and weak object responses that are easily disrupted by noise. Limited spatial support causes small-object signals to lose temporal continuity, resulting in fragmented and unstable predictions. To address this issue, we propose a physics-guided advection-consistent modeling framework, termed PACT, which formulates event evolution as a motion-driven feature transport process. Instead of relying solely on local spatio-temporal aggregation, PACT propagates features along estimated velocity fields and enforces trajectory-level consistency through advection constraints. This design preserves weak event responses over time and prevents their degradation under complex background interference. Technically, PACT integrates motion-aware feature extraction with a differentiable advection-based transport operator, enabling coherent motion representation and effective noise suppression during temporal evolution. Extensive experiments on benchmark event-based datasets demonstrate that PACT consistently outperforms state-of-the-art methods, achieving improvements of 20.72% in IoU and 15.03% in accuracy while maintaining comparable computational efficiency. The code will be made publicly available.

Keywords: Event cameras · advection-consistent transport · temporal continuity

1 Introduction

Event cameras represent a new generation of vision sensors that respond asynchronously to changes in scene brightness, producing measurements with microsecond level temporal resolution and a wide dynamic range [6, 14]. These properties make event cameras particularly effective at capturing rapid scene dynamics and fine-grained temporal variations driven by brightness changes,

* Corresponding author.

which are difficult to observe with traditional frame-based cameras [11, 22]. Instead of being a sequence of images, an event stream can be interpreted as a spatio-temporal signal whose structure continuously evolves over time. While this representation provides high temporal resolution, it also makes the resulting measurements sparse, irregular, and sensitive to noise [30]. Thus, event responses triggered by small or low-contrast entities are often fragmented in space and intermittent in time, even when their motion follows coherent kinematic patterns, especially under strong background activity and noise [20]. Accurately modeling and maintaining these weak event responses is therefore critical for reliable and generalizable event-based perception under challenging conditions with strong noise.

Existing approaches generally fall into two paradigms. Frame-based pipelines first aggregate events into static representations [17, 19, 36]. This aggregation inevitably collapses the continuous temporal evolution and removes cues needed to distinguish weak signals from noise. Alternatively, Spiking Neural Networks (SNNs) [5, 16, 21, 32] perform temporal integration through neuronal dynamics, yet weak and intermittent responses often fail to elicit stable spiking activity and become indistinguishable from background noise. A critical shared limitation is that both paradigms model temporal continuity implicitly. Without an explicit propagation rule, these approaches struggle to preserve fragmented weak trajectories as coherent structures under strong background activity.

Crucially, although weak entities induce sparse and intermittent event responses, the underlying physical motion remains continuous. This property fundamentally distinguishes such weak responses from background noise, which lacks a persistent motion source [6]. Within sufficiently short time intervals, event responses induced by the same entity tend to be temporally coherent and continuously displaced along a consistent motion direction, whereas noise-induced triggers exhibit no such directional persistence. In this regime, motion continuity becomes the only remaining cue capable of separating entity-induced weak responses from random background activity.

To formalize this distinction, we adopt the weakest physical constraint that captures motion continuity without imposing stronger assumptions on appearance, shape, or intensity. Over short time scales, the evolution of event responses can be well approximated by advection under a local velocity field [2, 7, 38], while higher-order effects such as deformation or acceleration are negligible. This velocity-driven spatiotemporal evolution naturally aligns with advection in physics. Importantly, advection imposes a minimal and unavoidable constraint: responses originating from the same moving entity remain under a consistent transport field, whereas scattered noise triggers, lacking a coherent motion source, cannot satisfy such consistency.

Motivated by this observation, we introduce an advection-consistency constraint to explicitly model the propagation of event responses over time. Under this constraint, responses that can be consistently transported along a local velocity field remain temporally aligned and accumulate coherently, whereas responses that violate advection consistency lose alignment and are progressively

suppressed. Advection consistency therefore provides an operational criterion for preserving weak yet physically meaningful trajectories in highly noisy event streams.

Building upon this formulation, we propose **PACT**, a framework that leverages a Physics-consistent Advection-Consistent Transport mechanism to maintain weak yet motion-consistent event structures. PACT consists of three complementary components that work in concert. The Trajectory-Guided Feature Extraction (T-FE) module distills motion-aware features from sparse event measurements to support trajectory reasoning. The Advection-based Trajectory Consistency (ATC) module estimates a locally consistent velocity field, supervised by a trajectory-consistent velocity loss, and enforces temporal coherence via advection consistency. Finally, the Advection-Consistent Feature Reconstruction (A-FR) module aligns and reconstructs features along the estimated velocity field, turning fragmented responses into continuous trajectory structures, while background clutter is progressively suppressed when transport consistency is absent. In practice, weak event responses are most prominent in small-object scenarios, where sparse and intermittent activations are easily dominated by background activity. We evaluate PACT on event-based small object detection benchmarks as a concrete instantiation of the proposed formulation, demonstrating its ability to recover weak trajectories and suppress noise under challenging conditions. Although our experiments are conducted in the context of detection, the proposed advection-based formulation is defined independently of any specific task and provides a general mechanism for maintaining weak spatio-temporal structures in event streams.

2 Related Work

2.1 Event Representation and Temporal Association

A common way to process event streams is to convert asynchronous events into regular tensors so that standard backbones can be applied [17]. Early designs aggregate polarities or counts within a window, while later ones inject timing cues, such as time surfaces and kernelized timestamp encoding [8, 13]. Reconstruction-based pipelines further map events to intensity-like proxies to reintroduce photometric cues [26]. To preserve temporal structure more explicitly, voxel-based methods discretize events into spatiotemporal volumes and apply 3D operators or recurrent updates to extend the temporal support [9, 25, 38]. Other approaches keep state evolution with spiking dynamics or sparse relational reasoning [3, 28, 32, 37]. Despite different forms, most methods rely on implicit continuity. Cross-temporal linkage is largely driven by local neighborhoods or short-range propagation. Under strong background activity, weak responses drift and appear intermittently, so evidence fragments rather than accumulating into motion-consistent traces.

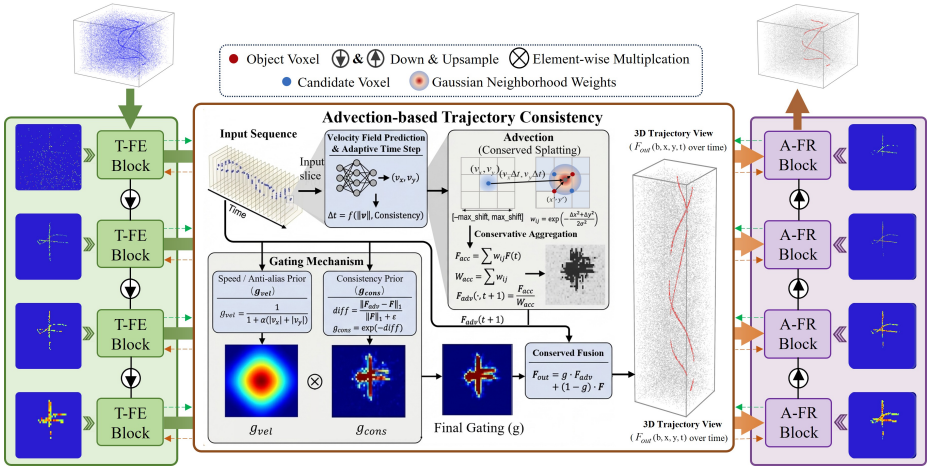


Fig. 1: Overall architecture of PACT. The encoder stacks four T-FE modules. ATC estimates velocity fields and consistency scores at each scale. The decoder uses four A-FR modules with skip fusion to reconstruct temporally continuous trajectories.

2.2 Physics-guided Dynamics

To stabilize motion cues in low-SNR event streams, many works introduce physical priors. A classic line is contrast maximization, which warps events to a reference time under a motion model and optimizes the motion parameters to sharpen the compensated accumulation [7, 18, 31]. Other formulations propagate sparse evidence with dynamical processes, including diffusion-like smoothing and probabilistic motion modeling [29,35]. Diffusion can denoise, but isotropic smoothing may also spread clutter and weaken directional patterns. Advection provides a motion-aligned alternative. A local velocity field specifies how information should move over short time spans [2, 7, 38]. PACT builds on this view and models representation evolution as advection-consistent transport, so aligned responses persist and accumulate, while inconsistent triggers lose alignment and are suppressed.

3 Methods

3.1 Advection-Consistent Representation

Given sparse, asynchronous event streams, informative responses often appear as sparse activations with fragmented temporal support, yet reveal coherent structure when traced over time. Coherent activations remain mutually alignable when propagated by a local motion field, whereas spurious triggers from background activity quickly lose temporal consistency over time. This motivates an advection-consistent representation, where the feature field is approximately conserved under advection driven by the local motion field.

We treat the sparse voxel features as samples of a C-channel field $u(\mathbf{x}, t)$ defined over space and time. Within a short time interval, its evolution is approximated by the advection equation [2]:

$$\frac{\partial u(\mathbf{x}, t)}{\partial t} + \mathbf{v}(\mathbf{x}, t) \cdot \nabla u(\mathbf{x}, t) \approx 0, \tag{1}$$

where $\mathbf{v}(\mathbf{x}, t)$ denotes a local velocity field. Along a characteristic curve $\mathbf{x}(t)$ satisfying $\frac{d\mathbf{x}(t)}{dt} = \mathbf{v}(\mathbf{x}(t), t)$, the transported feature is approximately preserved,

$$\frac{d}{dt} u(\mathbf{x}(t), t) \approx 0. \tag{2}$$

For a discrete temporal step τ and locally constant velocity, the transport admits a displacement operator:

$$\mathcal{T}_{\mathbf{v}}(u)(\mathbf{x}, t) \;=\; u(\mathbf{x} - \tau\mathbf{v}, t - \tau). \tag{3}$$

Advection consistency corresponds to a small deviation under $\mathcal{T}_{\mathbf{v}}$ [38]. We therefore measure the transport residual as:

$$\mathcal{R}_{adv}(\mathbf{x}, t) \;=\; \left\| u(\mathbf{x}, t) - \mathcal{T}_{\mathbf{v}}(u)(\mathbf{x}, t) \right\|, \tag{4}$$

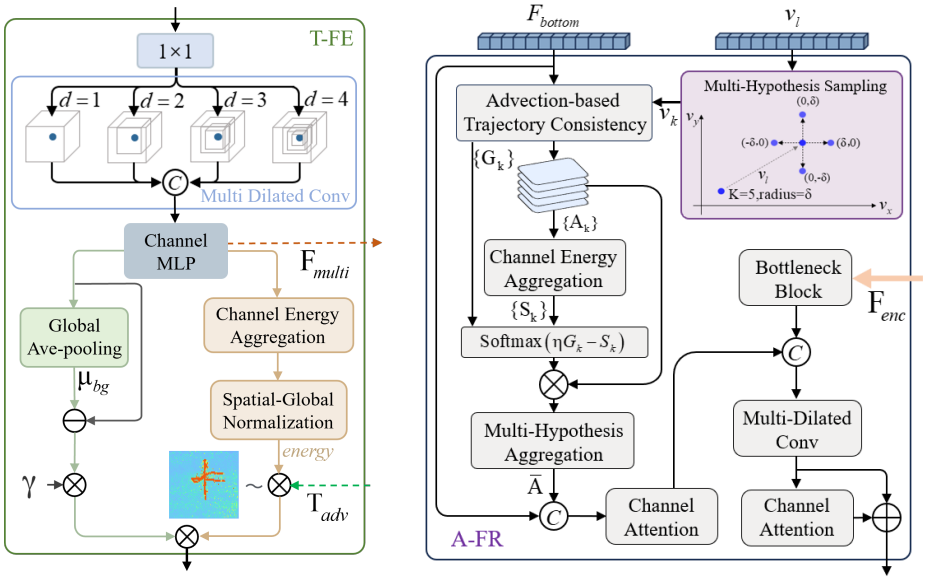
where $\|\cdot\|$ is a channel-wise discrepancy measure. Throughout the pipeline, $\mathcal{R}_{adv}$ provides a consistency signal. Features with smaller residuals receive higher confidence and are propagated more reliably under the estimated local velocity field.

3.2 Trajectory-Constrained Feature Encoding

As illustrated in Fig. 1, the pipeline follows an encoder–decoder architecture with multi-scale sparse processing. The input is a sparse voxel tensor $\mathrm{X} = \{\mathrm{C}, \mathrm{A}\}$, where C stores voxel indices (b, x, y, t) and A stores normalized attributes $(\tilde{x}, \tilde{y}, \tilde{t}, p)$ with polarity p. The encoder progressively extracts motion-aware features using four stages of sparse processing, producing a pyramid $\{\mathrm{F}^i_{enc}\}_{i=1}^4$. Rather than treating motion priors as an external post-processing step, we impose trajectory consistency *inside* feature encoding. At each scale, the representation is shaped to preserve components that remain coherent under transport, so that the subsequent decoding can propagate and connect weak responses more reliably.

At each scale, as shown in Fig. 2(a), T-FE aggregates spatiotemporal context to support transport assessment. Specifically, we form F_{multi} via multi-branch sparse convolutions with dilation rates $\mathcal{D} = \{1, 2, 3, 4\}$. ATC then evaluates transport feasibility on F_{multi} and outputs participation weights for subsequent propagation.

At the core is a trajectory-consistency evaluator that predicts a local displacement field and measures the agreement between the observed feature and



(a) T-FE module. **(b)** A-FR module.

Fig. 2: Architectures of the proposed modules. (a) T-FE filters and enhances sparse responses under trajectory-consistency guidance. (b) A-FR propagates features along the estimated velocity field to reconstruct continuous trajectories while suppressing incoherent triggers.

its transported prediction. For each active voxel i with feature F_i, we predict a bounded 2D displacement $\Delta\mathbf{p}_i$ on the spatial grid:

$$\Delta\mathbf{p}_i = \mathrm{v}_{max}\tanh(\mathrm{MLP}(\mathrm{F}_i)), \quad \mathbf{p}'_i = \mathbf{p}_i + \Delta\mathbf{p}_i, \tag{5}$$

where v_{max} limits the maximum displacement for stability.

Since $\mathbf{p}'_i$ is generally off-grid, we obtain F_{adv} by Gaussian-kernel sampling on the (x, y) lattice with normalized weights.

We quantify trajectory consistency using a gating weight g that combines the agreement between F_{adv} and F and a mild penalty on large velocities,

$$g = \exp\left(-\frac{\|\mathrm{F}_{adv} - \mathrm{F}\|_1}{\|\mathrm{F}\|_1 + \varepsilon}\right) \odot \frac{1}{1 + \alpha\,(|v_x| + |v_y|)}. \tag{6}$$

This weight yields a conservative fusion between the transported prediction and the original observation,

$$\mathrm{T}_{adv} = g \cdot \mathrm{F}_{adv} + (1 - g) \cdot \mathrm{F}. \tag{7}$$

Intuitively, g acts as a trajectory-consistency score: components that remain stable under transport receive stronger support from the transported prediction,

while inconsistent components are prevented from dominating the representation. We use T_{adv} and g to constrain encoding at each scale, producing features that are better aligned with the estimated velocity field and thus more amenable to transport-based propagation.

3.3 Advection-Guided Trajectory Propagation

The decoder aims to transform deep motion-aware features into temporally continuous trajectory structures. This is challenging in event data because coherent activations are intermittent, and naive upsampling tends to amplify isolated triggers. We therefore perform decoding as *advection-guided propagation* by extending features along the velocity field to connect fragmented responses over time, while naturally down-weighting components that do not admit coherent transport.

At each decoding stage, as illustrated in Fig. 2 (b), we take the lower-resolution feature F_{bottom} and the corresponding velocity guidance v_i from the encoder, and perform transport-based alignment. Because velocity estimation can be uncertain in sparsely activated regions, we form a set of perturbed hypotheses around v_i,

$$\mathrm{v}_i^k = \mathrm{v}_i + \delta_k, \quad k = 1, \ldots, K, \tag{8}$$

where δ_k are fixed small perturbations. For each hypothesis, we compute a transported candidate and its trajectory consistency,

$$\{\mathrm{A}_k, \mathrm{G}_k\} = \mathrm{ATC}(\mathrm{F}_{bottom}, \mathrm{v}_i^k), \tag{9}$$

where A_k is the transported feature and G_k is the corresponding confidence derived from transport agreement. To avoid linking trajectories through abnormally high activations, we introduce a magnitude penalty S_k and compute consistency-aware aggregation weights:

$$\pi_k = \mathrm{Softmax}\left(\eta\, \mathrm{G}_k - \mathrm{S}_k\right), \qquad \bar{\mathrm{A}} = \sum_{k=1}^{K} \pi_k \mathrm{A}_k. \tag{10}$$

Here S_k is the channel-averaged energy of A_k.

We then combine the aligned feature with the original bottom feature to avoid information loss,

$$\mathrm{F}_{align} = \mathrm{CA}\Big(\mathrm{Concat}(\mathrm{F}_{bottom}, \bar{\mathrm{A}})\Big), \tag{11}$$

where CA is a lightweight channel attention. Finally, we inject the corresponding encoder skip feature F_{enc} to recover spatial details. The skip feature is first compressed by a bottleneck transform, concatenated with F_{align}, and refined by multi-dilated sparse convolutions with a residual connection,

$$\mathrm{F}_{traj} = \mathrm{CA}\Big(\mathrm{DConv}_d\big(\mathrm{Concat}(\mathrm{F}_{align}, \mathrm{Bottleneck}(\mathrm{F}_{enc}))\big) + \mathrm{F}_{align}\Big). \tag{12}$$

Repeating this process across scales produces a trajectory feature pyramid $\{\mathrm{F}_{traj}^{i}\}_{i=1}^{4}$, whose responses are progressively connected along time by advection-guided transport, yielding temporally coherent trajectory representations for downstream prediction.

3.4 Learning Objectives

We train the network with a joint objective that combines voxel-level segmentation supervision and a regularizer on the predicted velocity field,

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{seg} + \lambda\mathcal{L}_{vel}, \tag{13}$$

where λ balances the two terms. We use binary cross-entropy for $\mathcal{L}_{seg}$ over predicted mask probabilities.

Dense flow annotations are unavailable in typical event settings. We therefore construct **pseudo velocity targets** via sparse temporal matching between foreground voxels across nearby time slices. For a foreground voxel i at time t with center p_i, we find a nearest matched voxel j from the same instance at a neighboring time t', and define:

$$\mathrm{v}_i^* = \frac{\mathrm{p}_j - \mathrm{p}_i}{t' - t}. \tag{14}$$

We supervise the predicted velocity with the Smooth-ℓ_1 loss,

$$\mathcal{L}_{vel} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \mathrm{SmoothL1}\left(\mathbf{v}_i - \mathbf{v}_i^*\right), \tag{15}$$

where $\mathcal{V}$ is the set of supervised foreground voxels. This regularizer stabilizes the learned transport field and improves temporal coherence, while the main supervision remains on the final task output.

4 Experiments

4.1 Implementation Details

All experiments are conducted on the EV-UAV dataset [4], which contains 147 sequences and over 2.3M event-level annotations. The targets are extremely small, with an average size of 6.8×5.4 pixels, and appear under complex backgrounds and challenging illumination. We follow the official split with 99 training and 24 test sequences. All compared methods are retrained on the same split. For fairness, all methods are evaluated on the same raw event streams, annotations, and metrics, while each baseline keeps its native event representation.

Input streams are sliced into 8s windows and voxelized at a temporal resolution of 1 ms. We train PACT for 50 epochs using Adam [12] with an initial learning rate of 10^{-3} and decay on an NVIDIA RTX 4060 Ti GPU. For all experiments, we use the same hyperparameter setting: $\alpha = 0.25$, $\epsilon = 10^{-6}$, $K = 5$, $\delta_k = 0.6$, $\eta = 0.75$, and $\lambda = 0.3$. We report IoU and Acc for event-level segmentation quality, and P_d and F_a for localization performance.

Table 1: Quantitative comparison of the proposed method to state-of-the-art methods. The **bold** and the underline represent the best and second-best performance, respectively.

Methods	Publications	Event Rep.	Temporal	$IoU(\%)\uparrow$	$ACC(\%)\uparrow$	$P_d(\%)\uparrow$	$F_a(10^{-4})\downarrow$	#Params.	Runtime(ms)
SSD	ECCV 2016	Event Count	No	25.31	28.56	26.31	486.63	25.2M	2113
Faster RCNN	TPAMI 2016	Event Count	No	26.93	29.68	27.39	689.68	41.2M	3962
DETR	ICLR 2020	Event Count	No	30.35	33.63	31.64	631.37	39.8M	3136
YOLOv10-S	NIPS 2025	Event Count	No	32.55	33.39	32.18	589.67	7.3M	1627
EMS-YOLO	ICCV 2023	SNN	Yes	36.77	42.92	50.68	112.36	3.3M	1229
Spike-YOLO	ECCV 2024	SNN	Yes	43.94	48.26	59.62	55.38	69.0M	1883
GET	ICCV 2023	Group Token	Yes	40.31	48.91	60.73	46.35	18.4M	2168
RED	NeurIPS 2020	Voxel Grid	Yes	35.99	45.54	53.76	102.27	24.1M	3427
RVT	CVPR 2023	Voxel Grid	Yes	43.21	51.38	60.35	55.68	9.9M	1737
SAST	CVPR 2024	Voxel Grid	Yes	34.31	40.22	51.21	150.32	18.5M	3075
KPConv	ICCV 2019	Points	Yes	48.19	57.28	68.59	16.32	50.1M	562
RandLA-Net	CVPR 2020	Points	Yes	50.32	59.29	70.56	6.95	**1.2M**	353
COSeg	CVPR 2024	Points	Yes	51.89	60.93	71.32	9.21	23.4M	364
EV-SpSegNet	ICCV 2025	Points	Yes	55.18	65.02	77.53	1.63	4.0M	**36**
Ours	-	Points	Yes	**75.90**	**80.05**	**91.84**	**0.76**	2.9M	58

4.2 Benchmark Comparisons

We divide the compared methods into three groups and select several state-of-the-art approaches in each group. The first group consists of frame-based generic object detectors, including SSD [15], Faster R-CNN [27], Deformable DETR [39], and YOLOv10 [34]. The second group covers event-based detection methods, including RVT [9], SAST [23], GET [24], RED [25], EMS-YOLO [32], and Spike-YOLO [16]. The third group contains point-cloud segmentation methods, including KPConv [33], RandLA-Net [10], COSeg [1], and EV-SpSegNet [4]. For the frame-based methods, the input is an event frame aggregated over a 50 ms time window.

Table 1 reports the results of frame-based, event-based, and point-based detectors. Frame-based methods perform the worst, as 50 ms event aggregation removes much of the within-window temporal structure. Event-based models improve temporal modeling but remain limited in low-SNR scenes with heavy clutter. Point-based segmentation baselines perform best among existing methods, with EV-SpSegNet reaching 55.18% IoU and 77.53% P_d. PACT further increases the IoU to 75.90% and P_d to 91.84%, while reducing F_a from 1.63 to 0.76 ($\times 10^{-4}$). With 2.9M parameters and 58 ms inference time per window, PACT achieves this gain without increasing model complexity.

Qualitative Results: Fig. 3 shows representative qualitative results. RVT frequently misses targets and produces scattered detections under clutter. Point-based methods preserve sparse geometry better, but still suffer from trajectory breaks and residual background triggers. PACT produces denser and more continuous target trajectories with fewer missed and false detections, confirming the benefit of advection-consistent propagation.

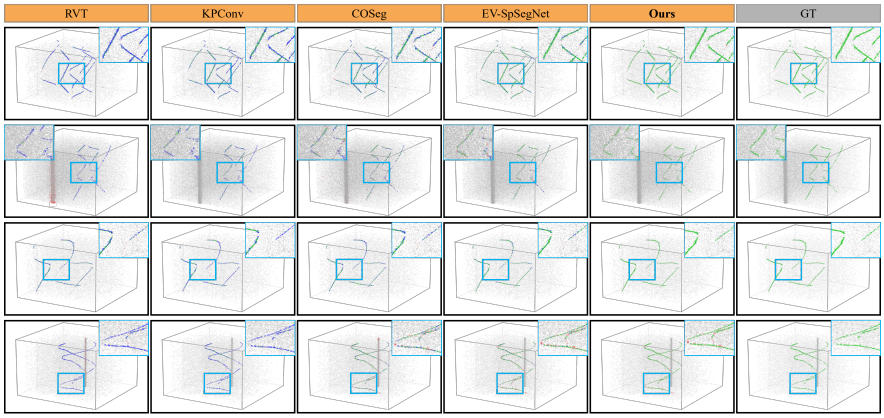


Fig. 3: Qualitative results of different methods. Correct detections are marked in green, false detections in red, missed detections in blue, and noise points in gray.

Table 2: Ablation study on framework scheme.

T-FE	ATC	A-FR	IoU↑	ACC↑	P_d ↑	F_a ↓	#Params.
			59.90	63.18	78.80	0.76	2.3M
✓			66.54	69.67	84.48	0.64	2.6M
	✓		55.11	69.95	83.35	4.76	2.4M
		✓	63.26	65.34	79.97	0.37	2.5M
✓	✓		71.95	73.48	85.61	**0.18**	2.7M
✓		✓	70.55	76.04	88.76	1.54	2.8M
✓	✓	✓	**75.90**	**80.05**	**91.84**	0.76	2.9M

4.3 Ablation Studies

We evaluate the physical formulation of PACT in Table 2. The Baseline has no explicit transport constraint, so weak responses remain temporally fragile and are easily overwhelmed. When ATC is attached to the Baseline, the false alarm rate rises to 4.76×10^{-4}. This indicates that enforcing advection consistency on unstructured features induces spurious temporal correspondences, because the baseline features provide no a stable substrate for transport consistency. With only T-FE or only A-FR, the model behaves conservatively. These variants suppress noise effectively, yet they do not aggregate intermittent responses into coherent temporal traces, which limits the detection gain. The full benefit emerges only after closing the physical loop. T-FE produces motion-consistent features, and ATC imposes a transport-feasibility constraint on their propagation. A-FR then extends the consistent responses along the estimated flow to form stable temporal traces. Overall, the ablation provides evidence for a coupled cycle of extraction, constraint, and propagation for sustaining weak temporal traces.

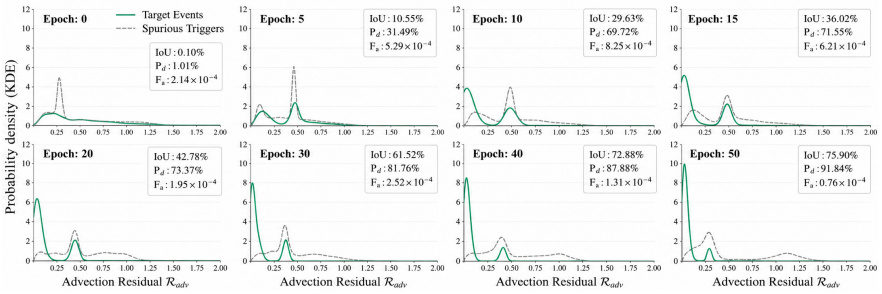


Fig. 4: Advection residual density across training stages. Kernel density estimates show the advection residual $\mathcal{R}_{adv}$ for target events and spurious triggers, with the corresponding detection metrics reported next to each stage.

4.4 Advection-Guided Representation Evolution

PACT uses advection consistency to preserve weak, fragmented temporal responses while suppressing triggers that do not support coherent transport. A key question is whether this constraint actually reshapes the internal representation, rather than merely improving the final metrics. To examine this, we track the distribution of advection residuals during training and test whether foreground responses progressively concentrate at low residuals while background activity remains broadly distributed at higher residuals.

For each checkpoint, we measure the advection residual between features before and after transport under the estimated velocity field,

$$\mathcal{R}_{adv} = \|\mathbf{F} - \mathcal{T}_{\mathbf{v}}(\mathbf{F})\|\,, \tag{16}$$

and plot kernel density estimates for foreground (green) and background (gray) responses using ground-truth labels. Fig. 4 shows four stages from initialization to convergence, together with the corresponding detection metrics.

At early stages, foreground and background residuals largely overlap, indicating that transport does not yet provide reliable alignment for weak responses. As training proceeds, foreground residuals form a sharp peak near zero and separate clearly from the background, whereas background residuals remain broad at higher values. A near-zero residual means that the transported prediction agrees with the observation, allowing intermittent target activations to accumulate into coherent temporal traces, while clutter remains in the high-residual regime. Overall, the growing separation is consistent with the metric improvement and provides direct evidence that transport consistency stabilizes weak traces under heavy clutter.

4.5 Temporal Continuity Analysis

To quantitatively evaluate temporal consistency, we analyze the distribution of continuity metrics over all test sequences. For each target, we convert the frame-wise hit status into a binary hit sequence and compute two complementary

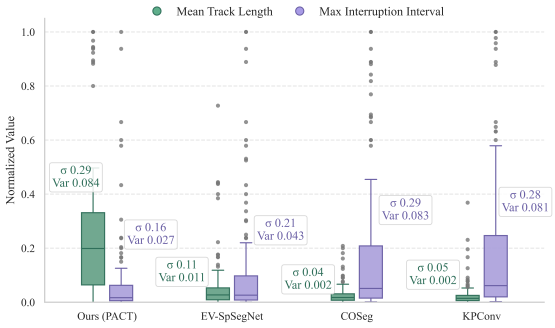


Fig. 5: Comparison of temporal continuity statistics across methods. We report the average length of continuous successful detection and the maximum length of continuous detection loss per target.

measures in Fig. 5. The first metric is the mean track length, defined as the average length of continuous successful detection segments. The second metric is the max interruption interval, defined as the maximum length of consecutive missed detections. Both metrics are normalized by the ground-truth target duration so that values lie in $[0, 1]$, enabling fair comparisons across targets with different temporal extents.

Fig. 5 shows that our method yields stronger temporal continuity than the baselines. For mean track length, our method attains a higher median and a distribution shifted toward larger values, whereas EV-SpSegNet [4], COSeg [1], and KPConv [33] concentrate at smaller values. This indicates that baseline predictions are frequently fragmented into short segments that appear only when local evidence is strong and break once the evidence weakens. In contrast, our approach maintains connected detections over longer spans, producing more coherent trajectories.

For max interruption interval, our method achieves a lower mean and reduced dispersion. For example, the standard deviation decreases from 0.28 for KPConv [33] to 0.15 for our method. These results align with our design motivation. By modeling advection-based transport, our method preserves temporal connections through physics-consistent constraints, which helps bridge gaps caused by fluctuating event rates rather than relying on transient peaks.

Beyond the statistics, Fig. 6 provides a complementary visual validation. We align each target's detection trajectory to the normalized target duration and visualize the frame-wise hit ratio, defined as the fraction of ground-truth target points correctly detected at each frame. Colors closer to blue indicate fewer hits, while warmer colors indicate more hits. Across targets and scenes, the baselines often exhibit bursty and intermittent responses, with short high-hit segments separated by frequent low-hit periods. Our method instead forms longer and more continuous high-hit bands and markedly reduces prolonged low-hit stretches, suggesting that the predictions remain temporally connected rather than driven by brief local spikes.

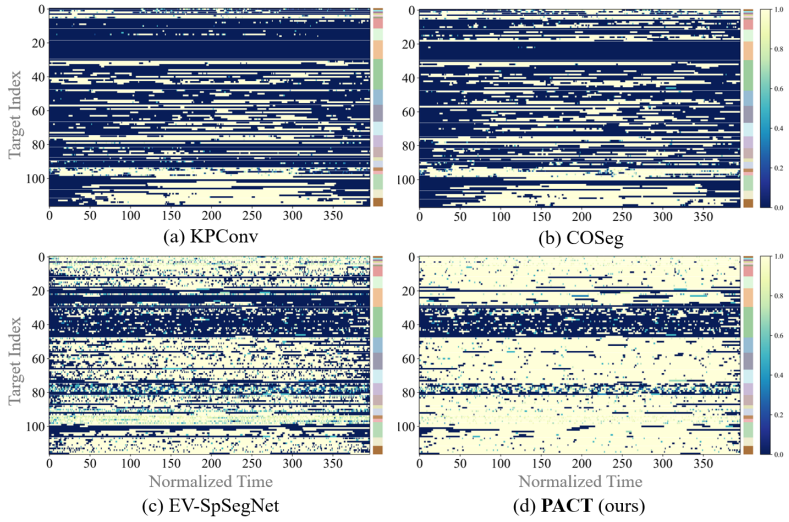


Fig. 6: Comparison of temporal continuity visualizations across methods. We align each target's detection trajectory to a unified time scale and visualize the hit ratio at each time step with color, where dark blue indicates no hit and yellow indicates a higher hit ratio. The right-side color strip groups targets by scene, with each block corresponding to one scenario, which makes it easy to compare continuous detection segments and interruption patterns across methods.

Crucially, the 20% relative IoU improvement in Table 1 is consistent with this enhanced temporal integrity. Longer continuous detections and fewer extended interruptions reduce temporal breakages and fragmented accumulation over time. As a result, predicted regions evolve more coherently across the sequence and yield more complete masks even when foreground evidence is intermittent.

4.6 Multi-Target Motion with Large Velocity Variation

Fig. 6 shows that PACT improves temporal continuity in most scenes, but the gain over EV-SpSegNet shrinks for the scene group around target index ∼40. We analyze this group to identify motion patterns under which single-field advection brings limited additional benefit.

Qualitative evidence (Fig. 7(a)). We visualize a three-frame window, where marker shapes indicate time and colors indicate target IDs. Multiple targets co-occur within the same window, yet their inter-frame displacements differ substantially. PACT propagates evidence with a single first-order velocity field; this works well when co-occurring targets share similar motion, but it cannot align multiple trajectories with markedly different velocities. As a result, temporal aggregation weakens for a subset of targets, reducing the continuity gain in this group.

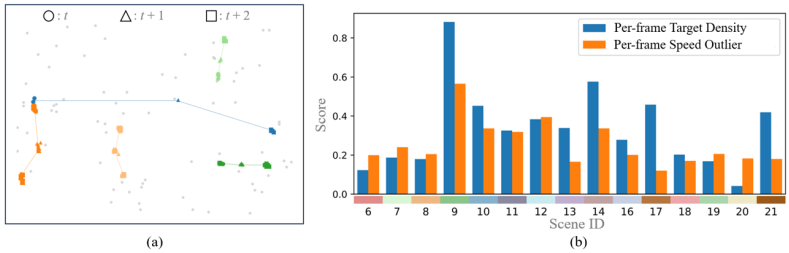


Fig. 7: Mechanism study of challenging scenes. (a) Three-frame visualization of a representative case. Marker shape indicates time (circle: t, triangle: t+1, square: t+2), and color indicates target ID. Co-occurring targets exhibit markedly different displacements within the same window. (b) Scene-level statistics after removing simple scenes without temporal target co-occurrence, since they do not form competing motions within a window. Per-frame Target Density counts how many target IDs are active simultaneously in a frame. Per-frame Speed Outlier measures how strongly the fastest targets deviate from the typical motion among co-occurring targets.

Scene-level statistics (Fig. 7(b)). We compute statistics only on scenes with temporal target co-occurrence, and each color block matches a scene group in Fig. 6. The target-index-$\sim$40 group (Scene 9) exhibits higher *Per-frame Target Density* and the highest *Per-frame Speed Outlierness*. This indicates that more targets compete within the same window and that a few targets frequently move much faster than the rest, which makes a single first-order field insufficient to align all trajectories consistently. The statistics agree with the reduced margin in Fig. 6 and support the above mechanism.

Overall, the hard cases are characterized by within-window motion inconsistency, especially the presence of fast outliers or abrupt speed changes. Extending the transport model with time-varying motion may better handle such mixed-motion windows.

5 Conclusion and Future Work

In this work, we formulate event streams as a continuously evolving spatiotemporal process. Under heavy clutter, the critical challenge is not merely sparsity, but the lack of an explicit propagation mechanism. We cast the mechanism as a local transport process, where motion continuity is expressed as approximate feature conservation along an estimated velocity field. Guided by this view, PACT enforces advection consistency to sustain weak responses and suppress noise.

In our experiments, we examine the mechanism by tracking the evolution of transport-residual distributions across checkpoints. As the model converges, target responses increasingly concentrate in a low-residual regime, whereas background triggers remain broadly distributed at higher residuals. The growing separation mirrors the improvement in downstream metrics, providing indirect

evidence that the model relies on transport validity to maintain weak responses instead of memorizing static patterns.

However, the current formulation models transport with locally constant velocity, which is best suited to short, near-linear motion. When acceleration or other non-linear dynamics become prominent, this approximation may become inaccurate. Future work will explore transport models beyond first-order motion so that the same consistency principle extends to more complex dynamics.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No.62072286 and in part by Shandong Province Graduate Education Innovation Program Project under Grant No.SDYAL21211.

References

1. An, Z., Sun, G., Liu, Y., Liu, F., Wu, Z., Wang, D., Van Gool, L., Belongie, S.: Rethinking few-shot 3d point cloud semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3996–4006 (2024)
2. Benosman, R., Clercq, C., Lagorce, X., Ieng, S.H., Bartolozzi, C.: Event-based visual flow. IEEE Transactions on Neural Networks and Learning Systems **25**(2), 407–417 (2014)
3. Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., Andreopoulos, Y.: Graph-based object classification for neuromorphic vision sensing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 491–501 (2019)
4. Chen, N., Xiao, C., Dai, Y., He, S., Li, M., An, W.: Event-based tiny object detection: A benchmark dataset and baseline (2025), arXiv:2506.23575
5. Cordone, L., Miramond, B., Thierion, P.: Object detection with spiking neural networks on automotive event data. In: Proceedings of the International Joint Conference on Neural Networks. pp. 1–8. IEEE (2022)
6. Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(1), 154–180 (2022)
7. Gallego, G., Rebecq, H., Scaramuzza, D.: A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3867–3876 (2018)
8. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5632–5642 (2019)
9. Gehrig, M., Scaramuzza, D.: Recurrent vision transformers for object detection with event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13884–13893 (2023)
10. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11108–11117 (2020)

11. Kim, H., Leutenegger, S., Davison, A.J.: Real-time 3d reconstruction and 6-dof tracking with an event camera. In: Proceedings of the European Conference on Computer Vision. Lecture Notes in Computer Science, vol. 9910, pp. 349–364. Springer (2016)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014), arXiv:1412.6980
13. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: Hots: A hierarchy of event-based time-surfaces for pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(7), 1346–1359 (2017)
14. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128×128 120 db 15 µs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits **43**(2), 566–576 (2008)
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision. pp. 21–37. Springer (2016)
16. Luo, X., Yao, M., Chou, Y., Xu, B., Li, G.: Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. In: Proceedings of the European Conference on Computer Vision. pp. 253–272. Springer (2024)
17. Maqueda, A.I., Loquercio, A., Gallego, G., García, N., Scaramuzza, D.: Event-based vision meets deep learning on steering prediction for self-driving cars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5419–5427 (2018)
18. Mitrokhin, A., Fermüller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1–9 (2018)
19. Mitrokhin, A., Hua, Z., Fermüller, C., Aloimonos, Y.: Learning visual motion segmentation using event surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14414–14423 (2020)
20. Mondal, A., Giraldo, J.H., Bouwmans, T., Chowdhury, A.S.: Moving object detection for event-based vision using graph spectral clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 876–884. IEEE (2021)
21. Neftci, E.O., Mostafa, H., Zenke, F.: Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. IEEE Signal Processing Magazine **36**(6), 51–63 (2019)
22. Pan, L., Hartley, R., Scheerlinck, C., Liu, M., Yu, X., Dai, Y.: High frame rate video reconstruction based on an event camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2519–2533 (2022)
23. Peng, Y., Li, H., Zhang, Y., Sun, X., Wu, F.: Scene-adaptive sparse transformer for event-based object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16794–16804 (2024)
24. Peng, Y., Zhang, Y., Xiong, Z., Sun, X., Wu, F.: Get: Group event transformer for event-based vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6038–6048 (2023)
25. Perot, E., de Tournemire, P., Nitti, D., Masci, J., Sironi, A.: Learning to detect objects with a 1 megapixel event camera. In: Advances in Neural Information Processing Systems. vol. 33, pp. 16639–16652 (2020)
26. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(6), 1964–1980 (2021)

27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems **28** (2015)
28. Schaefer, S., Gehrig, D., Scaramuzza, D.: Aegnn: Asynchronous event-based graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12371–12381. IEEE (2022)
29. Sekikawa, Y., Nagata, J.: Tangentially elongated gaussian belief propagation for event-based incremental optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 21940–21949 (2023)
30. Shariff, W., Dilmaghani, M.S., Kielty, P., Moustafa, M., Lemley, J., Corcoran, P.: Event cameras in automotive sensing: A review. IEEE Access **12**, 51275–51306 (2024)
31. Stoffregen, T., Gallego, G., Drummond, T., Kleeman, L., Scaramuzza, D.: Event-based motion segmentation by motion compensation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7244–7253 (2019)
32. Su, Q., Chou, Y., Hu, Y., Li, J., Mei, S., Zhang, Z., Li, G.: Deep directly-trained spiking neural networks for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6555–6565 (2023)
33. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6411–6420 (2019)
34. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J.: Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems **37**, 107984–108011 (2024)
35. Wang, X., Jin, Y., Wu, W., Zhang, W., Zhu, L., Jiang, B., Tian, Y.: Object detection using event camera: A moe heat conduction based detector and a new benchmark dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29321–29330 (2025)
36. Yang, C., Huang, Z., Wang, N.: Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13668–13677 (2022)
37. Zhang, J., Dong, B., Zhang, H., Ding, J., Heide, F., Yin, B., Yang, X.: Spiking transformers for event-based single object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8801–8810. IEEE (2022)
38. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)
39. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection (2020), arXiv:2010.04159