Unsupervised 3d Motion Estimation Using Event Camera

Han Han[1], Wei Zhai[1†], Tiesong Zhao[2], Bin Li[1], Yang Cao[1], Zheng-jun Zha[1]

[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China [2]Fuzhou University

{hanh@mail, wzhai056@}ustc.edu.cn, t.zhao@fzu.edu.cn

{binli, forrest, zhazj}@ustc.edu.cn

Abstract

Estimating the 3D motion of scene points from 2D observations, typically parameterized by optical flow and motion in depth, is a fundamental problem in computer vision. Existing learning-based methods usually rely on supervised regression from densely labeled data, but their dependence on annotations and limited use of geometric constraints restricts generalization, motivating unsupervised solutions. Unsupervised 3D motion estimation is challenging because motion along the viewing direction is unobservable, and optical flow and motion in depth are geometrically coupled, making their separation ambiguous. Event cameras capture per-pixel brightness changes asynchronously with microsecond latency, providing high temporal resolution and motion continuity. Projecting event streams along different axes reveals spatiotemporal expansion and contraction patterns that encode depth variation and geometric structure, offering rich cues for unsupervised estimation. Leveraging these properties, we propose an unsupervised event-based 3D motion estimation framework that jointly models optical flow and motion in depth. We first derive an analytical relationship to infer initial motion in depth from estimated flow and further refine it using a directional expansion modulation module that captures horizontal and vertical expansion–contraction patterns in event projections. Finally, motion in depth is incorporated into optical flow warping under a contrast maximization objective. Experiments on the CarlaEvent3D dataset show that our method achieves competitive accuracy and strong generalization, advancing unsupervised 3D motion estimation in the event domain.

1. Introduction

Visual 3D motion estimation aims to recover the motion of scene points in 3D space from their 2D projections over time, typically represented by optical flow on 2D plane and motion in depth along the viewing direction [19, 23]. It serves as a cornerstone for perception in applications such as robotic navigation and autonomous driving, where reliable and precise modeling of dynamic scenes is essential to spatial intelligence [1, 18].

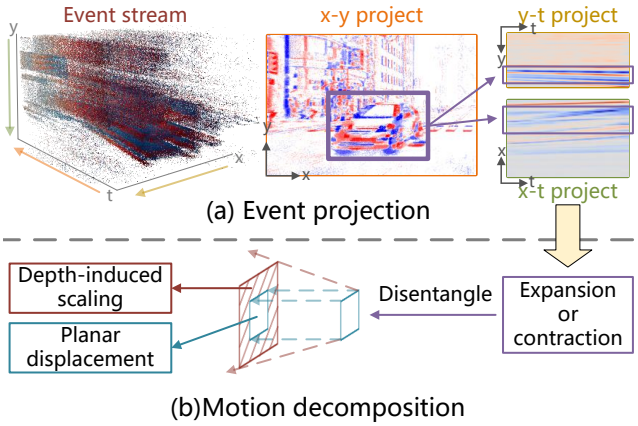

Figure 1. (a) Event projections along different axes. X-t and y-t projection reveal local expansion–contraction patterns that encode depth variation and motion geometry, providing complementary cues beyond x-y projection. (b) Motion decomposition into planar displacement and depth-induced scaling. Event projections provide complementary cues that help disentangle these two components, enabling more reliable estimation of motion in depth.

Existing methods [19, 22, 23, 35] follow a supervised paradigm, directly regressing 3D motion fields from densely labeled data. However, such models tend to rely heavily on the distribution of annotated samples and do not explicitly incorporate the geometric constraints that inherently govern motion. As a result, they often overfit to dataset-specific motion patterns, leading to limited generalization when encountering unseen environments. These limitations motivate the development of unsupervised ap-

† Corresponding author.

proaches that leverage intrinsic geometric and structural constraints, rather than depending solely on labeled data.

However, unsupervised 3D motion estimation faces two key challenges. First, motion along the viewing direction is unobservable, so estimating flow and motion in depth simultaneously requires strong priors that are unavailable in an unsupervised setting. Second, optical flow and motion in depth are coupled through projection geometry: each pixel's motion depends on 2D displacement and depth-induced scaling, making their disentanglement from observations inherently ambiguous.

Despite these challenges, event cameras provide unique advantages for unsupervised 3D motion estimation due to their high-temporal-resolution measurements, which capture continuous scene dynamics. Leveraging this property, projecting events along different axes indirectly reveals motion along the viewing direction in two complementary ways. First, the local expansion and contraction of patterns encode relative depth changes, providing cues for refining motion in depth. Second, the spatial distribution and temporal evolution of patterns reveal geometric correlations between planar motion and depth-induced scaling, offering signals to partially disentangle the two. As illustrated in Fig. 1(a), projections onto the x-y, y-t, and x-t planes expose spatial structure as well as fine-grained temporal evolution, with line variations indicating local expansion–contraction dynamics. Building on these cues, Fig. 1(b) summarizes how planar displacement and depth-induced scaling jointly contribute to motion, and how event projections help separate them for improved interpretation.

Motivated by these observations, this paper proposes an unsupervised event-based 3D motion estimation framework. We first derive the analytical relationship between optical flow and Motion in Depth (MID), revealing their scale dependency in the projection model and enabling an initial MID estimate to be inferred from flow. A Directional Expansion Modulation (DEM) module then refines this estimate by capturing horizontal and vertical expansion–contraction patterns in the x-t and y-t event projections. In the learning objective, MID is explicitly incorporated into the flow-based warping process to account for perspective scaling, where depth variations lead to near–far size changes. Optical flow and MID are jointly optimized under a contrast maximization framework, effectively reducing ambiguity between planar and depth-wise motion. By combining the estimated flow and MID to reconstruct scene flow, the proposed approach achieves highly competitive performance on the CarlaEvent3D dataset, demonstrating strong accuracy and generalization.

The contributions can be summarized as follows:

1. An unsupervised event-based 3D motion estimation framework is proposed to jointly model optical flow and motion in depth for complete scene motion representation.

2. A geometric analysis of the relationship between optical flow and motion in depth is conducted, and a directional expansion modulation module is introduced to refine motion in depth by capturing horizontal and vertical expansion–contraction patterns in event projections.

3. Motion in depth is incorporated into the optical flow warping process, explicitly modeling perspective scaling effects, and both optical flow and motion in depth are jointly optimized under a contrast maximization framework, enabling robust unsupervised learning.

4. Extensive experiments demonstrate the competitiveness of the proposed approach, validating its effectiveness and generalization in unsupervised 3D motion estimation.

2. Related Work

2.1. 3d Motion Estimation

Estimating the three-dimensional motion of scene points from two-dimensional observations has long been a fundamental problem in computer vision [31, 32]. Classical approaches rely on multi-view geometry or depth sensors to recover 3D motion fields [3, 36, 38]. More recently, learning-based methods have been proposed to directly regress 3D scene flow from image sequences [20, 30]. A paradigm shift was introduced by [37], which formulated normalized scene flow to represent 3D motion using optical flow and motion in depth jointly [19, 22, 23]. However, these methods typically rely on dense supervision or high-quality depth inputs, limiting their generalization to new environments.

To mitigate these issues, unsupervised approaches have been developed that jointly estimate scene flow by enforcing photometric and geometric consistency [2, 16, 17]. Nevertheless, they still rely on synchronized stereo pairs that require precise extrinsic calibration and remain vulnerable to geometric distortions and projection ambiguities inherent in image-based representations.

2.2. Event-based Unsupervised Motion Estimation

The advent of event cameras, which capture per-pixel brightness changes asynchronously with microsecond latency, has opened new possibilities for motion estimation [4, 9, 14, 21]. Early studies employed model-based approaches that relied on physical priors such as intensity consistency and local motion smoothness [5, 15, 24]. Gallego et al. [7] introduced the contrast maximization framework, which formulates event alignment as an optimization problem over optical flow, eliminating the need for local single-motion constraints and inspiring a series of subsequent unsupervised optical flow estimation methods [11, 13, 27]. Learning-based methods have since emerged, such as EV-FlowNet [39] and its extensions [12, 34, 40], which directly regress dense optical flow using contrast-based losses.

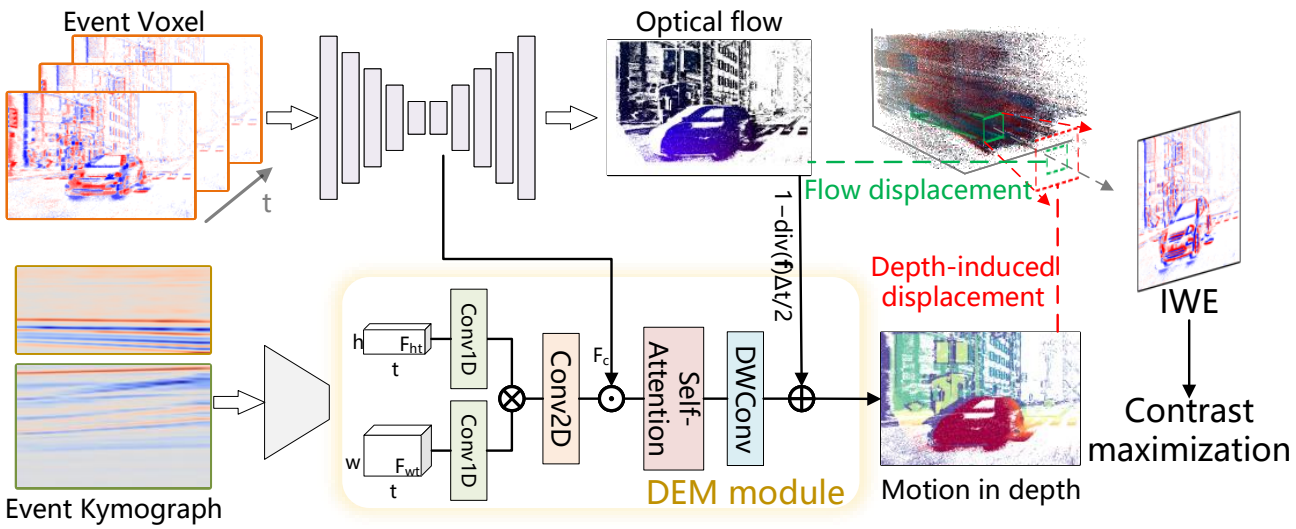

Figure 2. An overview of the event-based unsupervised 3D motion estimation framework. Event voxels are processed through a U-Net–like architecture to predict optical flow, from which a coarse motion-in-depth (MID) is derived according to Eq. (7). The coarse MID is then refined jointly with the event kymograph through the proposed DEM module. During the event warping process, the coordinate displacement induced by optical flow is illustrated by the green block, while the coordinate shift resulting from depth-induced scale variation is represented by the red block. The combined effect of these two components produces the Image of Warped Events (IWE), which is optimized via a contrast maximization loss.

While event-based vision has seen significant progress in motion estimation, including recent explorations into joint depth and ego-motion recovery [29], comprehensive 3D motion estimation from sparse event streams remains a formidable challenge. Our approach tackles this problem by deriving an analytical relationship between optical flow and motion in depth within the event domain, allowing the network to exploit the temporal continuity and geometric consistency of event patterns to infer and refine depth-related motion in a fully unsupervised manner.

3. Method

Event-based 3D motion estimation aims to recover the 3D motion (u, v, w) at each pixel from a sequence of events $\epsilon = \{x_i, y_i, t_i, p_i\}_{i=1}^{N}$, where (u, v) denotes the 2D optical flow and w represents the motion in depth (MID) along the viewing direction. In this work, the event stream is first converted into event voxels [39] and fed into a U-Net-like network to obtain multi-scale optical flow. A preliminary MID is then derived from the estimated flow and further refined using a directional expansion modulation module together with Event Kymograph [35] to capture depth-wise motion patterns. In the loss design, depth-induced scale variations are incorporated into coordinate warping and jointly constrained with a contrast maximization framework. The overall architecture is provided in Fig. 2.

The following sections first describe the event representation, then establish the relation between optical flow and MID under a pinhole camera model, based on which the network is designed to jointly estimate both. Finally, we detail the loss functions used to supervise flow and MID.

3.1. Event Representation

To capture the underlying scene structure, the event stream is aggregated into a voxelized representation on x–y plane [6], forming $V \in \mathbb{R}^{B\times H\times W}$, where H and W denote the sensor height and width, and B correspond to bins.

While preserving spatial information, voxelization inevitably compresses fine temporal dynamics. To retain temporal precision, events are also projected onto the x–t and y–t planes, yielding kymographs $K_x \in \mathbb{R}^{T\times W}$ and $K_y \in \mathbb{R}^{T\times H}$ [35], which decouple spatial dimensions while providing microsecond-level temporal resolution for detailed observation of dynamic scene evolution.

3.2. Flow and MID Modeling

We derive the relationship between optical flow and motion in depth in this section. Under the pinhole camera model, a 3D point (X, Y, Z) projects onto the 2D plane as::

$$x = f\frac{X}{Z}, \quad y = f\frac{Y}{Z}. \tag{1}$$

where f denotes the focal length. Assuming purely longitudinal motion along the viewing direction, the instantaneous optical flow satisfies:

$$u = -x\frac{\dot{Z}}{Z}, \quad v = -y\frac{\dot{Z}}{Z}. \tag{2}$$

Letting $\alpha(x, y) = \frac{\dot{Z}}{Z}$, the divergence of the optical flow field becomes:

$$\mathrm{div}(\mathbf{f}) = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = -2\alpha - (x\partial x\alpha + y\partial y\alpha). \quad (3)$$

In a local neighborhood where depth variations are small, the spatial derivative term can be neglected, yielding:

$$\frac{\dot{Z}}{Z} \approx -\frac{1}{2}\mathrm{div}(\mathbf{f}) \quad (4)$$

The above establishes a continuous relation between instantaneous depth variation and optical flow divergence. To extend this to a discrete temporal setting, we consider the depth ratio between two consecutive time instants separated by Δt, denoted as the motion in depth τ:

$$\tau = \frac{Z(t + \Delta t)}{Z(t)}. \quad (5)$$

By applying a first-order Taylor expansion to $Z(t + \Delta t)$ around t, the discrete ratio can be approximated as:

$$\tau \approx 1 + \frac{\dot{Z}}{Z}\,\Delta t. \quad (6)$$

Substituting Eq. (4) yields a discrete approximation linking motion in depth to optical flow divergence:

$$\tau \approx 1 - \frac{1}{2}\,\mathrm{div}(\mathbf{f})\,\Delta t. \quad (7)$$

This provides a practical bridge between the continuous motion field and its discrete observable counterpart, under simplified assumptions of locally rigid scene patches and dominant translational camera motion. These assumptions, while useful for deriving the analytical link between optical flow and depth variation, are later relaxed in our network design to better handle real-world motion complexity.

3.3. Network architecture

The network takes a sequence of event voxels as input and employs an encoder–decoder architecture with recurrent units, similar in [12], to extract multi-scale optical flow. The encoder captures hierarchical spatial features while preserving temporal memory across time, and the decoder fuses hidden states with upsampled features for refined flow estimation. From the highest-resolution flow, a coarse motion in depth map $\hat{\tau}$ is derived using Eq. (7). In parallel, a temporal encoder obtains motion cues F_{wt}, F_{ht} from event kymographs along the x-t and y-t planes.

To refine the coarse motion-in-depth (MID) estimates and relax the simplified motion constraints in Eq. (7), we propose the **Directional Expansion Modulation (DEM)** module. DEM leverages both spatial and temporal cues to adaptively enhance or suppress local MID responses, leading to more accurate and spatially coherent depth motion.

Specifically, DEM first extracts directional expansion rates from the temporal event projections F_{ht} and F_{wt}:

$$\begin{aligned} e_h &= \tanh\big(\mathrm{Conv}^{h}_{1\mathrm{D}}(F_{ht})\big), \\ e_w &= \tanh\big(\mathrm{Conv}^{w}_{1\mathrm{D}}(F_{wt})\big), \end{aligned} \quad (8)$$

where e_h and e_w are broadcast across spatial dimensions and concatenated to form a dual-axis expansion prior $E \in \mathbb{R}^{2\times H\times W}$. A lightweight 2-D projection then embeds E into the feature domain, producing an expansion-aware map that modulates the contextual features F_c extracted by the optical-flow encoder:

$$F_m = F_c \odot \mathrm{Conv}_{2\mathrm{D}}(E). \quad (9)$$

To propagate these directional cues beyond local neighborhoods, the modulated features F_m are processed by a compact self-attention block that aggregates long-range spatial context. This attentive output is subsequently passed through a set of dilated depthwise convolutions (DWConv) with dilation factors $\{1, 2, 4\}$, enabling directional information to be propagated across multiple spatial scales. The multi-dilation outputs are then fused to produce a dense MID residual map R, which captures both fine-grained and extended expansion behaviors. This residual term is added to the intermediate MID estimate $\hat{\tau}$, derived from optical flow via Eq. (7), producing the refined estimate $\tilde{\tau}$. Finally, the MID prediction is linearly rescaled as:

$$\tau = 0.75\,\tilde{\tau} + 1.25. \quad (10)$$

This rescaling adjusts the relative scale of local depth variations while maintaining numerical stability.

Through this refinement, the DEM combines contextual cues from the optical-flow encoder with direction-aware expansion priors, yielding a compact and effective mechanism for accurate, temporally consistent motion-in-depth estimation under complex dynamic conditions.

3.4. Self-supervised Loss with MID

To connect motion in depth with 2D-space deformation, we first consider how an object's projected area changes as its depth varies. From the pinhole camera model in Eq. (1), the 2D projected area A^{2d} and the real-world surface area A^{3d} are related by:

$$A^{2d} = \left(\frac{f}{Z}\right)^2 A^{3d}, \quad (11)$$

where Z denotes the object depth and f is the focal length. Given two time instants with depths Z_0 and Z_1, the corresponding 2D-space scale ratio can be expressed as

$$s = \sqrt{\frac{A_1^{2d}}{A_0^{2d}}} = \frac{Z_0}{Z_1} = \frac{1}{\mathrm{MID}}. \quad (12)$$

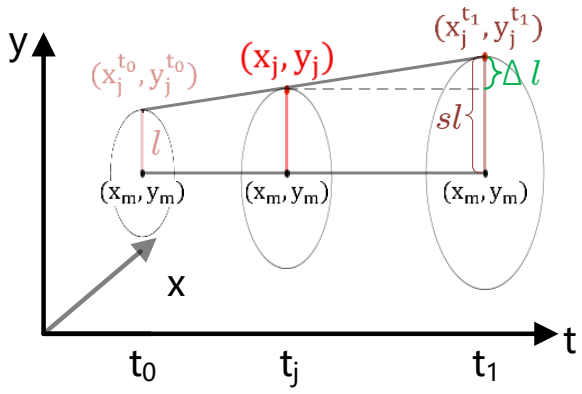

Figure 3. Illustration of depth-induced local expansion between two time instants t_0 and t_1. Each patch center (x_m, y_m) remains fixed, while events within the patch move outward as the scene depth decreases, resulting in an apparent scale change s. The displacement magnitude Δl represents the depth-induced expansion relative to the patch center when the event (x_j, y_j) is warped to t_1.

This scale ratio s represents the apparent magnification of local 2D regions due to depth variation.

Previous optical-flow-based formulations have incorporated such scale effects into the warping process by expanding or contracting local patches when computing photometric consistency [37]. However, directly applying this idea to event data is nontrivial. Unlike dense image grids, event streams consist of asynchronously triggered points, making it difficult to construct structured local patches. Moreover, improperly forcing spatial expansion or contraction directly on sparse events frequently triggers extreme geometric distortions and optimization degeneracies, widely known as event collapse [26, 28]. To address this, we reformulate the warping process at the event level.

In the contrast maximization framework, optical flow $\mathbf{f} = \{u, v\}$ warps all events toward a reference time t_r:

$$\begin{pmatrix} x_i^{t_r} \\ y_i^{t_r} \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} + (t_r - t_i) \begin{pmatrix} u(x_i, y_i) \\ v(x_i, y_i) \end{pmatrix}, \tag{13}$$

where (x_i, y_i, t_i) are the original event coordinates, and $(x_i^{t_r}, y_i^{t_r})$ are their warped positions at the reference time. After warping, all events are aggregated into an Image of Warped Events (IWE), where each pixel accumulates the timestamps of warped events falling within its spatial neighborhood. A well-aligned motion field leads to a sharp and high-contrast IWE, while misaligned motion produces temporal blur. Thus, network training aims to maximize the IWE contrast, encouraging events from the same physical edge to align temporally and spatially.

To incorporate depth-induced scale changes, we partition the 2D plane into non-overlapping patches. Given a global scale factor s derived from the motion in depth estimation according to Eq. (12), we introduce a time-dependent local scaling for each event to model gradual expansion or contraction over time. Specifically, as shown in Fig. 3, events within a local patch experience depth-induced scaling as the scene moves from t_0 to t_1, with the patch center (x_m, y_m) serving as a reference point. In practice, since events are asynchronously triggered, the depth-related scale change between t_0 and t_1 is linearly interpolated for each event according to its timestamp:

$$\lambda_j = \frac{t_1 - t_j}{t_1 - t_0}, \qquad s_j = 1 + \lambda_j (s - 1), \tag{14}$$

ensuring a smooth, time-continuous deformation model. Using this interpolated scale, the event's displacement toward t_1 is given by:

$$\begin{pmatrix} \Delta x_j^{t_1} \\ \Delta y_j^{t_1} \end{pmatrix} = \begin{pmatrix} (s_j - 1)x_j + (1 - s_j)x_m \\ (s_j - 1)y_j + (1 - s_j)y_m \end{pmatrix}, \tag{15}$$

where $\sqrt{(\Delta x_j^{t_1})^2 + (\Delta y_j^{t_1})^2}$ corresponds to the displacement magnitude Δl visualized in Fig. 3. By combining Eqs. (12), (13) and (15), the final warping position of each event is obtained by adding the depth-induced deformation to the optical-flow-based displacement, enabling a unified treatment of both translational and scale-varying motion. In practice, both forward and backward warpings (toward the end and initial time) are applied, resulting in slight notational differences but identical derivation logic.

Finally, the network is optimized under contrast-maximization objective, which measures the per-pixel temporal variance of warped timestamps, promoting temporally sharp and geometrically consistent IWE. A Charbonnier smoothness prior is additionally imposed to regularize neighboring motion estimates and preserve local coherence.

4. Experiment

4.1. Implementation Details

The CarlaEvent3D dataset contains six weather conditions: *Sunset*, *HardRain*, *Foggy*, *Night*, *Noon*, and *Cloudy*. Since the contrast maximization framework assumes constant illumination [7, 8, 25], where events are solely triggered by motion rather than lighting variations, we train the network exclusively on the *Sunset* sequences and evaluate it on all six weather conditions in the test set.

During training, each event window contains fixed 6000 events, which are voxelized on the x–y plane with 10 temporal bins. For the x–t and y–t projections, the temporal resolution is set to 120. The network is implemented in `PyTorch` and optimized using the Adam optimizer with an learning rate of $1e{-}4$. Gradient clipping is applied with a global norm of 100, and the model is trained for 100 epochs with a batch size of 4. Random horizontal, vertical, polarity

Table 1. Comparison of supervised (SL) and unsupervised (USL) methods across different weather conditions for 3D motion estimation. Our method achieves the best performance among unsupervised approaches and demonstrates strong generalization capability across diverse weather scenarios.

	Method	Sunset			Noon			Night		
		EPE↓	F1↓	log-mid↓	EPE↓	F1↓	log-mid↓	EPE↓	F1↓	log-mid↓
SL	E-RAFT [10]	2.022	17.164	-	3.160	23.977	-	2.356	20.229	-
	ScaleFlow [19]	3.295	41.323	271.081	4.253	42.271	262.894	3.271	32.508	280.086
	EMoTive [35]	1.852	19.223	106.938	2.846	24.594	147.086	2.008	20.023	104.734
USL	EV-FlowNet [39]	3.597	50.248	-	3.357	45.773	-	3.152	40.234	-
	Expansion [37]	8.499	56.627	774.114	8.188	52.158	749.043	7.963	49.501	771.360
	Ours	3.520	44.592	251.321	3.276	39.495	448.962	3.392	40.663	362.123
	Method	Cloudy			Foggy			Rainy		
		EPE↓	F1↓	log-mid↓	EPE↓	F1↓	log-mid↓	EPE↓	F1↓	log-mid↓
SL	E-RAFT [10]	2.709	23.458	-	3.031	29.436	-	3.104	30.100	-
	ScaleFlow [19]	3.711	44.098	278.209	5.943	43.855	264.882	6.615	45.278	251.929
	EMoTive [35]	2.629	24.163	116.434	3.054	26.587	142.342	3.125	31.263	144.123
USL	EV-FlowNet [39]	2.925	39.052	-	3.427	38.858	-	3.340	39.932	-
	Expansion [37]	8.583	59.856	769.685	7.655	47.022	740.813	8.149	57.436	720.142
	Ours	2.833	34.832	244.841	3.269	38.254	333.544	3.239	37.001	426.262

flips and crops from 320×960 to 320×320 are used for data augmentation. A backward update is performed every five forward passes to stabilize training.

4.2. 3D Motion Estimation

Metrics. Evaluation follows the protocol of [35]. The Average End-Point Error (EPE) measures the pixel-wise displacement error, while the F1-outlier ratio (F1) quantifies the percentage of outlier pixels beyond a predefined threshold. Since motion in depth values typically vary around unity, a logarithmic error metric (log-mid) is adopted to emphasize relative rather than absolute deviations, providing a more balanced evaluation of near–far motion changes.

Quantitative results. Table 1 summarizes the results on the CarlaEvent3D dataset. All methods take event voxels as input for fair comparison across all evaluated settings. Among unsupervised approaches, our method significantly outperforms the optical expansion baseline [37] on all weather conditions, highlighting the benefit of enforcing geometric consistency in the event domain. The optical expansion method estimates motion in depth via local affine transformations between projections, which are unreliable under asynchronous event sampling and complex motion. In contrast, our formulation captures scale variation continuously without explicit patch correspondences, producing smoother depth motion. Compared with unsupervised optical flow networks such as EV-FlowNet [39], our approach achieves lower EPE by explicitly modeling the coupling between planar and depth-wise motion.

While supervised methods achieve lower errors overall due to the use of ground-truth labels, our unsupervised framework attains comparable accuracy even when trained only on the *Sunset* sequence, demonstrating strong generalization across unseen scenes and illumination conditions.

Qualitative analysis. Section 10 presents qualitative comparisons of 3D motion estimation results under two weather conditions. The optical expansion baseline estimates motion in depth by fitting local affine transformations between two temporal projections in the event domain. However, due to the spatial sparsity of events, these local correspondences are often ambiguous, leading to noisy and distorted depth motion fields especially in low-texture regions. In contrast, our method leverages the high temporal resolution of events to disentangle planar and depth-wise motion components within a unified framework that enforces geometric consistency. As a result, it yields smoother and more accurate 3D motion predictions across diverse conditions.

4.3. Scene Flow Estimation

Metrics. The models is evaluated using the Average 3D End-Point Error (EPE_{3D}) and the Accuracy-at-10cm ($\text{ACC}_{0.1}$) metrics. EPE_{3D} measures the average deviation of the estimated 3D motion, reflecting the overall accuracy of the model on a global scale. In contrast, $\text{ACC}_{0.1}$ computes the proportion of points whose 3D error is below 10 cm, thus emphasizing the model's ability to recover fine-grained local motion precisely.

Quantitative results. As shown in Tab. 2, Despite similar EPE_{3D} scores to the Optical Expansion baseline, our

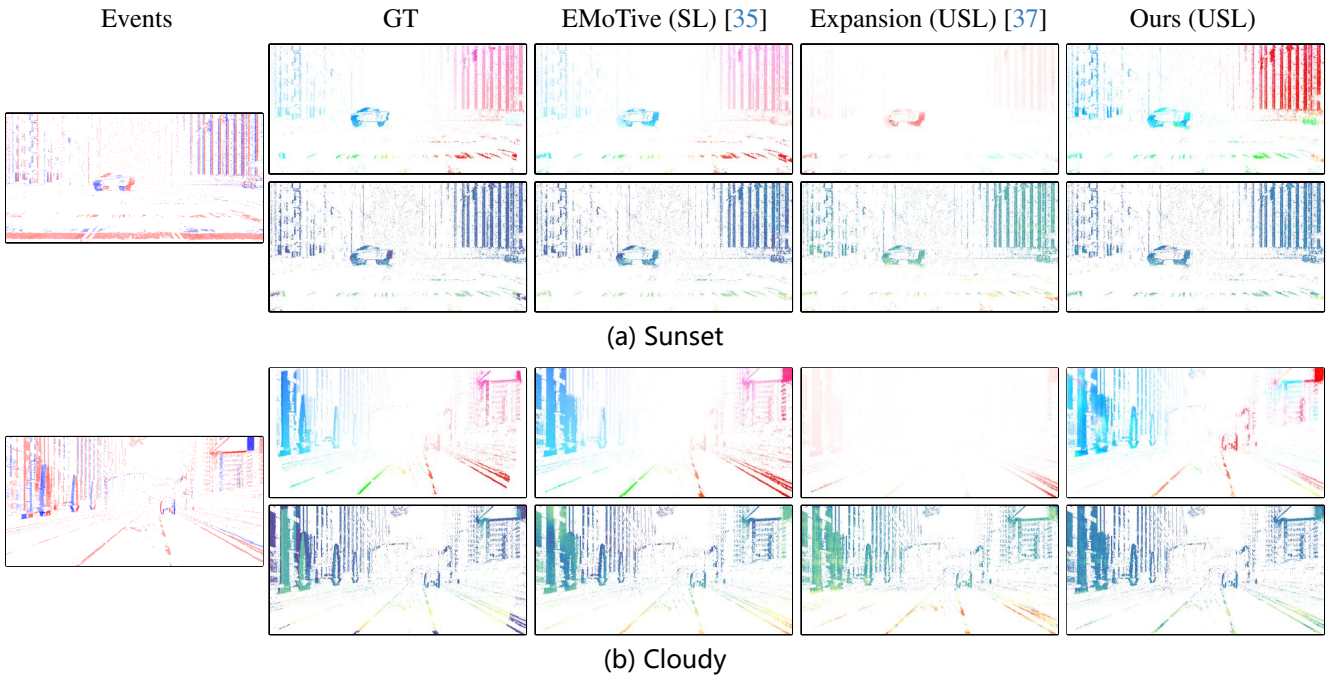

Figure 4. Visualization of 3D motion estimation under two weather conditions. For each sample, the leftmost column shows the event voxel, the first row illustrates the optical flow, and the second row presents the motion in depth. Both SL and USL methods are included for completeness. While the SL model produces cleaner motion fields owing to direct supervision, our unsupervised approach achieves visually consistent coherent results without using ground-truth labels, and maintains robustness across diverse illumination conditions.

Table 2. Scene flow estimation results under different weather conditions. Scene flow estimation results under different weather conditions. Our method achieves notably better performance than other unsupervised approach on the $ACC_{0.1}$ metric.

	Method	Sunset		Noon		Night	
		$EPE_{3D}\downarrow$	$Acc_{0.1}\uparrow$	$EPE_{3D}\downarrow$	$Acc_{0.1}\uparrow$	$EPE_{3D}\downarrow$	$Acc_{0.1}\uparrow$
SL	EMoTive	0.176	43.8%	0.200	41.0%	0.186	42.9%
USL	Expansion	0.812	2.3%	1.057	3.2%	1.690	1.4%
	Ours	1.062	12.7%	1.081	13.2%	1.130	13.2%
	Method	Cloudy		Foggy		Rainy	
		$EPE_{3D}\downarrow$	$Acc_{0.1}\uparrow$	$EPE_{3D}\downarrow$	$Acc_{0.1}\uparrow$	$EPE_{3D}\downarrow$	$Acc_{0.1}\uparrow$
SL	EMoTive	0.182	41.6%	0.208	37.6%	0.222	37.1%
USL	Expansion	0.793	3.1%	1.024	2.2%	1.272	2.4%
	Ours	1.084	13.5%	1.089	12.8%	1.058	14.3%

method attains a 6–7× improvement in $ACC_{0.1}$, reflecting a tighter and more consistent error distribution. This improvement stems from the proposed DEM, which refines motion in depth estimation by adapting the observation direction of event streams.

Qualitative analysis. Figure 5 shows results on two sequences under different weather conditions. Red points denote the previous scene, while blue points represent the warped scene obtained from the predicted scene flow, with darker colors indicating larger depths. The Optical Expansion method, constrained by its local affine formulation and the irregular timing of event samples, produces clearly divergent warped points, causing the underlying 3D structure to break down—most noticeably around the building facades. In contrast, our approach enforces temporally coherent geometric constraints at the event level, yielding stable and consistent warped geometry that better preserves the scene structure.

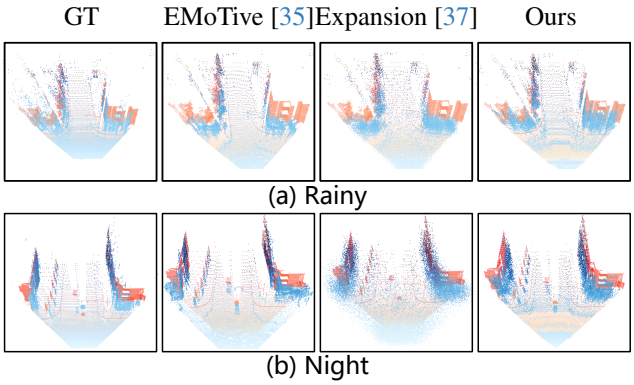

Figure 5. Scene flow visualization. Red points denote the initial scene, while blue points indicate the scene after scene flow warping. Darker colors correspond to greater depth.

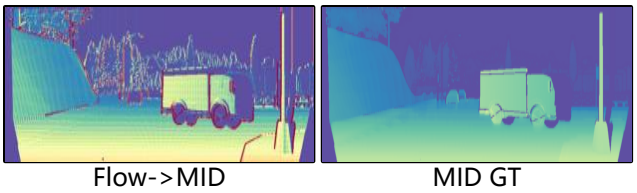

Figure 6. Visualization of motion in depth derived from Eq. (7).

Method	EPE	log-mid
✘	3.38	644.56
✔	**3.29**	**364.01**

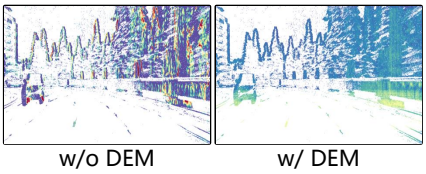

Figure 7. Effect of the DEM module. Left: performance averages across different weathers. Right: motion in depth visualization.

4.4. Ablation Study

To validate the rationality of our architectural design, we conduct ablation studies on the key components of our pipeline, namely the theoretical derivation from optical flow to motion in depth and the DEM module.

From optical flow to motion in depth. Figure 6 shows the motion in depth derived from Eq. (7). Compared with the ground truth on the right, the errors primarily occur in near-range ground regions and along object boundaries. Near-range ground regions exhibit the largest depth gradients due to perspective projection: although the ground plane itself is flat, its depth changes rapidly with image-space displacement near the bottom of the image, making the neglected spatial-derivative term non-negligible in Eq. (3). Depth also changes sharply at object boundaries, leading to the same effect and resulting in additional errors.

DEM module. DEM refines the motion in depth inferred from optical flow by altering the observation direction of events. Figure 7 illustrates the results after removing DEM, where the predicted MID becomes noticeably discontinuous. Figure 7 presents the corresponding quantitative results. The reduced MID accuracy further affects the integrated coordinate warping based on depth-motion and optical-flow cues, which in turn causes a slight degradation in the final optical-flow accuracy.

4.5. Limitations

Similar to previous unsupervised contrast-maximization frameworks, our method relies on the brightness-constancy assumption, which causes the model to struggle when events are triggered not by motion but instead by illumination changes. This limitation results in degraded performance under *noon* and *rainy* conditions, as shown in Tab. 1. In the former, the sun frequently enters the camera's field of view, and in the latter, reflections from accumulated water on the ground as well as raindrops introduce additional flickering events, both of which lead to illumination-induced noise that the model cannot reliably suppress.

Because the near-ground region in front of the vehicle contains very little texture, only a few of events are triggered and thus provide insufficient motion cues. This causes event-by-event optimization methods to produce inaccurate 3D motion estimates, which further propagate to the subsequent scene flow reconstruction. As shown in Fig. 5(a), points close to the camera (i.e., near the bottom of the image) exhibit noticeably larger positional errors, reflecting the adverse impact of sparse event observations

5. Conclusion

We introduce an unsupervised event-based framework for estimating 3D motion by jointly modeling optical flow and motion in depth. Building on the geometric relationship we derive between optical flow and depth-wise motion, our method obtains an initial MID estimate directly from flow and further refines it using a Directional Expansion Modulation module that leverages expansion–contraction patterns in event projections. Incorporating MID into the flow-warping process reduces ambiguity between planar and depth-induced motion under a contrast maximization objective. Experiments on CarlaEvent3D demonstrate competitive accuracy and strong generalization, underscoring the effectiveness of event-based sensing for unsupervised 3D motion estimation and highlighting the advantages of jointly reasoning over flow and motion in depth.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62436008, 62306295, and 62576328. The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

[1] Abhishek Badki, Orazio Gallo, Jan Kautz, and Pradeep Sen. Binary ttc: A temporal geofence for autonomous navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12946–12955, 2021. 1

[2] Bayram Bayramli, Junhwa Hur, and Hongtao Lu. Raft-msf: Self-supervised monocular scene flow using recurrent optimizer. *International Journal of Computer Vision*, 131(11): 2757–2769, 2023. 2

[3] Jan Čech, Jordi Sanchez-Riera, and Radu Horaud. Scene flow estimation by growing correspondence seeds. In *CVPR 2011*, pages 3129–3136. IEEE, 2011. 2

[4] Jinze Chen, Wei Zhai, Han Han, Tiankai Ma, Yang Cao, Bin Li, and Zheng-Jun Zha. Unbiased gradient estimation for event binning via functional backpropagation. *arXiv preprint arXiv:2602.12590*, 2026. 2

[5] Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *The 2011 International Joint Conference on Neural Networks*, pages 770–776. IEEE, 2011. 2

[6] Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1172–1181, 2022. 3

[7] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *CVPR*, pages 3867–3876, 2018. 2, 5

[8] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12280–12289, 2019. 5

[9] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 2

[10] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 197–206. IEEE, 2021. 6

[11] Shuang Guo, Friedhelm Hamann, and Guillermo Gallego. Unsupervised joint learning of optical flow and intensity with event cameras. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[12] Jesse Hagenaars, Federico Paredes-Vallés, and Guido De Croon. Self-supervised learning of event-based optical flow with spiking neural networks. *Advances in Neural Information Processing Systems*, 34:7167–7179, 2021. 2, 4

[13] Friedhelm Hamann, Ziyun Wang, Ioannis Asmanis, Kenneth Chaney, Guillermo Gallego, and Kostas Daniilidis. Motion-prior contrast maximization for dense continuous-time motion estimation. In *European Conference on Computer Vision (ECCV)*, pages 18–37, 2024. 2

[14] Han Han, Wei Zhai, Yang Cao, Bin Li, and Zheng-jun Zha. Mate: Motion-augmented temporal consistency for event-based point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2025. 2

[15] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[16] Junhwa Hur and Stefan Roth. Self-supervised monocular scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7396–7405, 2020. 2

[17] Junhwa Hur and Stefan Roth. Self-supervised multi-frame monocular scene flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2684–2694, 2021. 2

[18] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European conference on computer vision*, pages 349–364. Springer, 2016. 1

[19] Yuyang Leng, Renyuan Liu, Hongpeng Guo, Songqing Chen, and Shuochao Yao. Scaleflow: Efficient deep vision pipeline with closed-loop scale-adaptive inference. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1698–1706, 2023. 1, 2, 6

[20] Bing Li, Cheng Zheng, Silvio Giancola, and Bernard Ghanem. Sctn: Sparse convolution-transformer network for scene flow estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1254–1262, 2022. 2

[21] Bohao Liao, Wei Zhai, Zengyu Wan, Zhixin Cheng, Wenfei Yang, Tianzhu Zhang, Yang Cao, and Zheng-Jun Zha. Ef-3dgs: Event-aided free-trajectory 3d gaussian splatting. *arXiv preprint arXiv:2410.15392*, 2024. 2

[22] Han Ling and Quansen Sun. Scaleflow++: Robust and accurate estimation of 3d motion from video. *arXiv preprint arXiv:2407.09797*, 2024. 1, 2

[23] Han Ling, Yinghui Sun, Quansen Sun, and Zhenwen Ren. Learning optical expansion from scale matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5445–5454, 2023. 1, 2

[24] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981. 2

[25] Federico Paredes-Vallés, Kirk YW Scheper, Christophe De Wagter, and Guido CHE De Croon. Taming contrast maximization for learning sequential, low-latency, event-based optical flow. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9695–9705, 2023. 5

[26] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Event collapse in contrast maximization frameworks. *Sensors*, 22(14):5190, 2022. 5

[27] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *European Conference on Computer Vision*, pages 628–645. Springer, 2022. 2

[28] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. A fast geometric regularizer to mitigate event collapse in the contrast maximization framework. *Advanced Intelligent Systems*, 5(3):2200251, 2023. 5

[29] Shintaro Shiba, Yannick Klose, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow, depth and ego-motion estimation by contrast maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7742–7759, 2024. 3

[30] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8375–8384, 2021. 2

[31] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 722–729. IEEE, 1999. 2

[32] Sundar Vedula, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):475–480, 2005. 2

[33] Zhexiong Wan, Yuxin Mao, Jing Zhang, and Yuchao Dai. Rpeflow: Multimodal fusion of rgb-pointcloud-event for joint optical flow and scene flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10030–10040, 2023. 1

[34] Zengyu Wan, Ganchao Tan, Yang Wang, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Event-based optical flow via transforming into motion-dependent view. *IEEE Transactions on Image Processing*, 33:5327–5339, 2024. 2

[35] Zengyu Wan, Wei Zhai, Yang Cao, and Zhengjun Zha. Emotive: Event-guided trajectory modeling for 3d motion estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9342–9351, 2025. 1, 3, 6, 7, 8

[36] Andreas Wedel, Clemens Rabe, Tobi Vaudrey, Thomas Brox, Uwe Franke, and Daniel Cremers. Efficient dense scene flow from sparse or dense stereo data. In *European conference on computer vision*, pages 739–751. Springer, 2008. 2

[37] Gengshan Yang and Deva Ramanan. Upgrading optical flow to 3d scene flow through optical expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1334–1343, 2020. 2, 5, 6, 7, 8

[38] Haimei Zhao, Jing Zhang, Zhuo Chen, Bo Yuan, and Dacheng Tao. On robust cross-view consistency in self-supervised monocular depth estimation. *Machine Intelligence Research*, 21(3):495–513, 2024. 2

[39] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018. 2, 3, 6

[40] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019. 2