

BAT: Learning Event-based Optical Flow with Bidirectional Adaptive Temporal Correlation

Gangwei Xu^{1,2}, Haotong Lin³, Zhaoxing Zhang¹, Hongcheng Luo², Haiyang Sun², Xin Yang^{1*}

¹Huazhong University of Science and Technology

²Xiaomi EV

³Zhejiang University

Abstract

Event cameras deliver visual information characterized by a high dynamic range and high temporal resolution, offering significant advantages in estimating optical flow for complex lighting conditions and fast-moving objects. Current advanced optical flow methods for event cameras largely adopt established image-based frameworks. However, the spatial sparsity of event data limits their performance. In this paper, we present BAT, an innovative framework that estimates event-based optical flow using bidirectional adaptive temporal correlation. BAT includes three novel designs: 1) a bidirectional temporal correlation that transforms bidirectional temporally dense motion cues into spatially dense ones, enabling accurate and spatially dense optical flow estimation; 2) an adaptive temporal sampling strategy for maintaining temporal consistency in correlation; 3) spatially adaptive temporal motion aggregation to efficiently and adaptively aggregate consistent target motion features into adjacent motion features while suppressing inconsistent ones. Our BAT achieves state-of-the-art performance on the DSEC-Flow benchmark, outperforming existing methods by a large margin while also exhibiting sharp edges and high-quality details. Our BAT can accurately predict future optical flow using only past events, significantly outperforming E-RAFT’s warm-start approach.

Code — <https://github.com/gangweix/BAT>

Introduction

Optical flow has been an enduring fundamental task in the field of computer vision, which can be applied to a broad range of practical tasks, including video reconstruction (Tulyakov et al. 2021, 2022; Xu et al. 2024b) and visual odometry (Zhang et al. 2024; Shiba et al. 2024), and robotics (Xu et al. 2025a; Liang et al. 2025a,b; Li et al. 2025; Lin et al. 2025). Image-based optical flow estimation (Dosovitskiy et al. 2015; Teed and Deng 2020; Wang, Lipson, and Deng 2025) has demonstrated remarkable performance on public benchmarks (Butler et al. 2012; Menze and Geiger 2015; Mehl et al. 2023). Nonetheless, these methods struggle with motion blur and over-exposed or under-exposed regions due to the limited dynamic range of conventional

*Corresponding author.

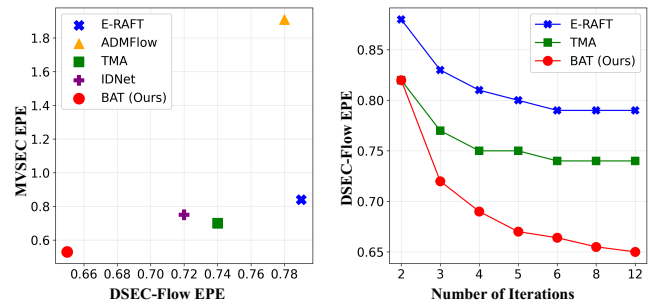


Figure 1: **Left:** Comparison with state-of-the-art event-based flow methods on DSEC-Flow and MVSEC benchmarks. Our BAT achieves the highest accuracy among all published methods. **Right:** Compared to E-RAFT and TMA, our BAT shows much higher efficiency in iterative optimization. Additionally, the aforementioned methods converge at a high error rate quickly due to limited unidirectional motion cues. In contrast, our BAT fully utilizes bidirectional and rich motion cues, allowing for continuous optimization for better results.

image sensors. Event cameras present a promising alternative for optical flow estimation. Compared to image-based cameras, event cameras continuously capture asynchronous brightness changes, providing high temporal resolution without motion blur (microseconds), along with a high dynamic range and low power consumption. Therefore, event-based optical flow estimation is more advantageous for fast-moving objects, high dynamic range scenes, and resource-constrained devices.

However, the data captured by event cameras is spatially sparse, irregular, and asynchronous, making it unsuitable for directly applying current state-of-the-art image-based flow methods to estimate event-based optical flow. To leverage established convolutional neural network architectures (Dosovitskiy et al. 2015; Ilg et al. 2017; Teed and Deng 2020), recent works (Zhu and Yuan 2018; Gehrig et al. 2021b) transform consecutive event streams into grid-like/tensor representations and input them into these networks to predict the optical flow.

A representative method is E-RAFT (Gehrig et al. 2021b). Inspired by RAFT (Teed and Deng 2020), E-RAFT employs

image-based optical flow architectures for event-based optical flow estimation. It first converts two event sequences into grid-like/tensor representations, constructs a correlation volume between the two representations, and then regresses the optical flow from the correlation volume. The correlation volume, which measures visual similarity between pixels in two images, has become an essential component in mainstream image-based/frame-based optical flow estimation methods. However, unlike dense and rich RGB or grayscale image data, event data is spatially sparse, resulting in the constructed correlation volume with insufficient motion cues, which significantly limits the performance of this approach. To provide more motion cues, TMA (Liu et al. 2023) computes the correlation volumes between the reference event frame and multiple forward event frames. Event data is inherently sequential. Although TMA has achieved improved results, it relies solely on forward temporal motion cues and overlooks backward temporal motion cues, which limits its overall performance (see Figure 1).

Our goal is to transform bidirectional temporally dense motion cues into spatially dense ones, thereby enabling detailed and accurate optical flow estimation. Our model is built on two key insights. First, events and motion are both temporally continuous and fine-grained, so forward and backward temporal motion cues are crucial for accurate optical flow estimation. Second, motion can vary non-uniformly over time, indicating the need to adaptively aggregate both forward and backward motion cues to ensure temporal consistency.

Building on these insights, we present bidirectional adaptive temporal (BAT) correlation for event-based optical flow estimation. Specifically, we split the reference event stream and target event stream into multiple groups, then compute bidirectional temporal correlation (BTC): forward correlation is applied between the reference event frame and multiple target event frames to obtain forward motion cues, while backward temporal correlation is performed between multiple reference event frames to capture backward motion cues. BTC transforms temporally dense motion cues into spatially dense ones, allowing for accurate and spatially dense optical flow estimation. We also propose an adaptive temporal sampling strategy for temporal consistency during computing correlation.

Linearly indexing adjacent event frames across different time spans and performing correlation can introduce inconsistent motion patterns when dealing with non-uniformly moving objects. To address this challenge, we propose spatially adaptive temporal motion aggregation (SATMA). Specifically, SATMA first applies deformable and sparse attention between target motion features and adjacent motion features to aggregate relevant target motion features within the regions of interest. It then generates an attention map to adaptively fuse the aggregated target motion features into the adjacent motion features, effectively suppressing inconsistent motion features and enhancing consistent ones.

Extensive experiments demonstrate that our approach outperforms previous methods by a large margin on public benchmarks. Specifically, our BAT ranks 1st on the DSEC-Flow (Gehrig et al. 2021a) benchmark, exceeding E-

RAFT (Gehrig et al. 2021b) and TMA (Liu et al. 2023) by 39.45% and 28.98%, respectively, for the IPE metric.

Additionally, the proposed backward temporal correlation offers two notable advantages: first, it enables us to predict future optical flow using only past events, a capability crucial for interpreting and interacting with dynamic environments, as shown in Figure 6; second, it effectively handles occlusions caused by objects moving out of the target frame, as shown in Figure 2.

To summarize, our main contributions are as follows:

- We present a novel framework, BAT, that learns to estimate event-based optical flow with bidirectional adaptive temporal correlation.
- We propose spatially adaptive temporal motion aggregation that efficiently integrates temporally consistent target motion features into adjacent motion features while suppressing inconsistent ones.
- Our BAT achieves state-of-the-art performance on the DSEC-Flow benchmark, outperforming all previously published methods.
- Our BAT can accurately predict future optical flow using only past events, while also effectively handling occlusions caused by objects moving out of the target frame.

Related Work

Deep learning has demonstrated remarkable success in image matching tasks (Xu et al. 2023b, 2025b; Cheng et al. 2024a, 2025; Wang et al. 2025; Xu et al. 2024a; Cheng et al. 2024b). The pioneering flow method, FlowNet (Dosovitskiy et al. 2015), introduced an end-to-end convolutional neural network that directly takes two images as input and outputs optical flow, achieving impressive performance. Subsequently, network architectures (Teed and Deng 2020; Huang et al. 2022) and training strategies (Wang, Lipson, and Deng 2025; Sun et al. 2022a) for optical flow estimation have been continuously improved, leading to consistent advancements in performance. For example, PWC-Net (Sun et al. 2018, 2019) and LiteFlowNet (Hui, Tang, and Loy 2018, 2020) introduced pyramid and warping-based cost volumes to estimate optical flow in a coarse-to-fine manner. These methods significantly reduce inference time while enhancing accuracy. RAFT presents a novel deep network architecture for optical flow, which first builds a 4D correlation volume for all pairs of pixels, and then iteratively regresses the optical flow from it. Combined with its new training strategy, RAFT sets a new state-of-the-art in optical flow estimation.

Inspired by RAFT, several novel designs, such as global attention (Jiang et al. 2021a), kernel patch attention (Luo et al. 2022), and super kernels (Sun et al. 2022b), have been proposed to improve the performance of motion aggregation. Recently, many works (Huang et al. 2022; Xu et al. 2022; Lu et al. 2023; Sui et al. 2022; Weinzaepfel et al. 2023) have introduced transformers into optical flow estimation to enhance global matching. These methods have all led to improvements in accuracy. Efforts (Xu et al. 2021; Zhao, Xu, and Wu 2024; Jiang et al. 2021b; Xu et al. 2023a) have also been made to overcome the excessive memory

consumption introduced by the high-resolution 4D correlation volume in RAFT. Instead of focusing on a two-frame setting, VideoFlow (Shi et al. 2023) concurrently estimates bidirectional optical flows for multiple consecutive frames. In contrast to VideoFlow, our approach builds on the sparse nature of event data and transforms the rich temporal motion cues into the target frame, focusing on estimating its motion. UnFlow (Meister, Hur, and Roth 2018) designs an unsupervised loss on bidirectional optical flow computed from two static images. In contrast, our method computes bidirectional temporal correlation for more accurate and detailed forward optical flow prediction.

Learning-based event-based flow methods also largely draw upon well-established image-based optical flow architectures. Moreover, adapting to these architectures often requires converting events into tensor representations, like event frames or voxel grids. These learning-based methods can be categorized into supervised (Gehrig et al. 2021b; Liu et al. 2023; Wu, Paredes-Vallés, and De Croon 2024), semi-supervised (Zhu and Yuan 2018; Lee et al. 2020; Ding et al. 2022), and unsupervised (Paredes-Vallés et al. 2023; Paredes-Vallés and De Croon 2021; Hagenaaers, Paredes-Vallés, and De Croon 2021; Zhu et al. 2019). Unsupervised methods rely solely on event data; semi-supervised methods use grayscale images as a supervisory signal; supervised methods require accurate ground truth flow for supervision. Currently, supervised methods dominate on standard benchmarks (Gehrig et al. 2021a; Zhu et al. 2018). A representative approach is E-RAFT (Gehrig et al. 2021b), which introduces the RAFT (Teed and Deng 2020) architecture into event-based optical flow estimation. Subsequently, TMA (Liu et al. 2023) extends the two-frame correlation volume of E-RAFT to a multi-frame correlation volume, providing more temporal motion cues. ADMFlow (Luo et al. 2023) renders high-quality datasets to facilitate event-flow learning. Without using a correlation volume, IDNet (Wu, Paredes-Vallés, and De Croon 2024) introduces iterative deblurring for accurate event-based flow estimation.

Contrary to existing learning-based event-based methods that only model unidirectional motion cues, we propose bidirectional adaptive temporal correlations to estimate spatially detailed optical flow. More importantly, by modeling backward motion cues, our method can predict future optical flow using only past events.

Method

Conventional optical flow estimation is typically defined as finding pixel correspondences between two image frames. However, unlike frame-based images, event data is recorded in a sequential and asynchronous manner by detecting per-pixel brightness changes. Therefore, given an event stream $\mathcal{E}(t_i, t_{i+1})$, the goal of event-based optical flow is to find pixel correspondences between two timestamps t_i and t_{i+1} .

Building on the spatial sparsity and temporal density of event data, we propose BAT, a novel framework that transforms temporally dense motion cues into spatially dense ones, enabling spatially dense and accurate optical flow estimation. The overview of the proposed network framework is illustrated in Figure 3. Before feeding the event data into

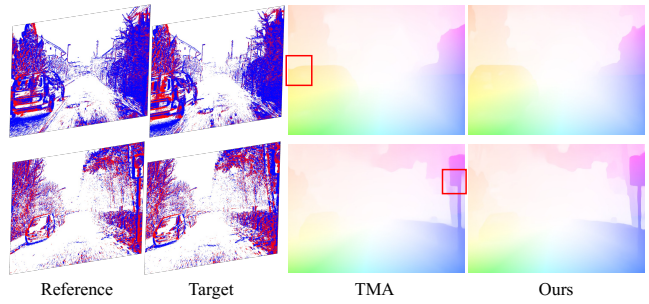


Figure 2: Benefiting from the proposed backward temporal correlation, our method effectively handles occlusions caused by objects moving out of the target frame.

the network, the event stream is first converted into an event voxel grid representation. In this section, we first describe the event voxel grid representation and then detail the network architecture.

Event Representation

Event cameras are novel bio-inspired vision sensors that generate an asynchronous event stream when the brightness change of a pixel exceeds a predefined threshold C . An event $e_k(t) = (x_k, y_k, t, p_k)$ is a tuple containing trigger pixel location (x_k, y_k) , the timestamp t_k and the polarity p_k that indicates whether the brightness change is positive or negative. The event stream $\mathcal{E}(t_i, t_{i+1})$ from timestamp t_i to timestamp t_{i+1} is defined as $\mathcal{E}(t_i, t_{i+1}) := \langle e(t) | t \in [t_i, t_{i+1}] \rangle$.

To ensure compatibility with CNNs and facilitate correlation volume construction, we follow previous works (Gehrig et al. 2021b; Liu et al. 2023) by transforming the event stream $\mathcal{E}(t_i, t_{i+1})$ into an event voxel grid $\mathbf{V} \in \mathbb{R}^{B \times H_0 \times W_0}$ with discretized time dimension B . Specifically, given an event stream $\mathcal{E}(t_i, t_{i+1})$, $\mathbf{V}(b, x, y)$ is obtained by:

$$t_j^* = (B - 1)(t_j - t_i) / (t_{i+1} - t_i)$$

$$\mathbf{V}(b, x, y) = \sum_j p_j k_b(x - x_j) k_b(y - y_j) k_b(b - t_j^*) \quad (1)$$

$$k_b(a) = \max(0, 1 - |a|),$$

where B denotes the number of time bins, and $b \in [0, B - 1]$ is the integer bin index.

Feature Extraction

Given two consecutive event streams $\mathcal{E}(t_{i-1}, t_i)$ and $\mathcal{E}(t_i, t_{i+1})$, where $t_{i+1} - t_i = t_i - t_{i-1} = \Delta t$, our goal is to estimate the optical flow $\mathbf{f}_{t_i \rightarrow t_{i+1}}$ from timestamp t_i to t_{i+1} . Firstly, we convert the event streams into voxel grid representations, denoted as \mathbf{V}_0 and \mathbf{V}_1 , respectively, using Equ. 1.

To exploit temporally dense motion cues, we split both \mathbf{V}_0 and \mathbf{V}_1 into N groups equally along the time bin dimension, and each group contains B/N time bins. Then we input the voxel grid event representations of these $2N$ groups into a shared-weight feature extraction network to produce a series of features \mathbf{F}_n ($\mathbf{F}_n \in \mathbb{R}^{D \times H \times W}$), $n = 1, 2, \dots, 2N$, where

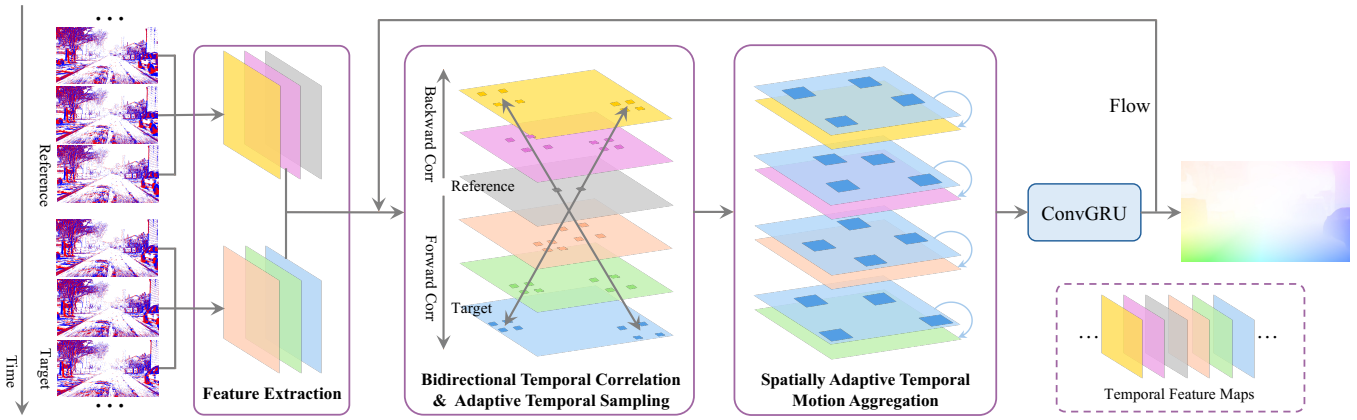


Figure 3: Network architecture of the proposed BAT.

D , H and W denote the channel, height, and width of feature F_n . The feature extraction network consists of 6 residual blocks, similar to RAFT (Teed and Deng 2020). Among these temporal features, F_N serves as the reference frame, while F_{2N} serves as the target frame.

Bidirectional Temporal Correlation

Correlation volume, which measures pixel-wise visual similarities between frames, provides crucial information for flow estimation. However, existing flow methods often focus on the two-frame setting (Gehrig et al. 2021b) or unidirectional correlation (Liu et al. 2023). In contrast, we propose bidirectional temporal correlation (BTC) to transform temporally rich motion cues into spatially dense ones. BTC consists of forward temporal correlation and backward temporal correlation. Forward temporal correlation computes the correlation between the reference frame F_N and its next frames F_{N+j} ($j = 1, \dots, N$), and backward temporal correlation computes the correlation between the reference frame F_N and its previous frames F_{N-j} ($j = 1, 2, \dots, N-1$).

Given the current flow estimate f from timestamp t_i to t_{i+1} , in order to facilitate the computation of temporal correlations, we assume that the optical flow is linear with respect to time. Thus, we can derive the optical flow df between adjacent temporal groups as:

$$df = f/N \quad (2)$$

Then, the forward temporal correlation $C_j^{fwd} \in \mathbb{R}^{(2r+1)^2 \times H \times W}$ is computed by:

$$C_j^{fwd} = \frac{F_N \cdot \mathcal{W}(F_{N+j}; jdf)}{\sqrt{D}}, j = 1, 2, \dots, N, \quad (3)$$

where \cdot denotes the dot product, $\mathcal{W}(\cdot; \cdot)$ denotes the warping of adjacent frames to the reference frame using optical flow. Specifically, for each pixel p in F_N , we use the current flow jdf to find its corresponding point p' in F_{N+j} , then define a local grid around p' within a radius r and bilinearly sample the values from the local grid. The number of samples for the local grid is $(2r+1)^2$.

Similarly, the backward temporal correlation $C_j^{bwd} \in \mathbb{R}^{(2r+1)^2 \times H \times W}$ is computed by:

$$C_j^{bwd} = \frac{F_N \cdot \mathcal{W}(F_{N-j}; (-j)df)}{\sqrt{D}}, j = 1, 2, \dots, N-1. \quad (4)$$

Finally, we obtain N groups of forward temporal correlation maps and $N-1$ groups of backward temporal correlation maps. Through this bidirectional temporal correlation, we transform temporally dense motion cues into spatially dense ones, enabling spatially detailed and accurate optical flow estimation. In particular, the backward temporal correlation maps are highly beneficial for handling occlusions of objects moving out of the target frame, as shown in Figure 2.

Adaptive Temporal Sampling Strategy. The sampling radius r is set as a hyperparameter in previous works (Gehrig et al. 2021b; Liu et al. 2023). However, as motion patterns change over time, a manually chosen r may not be robust for temporal consistent correlation. To address this, we introduce a learnable parameter α as a scale factor to obtain a learnable sampling radius lr ,

$$lr = \alpha \cdot r. \quad (5)$$

The lr can be adaptively learned during the training process, while the number of sampling points in the local grid remains unchanged.

Spatially Adaptive Temporal Motion Aggregation

Effectively utilizing temporally rich motion cues is crucial for achieving spatially dense and accurate optical flow estimation. However, due to the non-uniformity of optical flow over time, linearly computing correlations on adjacent temporal frames introduces motion features inconsistent with the target frame. To address this, we propose a spatially adaptive temporal motion aggregation (SATMA) module that adaptively enhances motion features consistent with the target frame while suppressing inconsistent ones. The architecture of SATMA is illustrated in the supplementary material.

Before feeding into STAMA, we encode the correlation

	Method	EPE↓	1PE↓	2PE↓	3PE↓	AE↓
MB	MultiCM	3.472	76.570	48.480	30.855	13.983
	E-RAFT	0.788	12.742	4.740	2.684	2.851
	ADMFlow	0.779	12.522	4.673	2.647	2.838
SL	EEMFlow+	0.751	11.403	3.932	2.145	2.669
	TMA	0.743	10.863	3.972	2.301	2.684
	IDNet	0.719	10.069	3.497	2.036	2.723
	BAT (Ours)	0.655	7.715	2.896	1.773	2.439

Table 1: DSEC (Gehrig et al. 2021a) benchmark evaluation. MB refers to model-based methods that do not require training data, and SL represents supervised learning methods that require ground truth optical flow. **Bold**: best.

feature C_j^{fwd} into motion feature M_j^{fwd} :

$$M_j^{fwd} = \text{MotionEncoder}(C_j^{fwd}, \mathbf{f}), j = 1, 2, \dots, N, \quad (6)$$

Similarly, we can obtain the backward motion feature M_j^{bwd} ($j = 1, 2, \dots, N - 1$). Among them, M_N^{fwd} represents the target motion feature, while the others are adjacent motion features.

The architecture of the spatially adaptive temporal motion aggregation (SATMA) module is presented in the supplementary material. Given the target motion feature M_N^{fwd} and an adjacent motion feature M_j^{fwd} (or M_j^{bwd}), we first concatenate them, then pass them through a convolutional layer followed by a sigmoid activation function to generate a spatial attention map A_{spa} . The spatial attention effectively enhances consistent motion features while suppressing inconsistent ones. We also adopt deformable attention (Xia et al. 2022) to efficiently integrate consistent or similar motion features from M_N^{fwd} into M_j^{fwd} . Specifically, M_j^{fwd} is first projected linearly to obtain the query Q , which is then fed into a lightweight offset network to generate sparse sampling locations. Based on these locations, we can sample M_N^{fwd} to obtain the key K and value V . Then we obtain the aggregated motion feature M_j^{agg} through attention:

$$M_j^{agg} = \sigma\left(\frac{QK^T}{\sqrt{D_m}}\right)V, \quad (7)$$

where $\sigma(\cdot)$ denotes the softmax function, and D_m is the dimension of the motion feature. Finally, we obtain fused motion feature M_j^{fuse} by:

$$M_j^{fuse} = A_{spa} \odot M_j^{agg} + M_j^{fwd}, \quad (8)$$

where \odot denotes the element-wise product. Similarly, we also perform SATMA between backward motion features and the target motion feature to obtain the backward fused motion features. Then, these forward and backward fused motion features are fed into the ConvGRU to update the optical flow.

Loss Function

Following RAFT, we supervised our network using the l_1 loss between the predicted and ground truth flow over the

	Method	$dt = 1$		$dt = 4$	
		EPE↓	%Out↓	EPE↓	%Out↓
MB	MultiCM	0.30	0.10	1.25	9.21
	EV-FlowNet	0.49	0.20	1.23	7.30
SSL	Spike-FlowNet	0.49	-	1.09	-
	STE-FlowNet	0.42	0.00	0.99	3.90
	E-RAFT	0.27	0.00	0.84	1.70
	DCEIFlow	0.91	0.71	1.87	19.1
SL	ADMFlow	0.52	0.00	1.91	19.2
	TMA	0.25	0.07	0.70	1.08
	IDNet	0.29	0.00	0.75	1.20
	BAT (Ours)	0.21	0.00	0.53	0.71

Table 2: MVSEC (Zhu et al. 2018) benchmark evaluation. SSL denotes semi-supervised learning methods that use grayscale images for supervision.

Model	EPE↓	1PE↓	2PE↓	3PE↓	AE↓
Baseline	0.712	9.123	3.332	2.033	2.544
BTC	0.680	8.279	3.130	1.889	2.527
BTC+ATS	0.671	8.179	3.014	1.804	2.482
Full (BAT)	0.655	7.715	2.896	1.773	2.439

Table 3: Ablation study on DSEC-Flow benchmark. BTC represents bidirectional temporal correlation, ATS represents adaptive temporal sampling. Full model includes SATMA, denoting spatially adaptive temporal motion aggregation.

entire sequence of predictions, $\{\mathbf{f}^1, \mathbf{f}^2, \dots, \mathbf{f}^K\}$, applying exponentially increasing weights. Given the ground truth flow \mathbf{f}^{gt} , the loss is defined as,

$$\mathcal{L} = \sum_{i=1}^K \gamma^{K-i} \|\mathbf{f}^i - \mathbf{f}^{gt}\|_1, \quad (9)$$

where \mathbf{f}^i denotes the predicted optical flow at the i -th iteration, and γ balances the weights of the loss terms for each iteration. K is the total number of iterations.

Experiments

Experimental Setup

Datasets and evaluation setups. Following previous work, we conduct extensive experiments on two popular event-based datasets: DSEC-Flow (Gehrig et al. 2021a) and MVSEC (Zhu et al. 2018). For DSEC-Flow, we train the models on the official training dataset and then evaluate them on the public benchmark. DSEC-Flow provides 8,170 training samples and 416 testing samples, with a resolution of 640×480 . For MVSEC, we perform training and evaluation on two types of event data, $dt = 4$ and $dt = 1$, which are divided based on time intervals.

Implementation details. We implement our model with PyTorch and perform our experiments using NVIDIA RTX

Experiment	Method	EPE↓	IPE↓	3PE↓
Attention	dense	0.670	8.049	1.868
	spatial-reduction	0.698	8.731	2.044
	<u>deformable</u>	0.655	7.715	1.773
Radius	r=0	0.677	8.02	1.985
	r=1	0.667	7.981	1.945
	<u>r=2</u>	0.655	7.715	1.773
Iterations	4	0.689	8.563	2.011
	<u>8</u>	0.655	7.715	1.773
	<u>12</u>	0.650	7.540	1.740

Table 4: Ablation experiments on DSEC-Flow benchmark. Settings used in our final model are underlined.

Input	Method	EPE↓	IPE↓
Tgt.+Ref. event	MultiCM	3.472	76.570
	TamingCM	2.330	68.293
	EV-FlowNet	2.320	55.400
only Ref. event	E-RAFT (warm-start)	4.518	85.378
	BAT (bwd corr)	1.163	33.026

Table 5: Future optical flow prediction. Ref. denotes reference event stream $\mathcal{E}(t_{i-1}, t_i)$, and Tgt. denotes target event stream $\mathcal{E}(t_i, t_{i+1})$. Our BAT (bwd corr) can predict future optical flow $\mathbf{f}_{t_i \rightarrow t_{i+1}}$ using only the past event stream $\mathcal{E}(t_{i-1}, t_i)$.

3090 GPUs. We set the temporal group N to 3 in the final BAT model. For experiments on DSEC-Flow, we set the event representation to $B = 15$ time bins for every 100ms of events. For MVSEC, B is set to 5 when $dt = 1$ and 15 when $dt = 4$. The length B of time bins is kept consistent with E-RAFT and TMA. The final channel dimension D for feature extraction is 128, and the flow updater is identical to that of RAFT. The sampling radius r is set to 2. For all training, we use the AdamW optimizer with a one-cycle learning rate schedule, setting the learning rate to $2e-4$. We use 8 update iterations during training, and the weight γ is set to 0.8. For DSEC-Flow, we train our BAT for 200k steps with a batch size of 8, while for MVSEC, we train it for 100k steps.

Comparisons with State-of-the-art

DSEC-Flow. Table 1 presents the comparison results on the DSEC-Flow benchmark. Our BAT ranks 1st among all published methods for all metrics. Compared to other iterative methods, E-RAFT (Gehrig et al. 2021b) and TMA (Liu et al. 2023), our BAT improves the IPE metric by 39.45% and 28.98%, respectively. The IPE metric is the most challenging evaluation metric, as it directly reflects the detailed performance of the predictions.

Qualitative results on the DSEC-Flow benchmark are shown in Figure 5. Due to the spatial sparsity of event data, previous methods (Gehrig et al. 2021b; Liu et al. 2023) tend to produce predictions with blurred edges and loss of details. In contrast, our BAT leverages bidirectional temporal



Figure 4: Future optical flow prediction. The first row illustrates events from timestamp t_{i-1} to t_i , while the second row presents the optical flow results from t_i to t_{i+1} . Given only past events, our method can predict future optical flow.

motion cues, enabling it to predict sharp edges and preserve fine structures. Additionally, benefiting from the proposed backward temporal correlation, our BAT effectively handles occlusions caused by objects moving out of the target frame, as shown in Figure 2.

MVSEC. Following previous works, we perform training on outdoor_day2 sequence and evaluate on outdoor_day1 sequence. Table 2 presents the evaluation results corresponding to $dt = 1$ grayscale frame and $dt = 4$ grayscale frame. Our method achieves the highest prediction accuracy and outperforms other methods by a large margin. Specifically, for the more challenging $dt = 4$ setting, our BAT surpasses the second-best method, TMA (Liu et al. 2023), by 34.26% for the %Out metric and 24.29% for the EPE metric.

Ablation Study

We conduct ablation studies to validate the effectiveness of the proposed components. All ablation models are trained on the DSEC-Flow training dataset and evaluated on the DSEC-Flow benchmark, as presented in Table 3. We use the TMA (Liu et al. 2023) with $N = 3$ temporal groups as the Baseline. Baseline only computes forward temporal correlation, neglecting the backward motion cues. In contrast, the proposed bidirectional temporal correlation (BTC) computes both forward and backward temporal correlations, enabling the model to leverage bidirectional temporally dense motion cues for accurate and spatially detailed optical flow estimation. As a result, BTC achieves a significant improvement in accuracy. Furthermore, we introduce an adaptive temporal sampling (ATS) strategy to ensure temporal consistency when computing correlations, which enhances the model’s performance.

Due to the presence of non-uniform motion, simply sampling the non-target event frames based on a linear motion assumption can introduce inconsistent temporal motion features. To mitigate this issue, we propose spatially adaptive temporal motion aggregation (SATMA), which efficiently and adaptively aggregates consistent target motion features while suppressing inconsistent motion features. As shown in Table 3, SATMA significantly improves the model’s prediction accuracy.

Attention type. To perform temporal motion aggregation effectively, we compare three attention types (Table 4). Dense

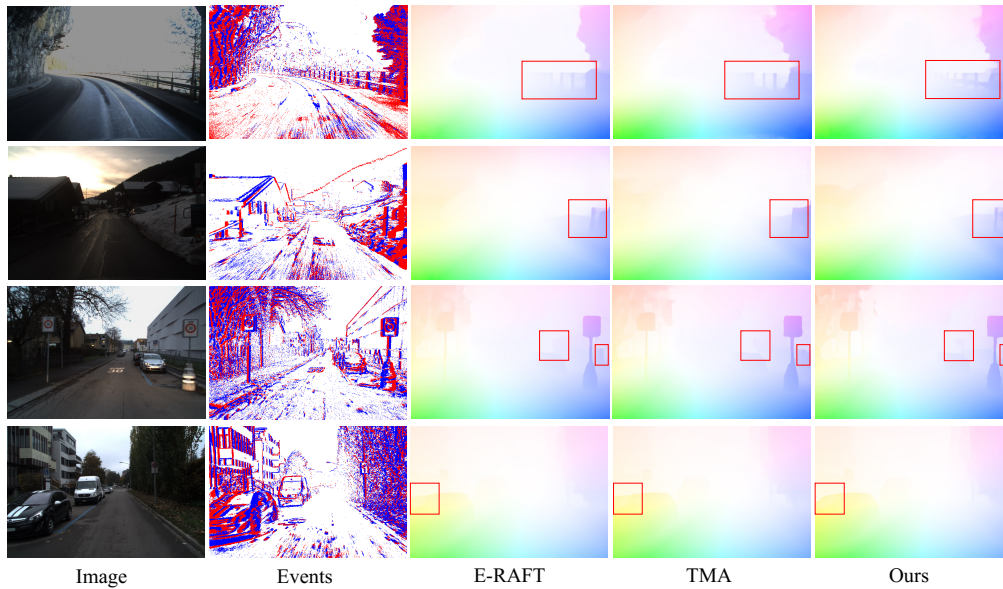


Figure 5: Qualitative results of optical flow predictions on DSEC-Flow. Significant improvements are highlighted by red boxes. Images are provided for visualization only, since the optical flow is event-based. Our method accurately distinguishes subtle details and sharp edges. Zoom in for a clearer view.

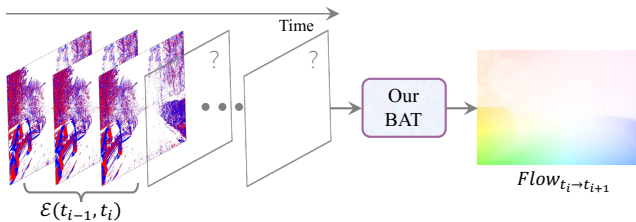


Figure 6: Predicting future optical flow $f_{t_i \rightarrow t_{i+1}}$ using only past events $\mathcal{E}(t_{i-1}, t_i)$.

attention incurs high memory/computation and introduces irrelevant features. Spatial-reduction attention lowers the cost by pooling K and V , but sacrifices detail and degrades accuracy. In contrast, deformable attention efficiently focuses on relevant motion features with low overhead.

Sampling radius. Due to the rich temporal motion cues, even with using a sampling radius of 0, computing correlation at a single point, our method can still yield good results, as shown in Table 4. Better results are obtained with an increased radius.

Iterations. By effectively integrating bidirectional and rich temporal motion cues, our results continuously improve with more iterations, as shown in Table 4 and Figure 1.

Future Optical Flow Prediction

Predicting future optical flow is a crucial task for interpreting and interacting with dynamic environments, such as assisting drones or autonomous vehicles in navigating through complex and dynamic environments. Previous methods (Zhu and Yuan 2018; Gehrig et al. 2021b; Liu et al. 2023) require the given $\mathcal{E}(t_i, t_{i+1})$ to predict flow $f_{t_i \rightarrow t_{i+1}}$, mak-

ing them difficult to predict future optical flow. Benefiting from the proposed backward temporal correlation, our model can predict future optical flow $f_{t_i \rightarrow t_{i+1}}$ using only the past event stream $\mathcal{E}(t_{i-1}, t_i)$, shown in Figure 6. Our BAT (bwd corr) produces promising optical flow results on the DSEC-Flow benchmark, significantly outperforming the model-based method MultiCM, the self-supervised method TamingCM, and E-RAFT’s warm-start approach that initializes the flow for the next frame, as shown in Table 5. Visual results are shown in Figure 4.

Conclusion

In this paper, we present BAT, an innovative framework designed to estimate accurate and detailed optical flow from events. BAT introduces three key innovations: a bidirectional temporal correlation, an adaptive temporal sampling strategy, and spatially adaptive temporal motion aggregation. Our approach ranks 1st on the DSEC-Flow benchmark, outperforming existing state-of-the-art methods by a large margin, while also preserving sharp edges and high-quality details. Additionally, our BAT can accurately predict future optical flow using only past events. We believe that our simple method, BAT, will become the baseline for the community and drive the future development of event-based optical flow.

Limitations. Our BAT does not bring significant gains under rapid motion changes. For example, when the event camera shakes violently, there is a significant difference between backward and forward temporal motion. In this case, backward temporal motion cues provide little help for forward optical flow estimation.

Acknowledgments

This research is supported by the National Key R&D Program of China (2024YFE0217700), the National Natural Science Foundation of China (623B2036,62472184), the Fundamental Research Funds for the Central Universities, and the Innovation Project of Optics Valley Laboratory (Grant No. OVL2025YZ005).

References

- Butler, D. J.; Wulff, J.; Stanley, G. B.; and Black, M. J. 2012. A naturalistic open source movie for optical flow evaluation. In *Eur. Conf. Comput. Vis.*, 611–625. Springer.
- Cheng, J.; Liu, L.; Xu, G.; Wang, X.; Zhang, Z.; Deng, Y.; Zang, J.; Chen, Y.; Cai, Z.; and Yang, X. 2025. Monster: Marry monodepth to stereo unleashes power. In *CVPR*, 6273–6282.
- Cheng, J.; Xu, G.; Guo, P.; and Yang, X. 2024a. Coatsrnet: Fully exploiting convolution and attention for stereo matching by region separation. *IJCV*, 132(1): 56–73.
- Cheng, J.; Yin, W.; Wang, K.; Chen, X.; Wang, S.; and Yang, X. 2024b. Adaptive fusion of single-view and multi-view depth for autonomous driving. In *CVPR*, 10138–10147.
- Ding, Z.; Zhao, R.; Zhang, J.; Gao, T.; Xiong, R.; Yu, Z.; and Huang, T. 2022. Spatio-temporal recurrent networks for event-based optical flow estimation. In *AAAI*, volume 36, 525–533.
- Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; and Brox, T. 2015. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2758–2766.
- Gehrig, M.; Aarents, W.; Gehrig, D.; and Scaramuzza, D. 2021a. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3): 4947–4954.
- Gehrig, M.; Millhäusler, M.; Gehrig, D.; and Scaramuzza, D. 2021b. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, 197–206. IEEE.
- Hagenaars, J.; Paredes-Vallés, F.; and De Croon, G. 2021. Self-supervised learning of event-based optical flow with spiking neural networks. *NIPS*, 34: 7167–7179.
- Huang, Z.; Shi, X.; Zhang, C.; Wang, Q.; Cheung, K. C.; Qin, H.; Dai, J.; and Li, H. 2022. Flowformer: A transformer architecture for optical flow. In *Eur. Conf. Comput. Vis.*, 668–685. Springer.
- Hui, T.-W.; Tang, X.; and Loy, C. C. 2018. LiteflowNet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, 8981–8989.
- Hui, T.-W.; Tang, X.; and Loy, C. C. 2020. A lightweight optical flow cnn—revisiting data fidelity and regularization. *IEEE TPAMI*, 43(8): 2555–2569.
- Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2462–2470.
- Jiang, S.; Campbell, D.; Lu, Y.; Li, H.; and Hartley, R. 2021a. Learning to estimate hidden motions with global motion aggregation. In *ICCV*, 9772–9781.
- Jiang, S.; Lu, Y.; Li, H.; and Hartley, R. 2021b. Learning optical flow from a few matches. In *CVPR*, 16592–16600.
- Lee, C.; Kosta, A. K.; Zhu, A. Z.; Chaney, K.; Daniilidis, K.; and Roy, K. 2020. Spike-flowNet: event-based optical flow estimation with energy-efficient hybrid neural networks. In *ECCV*, 366–382. Springer.
- Li, Y.; Xiong, K.; Guo, X.; Li, F.; Yan, S.; Xu, G.; Zhou, L.; Chen, L.; Sun, H.; Wang, B.; et al. 2025. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint arXiv:2506.08052*.
- Liang, D.; Feng, T.; Zhou, X.; Zhang, Y.; Zou, Z.; and Bai, X. 2025a. Parameter-efficient fine-tuning in spectral domain for point cloud learning. *IEEE TPAMI*.
- Liang, D.; Hua, W.; Shi, C.; Zou, Z.; Ye, X.; and Bai, X. 2025b. Sood++: Leveraging unlabeled data to boost oriented object detection. *IEEE TPAMI*.
- Lin, H.; Peng, S.; Chen, J.; Peng, S.; Sun, J.; Liu, M.; Bao, H.; Feng, J.; Zhou, X.; and Kang, B. 2025. Prompting depth anything for 4k resolution accurate metric depth estimation. In *CVPR*, 17070–17080.
- Liu, H.; Chen, G.; Qu, S.; Zhang, Y.; Li, Z.; Knoll, A.; and Jiang, C. 2023. TMA: Temporal motion aggregation for event-based optical flow. In *ICCV*, 9685–9694.
- Lu, Y.; Wang, Q.; Ma, S.; Geng, T.; Chen, Y. V.; Chen, H.; and Liu, D. 2023. Transflow: Transformer as flow learner. In *CVPR*, 18063–18073.
- Luo, A.; Yang, F.; Li, X.; and Liu, S. 2022. Learning optical flow with kernel patch attention. In *CVPR*, 8906–8915.
- Luo, X.; Luo, K.; Luo, A.; Wang, Z.; Tan, P.; and Liu, S. 2023. Learning optical flow from event camera with rendered dataset. In *ICCV*, 9847–9857.
- Mehl, L.; Schmalfluss, J.; Jahedi, A.; Nalivayko, Y.; and Bruhn, A. 2023. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *CVPR*, 4981–4991.
- Meister, S.; Hur, J.; and Roth, S. 2018. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, volume 32.
- Menze, M.; and Geiger, A. 2015. Object scene flow for autonomous vehicles. In *CVPR*, 3061–3070.
- Paredes-Vallés, F.; and De Croon, G. C. 2021. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *CVPR*, 3446–3455.
- Paredes-Vallés, F.; Scheper, K. Y.; De Wagter, C.; and De Croon, G. C. 2023. Taming contrast maximization for learning sequential, low-latency, event-based optical flow. In *ICCV*, 9695–9705.
- Shi, X.; Huang, Z.; Bian, W.; Li, D.; Zhang, M.; Cheung, K. C.; See, S.; Qin, H.; Dai, J.; and Li, H. 2023. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *ICCV*, 12469–12480.

- Shiba, S.; Klose, Y.; Aoki, Y.; and Gallego, G. 2024. Secrets of event-based optical flow, depth and ego-motion estimation by contrast maximization. *IEEE TPAMI*, 1–18.
- Sui, X.; Li, S.; Geng, X.; Wu, Y.; Xu, X.; Liu, Y.; Goh, R.; and Zhu, H. 2022a. Craft: Cross-attentional flow transformer for robust optical flow. In *CVPR*, 17602–17611.
- Sun, D.; Herrmann, C.; Reda, F.; Rubinstein, M.; Fleet, D. J.; and Freeman, W. T. 2022a. Disentangling architecture and training for optical flow. In *ECCV*, 165–182. Springer.
- Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 8934–8943.
- Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2019. Models matter, so does training: An empirical study of cnns for optical flow estimation. *IEEE TPAMI*, 42(6): 1408–1423.
- Sun, S.; Chen, Y.; Zhu, Y.; Guo, G.; and Li, G. 2022b. Sk-flow: Learning optical flow with super kernels. *NIPS*, 35: 11313–11326.
- Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 402–419. Springer.
- Tulyakov, S.; Bochicchio, A.; Gehrig, D.; Georgoulis, S.; Li, Y.; and Scaramuzza, D. 2022. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *CVPR*, 17755–17764.
- Tulyakov, S.; Gehrig, D.; Georgoulis, S.; Erbach, J.; Gehrig, M.; Li, Y.; and Scaramuzza, D. 2021. Time lens: Event-based video frame interpolation. In *CVPR*, 16155–16164.
- Wang, X.; Yang, H.; Xu, G.; Cheng, J.; Lin, M.; Deng, Y.; Zang, J.; Chen, Y.; and Yang, X. 2025. ZeroStereo: Zero-shot Stereo Matching from Single Images. *arXiv preprint arXiv:2501.08654*.
- Wang, Y.; Lipson, L.; and Deng, J. 2025. Sea-raft: Simple, efficient, accurate raft for optical flow. In *ECCV*, 36–54. Springer.
- Weinzaepfel, P.; Lucas, T.; Leroy, V.; Cabon, Y.; Arora, V.; Brégier, R.; Csurka, G.; Antsfeld, L.; Chidlovskii, B.; and Revaud, J. 2023. CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *ICCV*, 17969–17980.
- Wu, Y.; Paredes-Vallés, F.; and De Croon, G. C. 2024. Lightweight event-based optical flow estimation via iterative deblurring. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 14708–14715. IEEE.
- Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In *CVPR*, 4794–4803.
- Xu, G.; Chen, S.; Jia, H.; Feng, M.; and Yang, X. 2023a. Memory-efficient optical flow via radius-distribution orthogonal cost volume. *arXiv preprint arXiv:2312.03790*.
- Xu, G.; Lin, H.; Luo, H.; Wang, X.; Yao, J.; Zhu, L.; Pu, Y.; Chi, C.; Sun, H.; Wang, B.; et al. 2025a. Pixel-perfect depth with semantics-prompted diffusion transformers. *arXiv preprint arXiv:2510.07316*.
- Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023b. Iterative geometry encoding volume for stereo matching. In *CVPR*, 21919–21928.
- Xu, G.; Wang, X.; Zhang, Z.; Cheng, J.; Liao, C.; and Yang, X. 2025b. Igev++: Iterative multi-range geometry encoding volumes for stereo matching. *IEEE TPAMI*.
- Xu, G.; Wang, Y.; Cheng, J.; Tang, J.; and Yang, X. 2024a. Accurate and Efficient Stereo Matching via Attention Concatenation Volume. *IEEE TPAMI*, 46(4): 2461–2474.
- Xu, G.; Wang, Y.; Gu, J.; Xue, T.; and Yang, X. 2024b. HDRFlow: Real-Time HDR Video Reconstruction with Large Motions. In *CVPR*, 24851–24860.
- Xu, H.; Yang, J.; Cai, J.; Zhang, J.; and Tong, X. 2021. High-resolution optical flow from 1d attention and correlation. In *ICCV*, 10498–10507.
- Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; and Tao, D. 2022. Gmflow: Learning optical flow via global matching. In *CVPR*, 8121–8130.
- Zhang, Z.; Cheng, J.; Xu, G.; Wang, X.; Zhang, C.; and Yang, X. 2024. Leveraging Consistent Spatio-Temporal Correspondence for Robust Visual Odometry. *arXiv preprint arXiv:2412.16923*.
- Zhao, Y.; Xu, G.; and Wu, G. 2024. Hybrid Cost Volume for Memory-Efficient Optical Flow. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8740–8749.
- Zhu, A. Z.; Thakur, D.; Özaslan, T.; Pfrommer, B.; Kumar, V.; and Daniilidis, K. 2018. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robotics and Automation Letters*, 3(3): 2032–2039.
- Zhu, A. Z.; and Yuan, L. 2018. EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras. In *Robotics: Science and Systems*.
- Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2019. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, 989–997.