

# DEIO: Deep Event Inertial Odometry

Weipeng Guan\* Fuling Lin\* Peiyu Chen Peng Lu✉  
ArcLab, The University of Hong Kong

{wpguan, fuling, chenpyhk}@connect.hku.hk lupeng@hku.hk

## Abstract

*Event cameras show great potential for visual odometry (VO) in handling challenging situations, such as fast motion and high dynamic range. Despite this promise, the sparse and motion-dependent characteristics of event data continue to limit the performance of feature-based or direct-based data association methods in practical applications. To address these limitations, we propose Deep Event Inertial Odometry (DEIO), the first monocular learning-based event-inertial framework, which combines a learning-based method with traditional nonlinear graph-based optimization. Specifically, an event-based recurrent network is adopted to provide accurate and sparse associations of event patches over time. DEIO further integrates it with the IMU to recover up-to-scale pose and provide robust state estimation. The Hessian information derived from the learned differentiable bundle adjustment (DBA) is utilized to optimize the co-visibility factor graph, which tightly incorporates event patch correspondences and IMU pre-integration within a keyframe-based sliding window. Comprehensive validations demonstrate that DEIO achieves superior performance on 10 challenging public benchmarks compared with more than 20 state-of-the-art methods. We release the source code and qualitative results: <https://kwanwaipang.github.io/DEIO/>.*

## 1. Introduction

Event cameras are motion-activated sensors that only capture pixel-wise intensity changes with microsecond precision and report them as an asynchronous stream instead of the whole scene as an intensity image with a fixed frame rate. Due to their remarkable properties, such as high temporal resolutions, high dynamic range (HDR), and no motion blur, event cameras have the potential to enable high-quality perception in extreme lighting conditions and high-speed motion scenarios that are currently not accessible to standard cameras.

Despite such promises, integrating event cameras into visual odometry (VO) systems presents significant challenges. This is primarily due to the sparse, irregular, and asynchronous nature of event data, which conveys limited information and contains inherent noise. Moreover, due to the motion-dependent characteristic, both feature-based and direct-based methods easily fail in incomplete observation or sudden variation of the event camera. Therefore, current purely event-based VO systems [28, 36] generally lack the robustness requirement for real-world applications. Recent studies have introduced learning-based approaches [19, 25] as promising solutions for event-based VO, addressing the previously mentioned limitations by employing neural networks to establish robust associations. However, it is worth noting that visual/event-only systems have inherent limitations, making them vulnerable to low-textured environments or suffering from scale ambiguity. To mitigate visual degradations, a practical and promising strategy is to incorporate the Inertial Measurement Unit (IMU), which is low-cost and readily available in event cameras. Nevertheless, integrating an event-based network with IMU remains an unexplored territory due to the challenge of efficiently fusing the learning-based event association with IMU.

In this work, we propose Deep Event Inertial Odometry (DEIO), the first deep learning-based event-inertial odometry framework. It is developed based on a learning-optimization framework that leverages neural networks to predict event correspondences and tightly integrates IMU measurements to enhance the robustness of the odometry. More specifically, our framework decouples the neural network from IMU integration and operates in two phases: (i) an event-based recurrent network learns to provide robust data associations of sparse event patches. (ii) The network is tightly integrated with the IMU measurements within a factor graph optimization framework to achieve 6-DoF pose tracking. Our contributions are summarized as follows:

1. We propose a learning-optimization framework that seamlessly integrates the power of deep learning with the efficiency of factor graph optimization. To the best of our knowledge, this is the first event-inertial odometry framework that employs deep learning for event data

\*These authors contributed equally to this work.

- association and graph optimization for pose estimation.
2. An event-based co-visibility factor graph optimization is proposed to tightly integrate event patch correspondences and IMU pre-integration by deriving Hessian information from differentiable bundle adjustment (DBA).
  3. Extensive experiments on 10 challenging event-based real-world benchmarks demonstrate the superior performance of DEIO compared to over 20 advanced methods. We release the source code to facilitate further research.

## 2. Related Work

### 2.1. Traditional Event-based VO

To enhance the robustness of purely event-based VO, existing event-based SLAM methods have demonstrated good performance by incorporating additional sensors. Notably, event-inertial integration is a widely used approach to address the limitations of event-only SLAM, which provides scale awareness and continuity of estimation with minimal setup requirements. Zhu et al. [40] propose the first event-inertial odometry (EIO) method, which fuses events with IMU through the Extended Kalman Filter. Rebecq et al. [29] propose an optimization-based EIO that detects and tracks features in the edge image, generated from motion-compensated event streams, through traditional image-based feature detection and tracking. The tracked features are then combined with IMU measurements via keyframe-based nonlinear optimization. Mono-EIO [13] employs the event-corner features with IMU measurement to deliver real-time and accurate 6-DoF state estimation. Ultimate-SLAM [33] and PL-EVIO [14] investigate the complementary nature of events and images to present an event-image-IMU odometry (EVIO). ESvio [4] proposes the first stereo EIO and EVIO framework to estimate states through temporally and spatially event-corner feature association. ESVO2 [23, 24] extends ESVO [36] and presents a direct method for stereo event cameras with an IMU-aided solution. EVI-SAM [15] introduces the first event-based hybrid pose tracking framework, merging feature-based and direct-based methods, with IMU fusion.

### 2.2. Learning Event-based VO

Zhu et al. [39] pioneer the first learning-based event odometry framework, utilizing an unsupervised network with a contrast maximization loss [7]. Ye et al. [34] extend the SfMLearner [35], which employs a depth network and pose network for event-based optical flow estimation. However, these methods show poor generalization beyond the training scenarios. DEVO [19] extends the DPVO [32] to accommodate the event modality also through the voxel-based representation like E-RAFT [10], demonstrating great generalization from synthetic data to seven real-world event-based benchmarks. RAMP-VO [25] introduces an end-to-

end VO that also builds upon DPVO [32] using feature encoders to fuse events and image data. However, the absolute scale is not observable in these monocular event-only systems. Visual-IMU integration is the most common solution to address these limitations, which provides scale awareness and continuous estimation with minimal setup. Nevertheless, efficiently integrating learning-based event data association with IMU measurements remains an open problem. This work aims to bridge this gap by proposing a combined learning-optimization framework.

## 3. Methodology

The overall design of the framework aims to tightly fuse the learnable event-based data association with traditional IMU pre-integration. Fig. 1 depicts the overview of our system. For the front end, a deep neural network (Section 3.1) is utilized to estimate the sparse patch-based correspondences for the optical flow of events. On the back end, Hessian information derived from the learned DBA layer is tightly integrated with the IMU pre-integration (Section 3.2). This design leverages the representational power of deep neural networks to achieve robust event-based data association while simultaneously harnessing inertial measurement benefits without requiring IMU training data, thereby preserving the generalization capabilities of our DEIO.

### 3.1. Learning-based Event Data Association

**Event Encoding:** Each event is represented as a tuple  $(t, x, y, p)$ , where  $t$  denotes the trigger timestamp in microseconds at the pixel  $(x, y)$  and  $p_i$  indicates polarity with  $p = 1$  for an increase in brightness and  $p = -1$  for a decrease. The event streams are divided into segments based on a predefined temporal interval  $\Delta t$ . We preprocess the event segment within an interval  $[t_i - \Delta t, t_i)$  into a tensor  $\mathbf{E}_i \in \mathbb{R}^{D \times H \times W}$  using the voxel representation [39], where  $D$  represents the number of discretization steps in time. Therefore, the event-based optical flow estimation from segment  $i$  to segment  $j$  fundamentally involves establishing data correspondences between  $\mathbf{E}_i$  and  $\mathbf{E}_j$ .

**Patch Structure:** The patch-based architecture [32] is adopted to compute the flow for a set of sparse event patches. A  $p \times p$  event patch, sampled from the event voxel  $\mathbf{E}$ , is represented as a set of pixel coordinates  $\mathbf{P} = [\mathbf{x}, \mathbf{y}] \in \mathbb{R}^{p^2 \times 2}$ , where all pixels within the patch are assumed to have a constant inverse depth  $\mathbf{d} \in \mathbb{R}_+$ . The dynamic event patch graph  $\mathcal{G}$  is a bipartite graph where each edge is denoted as  $[(i, n), j]$ , indicating the relationship between event patch  $n$  from segment  $i$  and the target segment  $j$ . The network structure for event-based data association inherits the recurrent network from [19, 32] and consists of three primary components: (i) a feature encoder that extracts patch-based event feature representations; (ii) a correlation layer that computes the visual similarity; (iii) a recurrent update

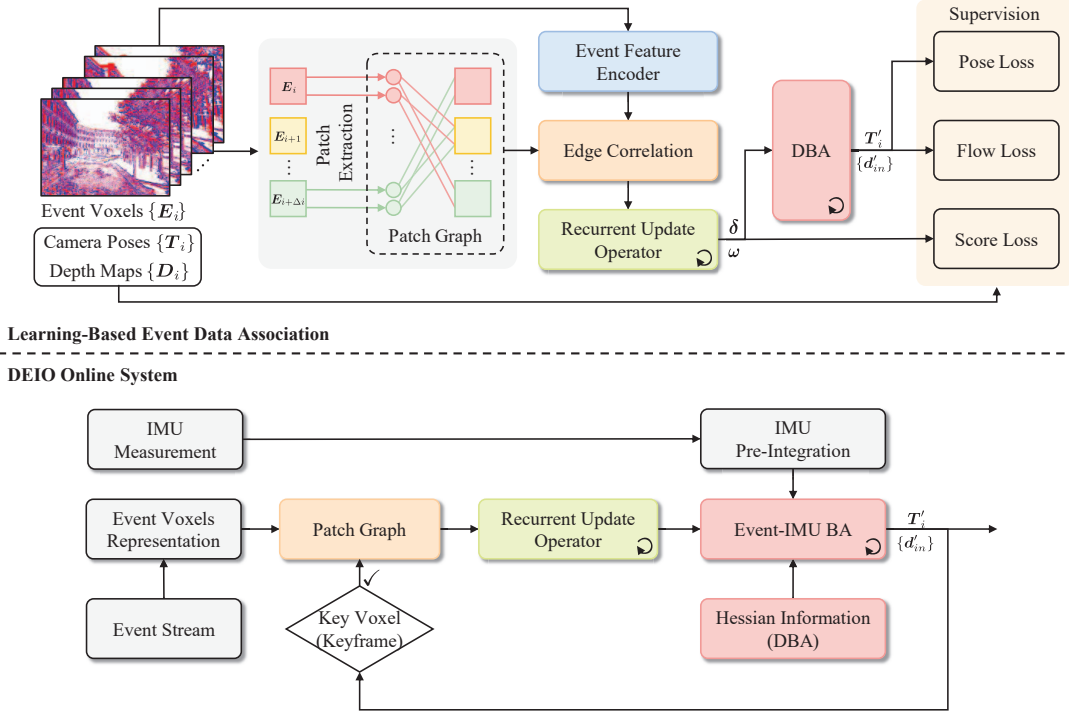


Figure 1. Overview of the DEIO system. It decouples network training from IMU integration and operates in two phases: offline training and online optimization. During training, a unified event-based optical flow network is trained to provide robust data associations of sparse event patches. At runtime, the Hessian information, derived from the DBA layer in the update operator, is utilized to tightly integrate event patch correspondence with IMU pre-integration through an event patch-based co-visibility factor graph optimization.

operator that handles event-patch correspondences, which estimates the 2D optical flow vector  $\delta$  and the confidence weights  $\omega$  for each patch.

**Differentiable Bundle Adjustment (DBA):** The DBA layer jointly refines camera poses and patch depth across the entire patch graph  $\mathcal{G}$  to match the predicted patch optical flow  $\delta$  by the following optimization objective:

$$\{\{T'_{ji}\}, \{d'_{in}\}\} = \arg \min_{T, d} \sum_{\mathcal{G}} \|F(T_{ji}, d_{in})\|_{\omega_{inj}}^2, \quad (1)$$

$$F(T_{ji}, d_{in}) = \pi(T_{ji} \cdot \pi^{-1}(\hat{P}_{in}, d_{in})) - (\hat{P}_{in} + \delta_{inj})$$

where  $(\delta_{inj}, \omega_{inj})$  is the patch-based optical flow field predicted by the event-based recurrent network. The term  $\delta_{inj} \in \mathbb{R}^2$  denotes a 2D flow vector that indicates how the reprojection of the event-patch center should be updated, and  $\omega_{inj}$  serves as the patch-wise confidence weight of the optical flow.  $F$  is the shorthand to denote the residual term on the patch center coordinates, and  $\|\cdot\|$  is the Mahalanobis distance.  $\pi$  and  $\pi^{-1}$  are the projection and back-projection functions of the event camera.  $T_{ji} = T_j^{-1}T_i$  represents the transformation from frame  $i$  to frame  $j$ , where  $T_i, T_j \in \mathbb{SE}(3)$  denote the camera poses in the camera-to-world format. In this work, the pre-trained network weights [19] are employed to estimate sparse event patch correspondences

due to their strong generalization.

### 3.2. Learnable Hessian Information Extraction

To extract the information from the learning-based event data association and integrate it with the IMU, we linearize Eq. (1) as follows:

$$F(\xi_{ji} \oplus T_{ji}, d_{in} + \Delta d_{in}) - F(T_{ji}, d_{in}) = [J_{ji} \quad J_{d_{in}}] \begin{bmatrix} \xi_{ji} \\ \Delta d_{in} \end{bmatrix} \quad (2)$$

where  $\xi_{ji}$  is the Lie algebras of the updated pose in  $\mathbb{SE}(3)$ .  $\Delta d_{in}$  denotes the updated state of the inverse depth. The Jacobians  $J_{ji}, J_{d_{in}}$  are the partial derivatives of  $F$  with respect to the pose  $T_{ji}$  and the inverse depth  $d_{in}$ , respectively. An event-patch  $P_{in}$  can be reprojected from segment  $i$  into segment  $j$  follows the warping function:

$$P'_{jn} = [x_{jn} \quad y_{jn} \quad z_{jn} \quad 1]^T = T_{ji} \cdot \pi^{-1}(\hat{P}_{in}, d_{in}) \quad (3)$$

where  $(x_{jn}, y_{jn}, z_{jn})$  is the center of the event patch at segment  $j$  in the event camera coordinate.  $J_{ji}$  is expressed as:

$$J_{ji} = \begin{bmatrix} \frac{f_x}{z_{jn} \cdot d_{in}} & 0 & -\frac{f_x x_{jn}}{z_{jn}^2 \cdot d_{in}} & -\frac{f_x x_{jn} y_{jn}}{z_{jn}^2} & f_x + \frac{f_x x_{jn}^2}{z_{jn}^2} & -\frac{f_x y_{jn}}{z_{jn}} \\ 0 & \frac{f_y}{z_{jn} \cdot d_{in}} & -\frac{f_y y_{jn}}{z_{jn}^2 \cdot d_{in}} & -f_y - \frac{f_y y_{jn}^2}{z_{jn}^2} & \frac{f_y x_{jn} y_{jn}}{z_{jn}^2} & \frac{f_y x_{jn}}{z_{jn}} \end{bmatrix} \quad (4)$$

where  $f_x$  and  $f_y$  are the given event camera intrinsic parameters.  $\mathbf{J}_{d_{in}}$  is given as:

$$\mathbf{J}_{d_{in}} = \begin{bmatrix} f_x \left( \frac{t_{ji}[0]}{z_{jn} \cdot d_{in}} - \frac{t_{ji}[2]x_{jn}}{z_{jn}^2 \cdot d_{in}} \right) \\ f_y \left( \frac{t_{ji}[1]}{z_{jn} \cdot d_{in}} - \frac{t_{ji}[2]y_{jn}}{z_{jn}^2 \cdot d_{in}} \right) \end{bmatrix} \quad (5)$$

where  $t_{ji}$  is the translation vector of the related transform between  $T_i$  and  $T_j$ . Therefore, the Hessian matrix of Eq. (2) can be computed as follows:

$$\mathbf{H}_{ji} = [\mathbf{J}_{ji} \quad \mathbf{J}_{d_{in}}]^T \mathbf{W}_{inj} [\mathbf{J}_{ji} \quad \mathbf{J}_{d_{in}}] \quad (6)$$

where  $\mathbf{W}_{inj} = \text{diag}(\omega_{inj})$ . To improve readability, the notation for  $[(i, n), j]$  edges in  $\mathbf{H}_{ji}$  and  $\mathbf{W}_{inj}$  will be omitted in subsequent equations unless otherwise specified.

By decoupling the pose and depth variables, the system can be solved efficiently using the Schur complement:

$$\begin{aligned} \mathbf{H} \begin{bmatrix} \xi_{ji} \\ \Delta d_{in} \end{bmatrix} &= -[\mathbf{J}_{ji} \quad \mathbf{J}_{d_{in}}]^T \mathbf{W} \mathbf{F}(T_{ji}, d_{in}) \\ \begin{bmatrix} \mathbf{B} & \mathbf{E} \\ \mathbf{E}^T & \mathbf{C} \end{bmatrix} \begin{bmatrix} \xi_{ji} \\ \Delta d_{in} \end{bmatrix} &= \begin{bmatrix} \mathbf{v} \\ \mathbf{u} \end{bmatrix} \end{aligned} \quad (7)$$

Therefore, the following equation can be obtained:

$$\xi_{ij} = \underbrace{[\mathbf{B} - \mathbf{E} \mathbf{C}^{-1} \mathbf{E}^T]^{-1}}_{\mathbf{H}_g} \underbrace{(\mathbf{v} - \mathbf{E} \mathbf{C}^{-1} \mathbf{u})}_{\mathbf{V}_g} \quad (8)$$

where  $\mathbf{B}$  is the matrix with size of  $60 \times 60$ ,  $\mathbf{E}$  is the residual matrix with size of  $60 \times 960$ , and  $\mathbf{C}$  is a diagonal matrix with size of  $960 \times 960$ . A damping factor of  $10^{-4}$  is also applied to  $\mathbf{C}$  as [32]. These matrices establish an interframe pose constraint (represented by  $\mathbf{H}_g$  and  $\mathbf{V}_g$ ) that integrates the DBA information. After updating the camera poses  $T'_{ji} = \text{Exp}(\xi_{ji})T_{ji}$ , the inverse depth of each event patch can be updated as:

$$\begin{aligned} d'_{in} &= \Delta d_{in} + d_{in} \\ \Delta d_{in} &= \mathbf{C}^{-1}(\mathbf{u} - \mathbf{E}^T \xi_{ji}) \end{aligned} \quad (9)$$

The calculations of Eq. (8) and Eq. (9) can be efficiently performed in parallel on a GPU with CUDA acceleration. All the Hessian information  $\mathbf{H}_g$  and the corresponding  $\mathbf{V}_g$ , derived from the co-visibility graph, are integrated into the factor graph where they are optimized on the CPU. The DBA contributes extensive geometric information, incorporating learned uncertainties, to the factor graph. The optimization results (updated poses and depths) are then iteratively fed back to refine the event-based optical flow network.

### 3.3. Event-IMU Combined Bundle Adjustment

Unlike end-to-end approaches that use deep networks to fuse the features from two modalities (visual and IMU)

and predict poses directly, our DEIO combines the neural networks with event-inertial bundle adjustment. To this end, we design a learning-optimization combined framework that tightly integrates the Hessian information from DBA and IMU pre-integration within keyframe-based sliding window optimization. The full state vector of the  $k$ th keyframe in the sliding window (with the total number of keyframes  $K = 10$  in our implementation), is defined as:

$$\chi = [T_{b_k}^w, \mathbf{v}_{b_k}^w, \mathbf{b}_{a_k}, \mathbf{b}_{g_k}], k = 1, 2, 3, \dots \quad (10)$$

where  $T_{b_k}^w = \begin{bmatrix} \mathbf{R}_{b_k}^w & \mathbf{t}_{b_k}^w \\ 0 & 1 \end{bmatrix} \in \mathbb{SE}(3)$  is the pose of the body (IMU) frame in the world frame, given by the translation  $\mathbf{t}_{b_k}^w$  and rotation matrix  $\mathbf{R}_{b_k}^w$ .  $\mathbf{v}_{b_k}^w$  is the velocity of the IMU in the world frame.  $\mathbf{b}_{a_k}$  and  $\mathbf{b}_{g_k}$  are the accelerometer bias and gyroscope bias, respectively. We solve the state estimation problem by constructing a factor graph with the GTSAM library and optimizing it with the Levenberg-Marquardt. The cost function can be written as:

$$\mathcal{J}(\chi) = \|\mathbf{r}_{\text{event}}^k\|_{W_{\text{event}}^k}^2 + \sum_{k=0}^{K-1} \|\mathbf{r}_{\text{imu}}^k\|_{W_{\text{imu}}^k}^2 + \|\mathbf{r}_m\|_{W_m}^2 \quad (11)$$

Eq. (11) contains the event-based residuals  $\mathbf{r}_{\text{event}}^k$  with weight  $W_{\text{event}}^k$ , the IMU pre-integration residuals  $\mathbf{r}_{\text{imu}}^k$  with weight  $W_{\text{imu}}^k$ , and the marginalization residuals  $\mathbf{r}_m$  with weight  $W_m$ . Given the Hessian information  $\mathbf{H}_g$  and the corresponding  $\mathbf{V}_g$ , the event residual factor can be written as:

$$\begin{aligned} \mathbf{r}_{\text{event}}^k &= \frac{1}{2} \begin{bmatrix} \xi_{e_0}^w & \dots & \xi_{e_{10}}^w \end{bmatrix} \mathbf{H}_g \begin{bmatrix} \xi_{e_0}^w \\ \vdots \\ \xi_{e_{10}}^w \end{bmatrix} \\ &\quad - \begin{bmatrix} \xi_{e_0}^w & \dots & \xi_{e_{10}}^w \end{bmatrix} \mathbf{V}_g \end{aligned} \quad (12)$$

where  $\xi_{e_k}^w = \xi_b^e \cdot \xi_{b_k}^w$ , and  $\xi_{b_k}^w = \log_{\mathbb{SE}(3)}(T_{b_k}^w) \cdot \xi_{e_k}^w$  and  $\xi_{b_k}^w$  are the Lie algebras of the event camera pose and IMU pose in  $k^{\text{th}}$  keyframe, respectively.  $\xi_b^e$  is the extrinsics between the event camera and IMU.

Eventually, the IMU residual factor can be derived as follows:

$$\mathbf{r}_{\text{imu}}^k = \begin{bmatrix} \mathbf{R}_{b_k}^{b_{k+1}} (\mathbf{t}_{b_{k+1}}^w - \mathbf{t}_{b_k}^w - \mathbf{v}_{b_k}^w \Delta t - \frac{1}{2} \mathbf{g}^w \Delta t^2) - \hat{\alpha}_{b_{k+1}}^{b_k} \\ \mathbf{R}_{b_k}^{b_{k+1}} (\mathbf{v}_{b_{k+1}}^w - \mathbf{v}_{b_k}^w - \mathbf{g}^w \Delta t) - \hat{\beta}_{b_{k+1}}^{b_k} \\ 2 \left[ (\mathbf{q}_{b_k}^w)^{-1} \otimes \mathbf{q}_{b_{k+1}}^w \otimes (\hat{\gamma}_{b_{k+1}}^{b_k})^{-1} \right]_{xyz} \\ \mathbf{b}_{a_{k+1}} - \mathbf{b}_{a_k} \\ \mathbf{b}_{g_{k+1}} - \mathbf{b}_{g_k} \end{bmatrix} \quad (13)$$

where  $\hat{\alpha}_{b_{k+1}}^{b_k}$ ,  $\hat{\beta}_{b_{k+1}}^{b_k}$ ,  $\hat{\gamma}_{b_{k+1}}^{b_k}$  are the IMU pre-integration term [14];  $\mathbf{g}^w$  is the gravity vector;  $\Delta t$  is the time interval between keyframe  $k$  and  $k+1$ ;  $\mathbf{q}_{b_k}^w$  is the quaternion of the corresponding rotation matrix  $\mathbf{R}_{b_k}^w$  with  $[\cdot]_{xyz}$  extracts the vector portion.

Table 1. Accuracy comparison [MPE(%)] of our DEIO in DAVIS240c dataset [22]. The estimated trajectory is aligned with the ground truth over the first 5 seconds.

Methods	Modality	boxes_translation	hdr_boxes	boxes_6dof	dynamic_translation	dynamic_6dof	poster_translation	hdr_poster	poster_6dof	Average
Zhu et al. [40]	E+I	2.69	1.23	3.61	1.90	4.07	0.94	2.63	3.56	2.58
Henri et al. [29]	E+I	0.57	0.92	0.69	0.47	0.54	0.89	0.59	0.82	0.69
Ultimate-SLAM [33]	E+I	0.76	0.67	0.44	0.59	0.38	0.15	0.49	0.30	0.47
Jung et al. [17]	E+I	1.50	2.45	2.88	4.92	6.23	3.43	2.38	2.53	3.29
HASTE-VIO [1]	E+I	2.55	1.75	2.03	1.32	0.52	1.34	0.57	1.50	1.45
EKLT-VIO [21]	E+F+I	0.48	0.46	0.84	0.40	0.79	0.35	0.65	0.35	0.54
Dai et al. [5]	E+I	1.0	1.8	1.5	0.9	1.5	1.9	2.8	1.2	1.58
Mono-EIO [13]	E+I	0.34	0.40	0.61	0.26	0.43	0.40	0.40	0.26	0.39
Kai et al. [31]	E+I	0.36	0.31	0.32	0.59	0.49	0.23	0.18	0.31	0.35
PL-EVIO [14]	E+F+I	0.06	0.10	0.21	0.24	0.48	0.54	0.12	0.14	0.24
Lee et al. [20]	E+F+I	0.74	0.69	0.77	0.71	0.86	0.28	0.52	0.59	0.65
EVI-SAM [15]	E+F+I	0.11	0.13	0.16	0.30	0.27	0.34	0.15	0.24	0.21
DPVO [32]	F	<b>0.02</b>	0.71	0.59	0.09	0.05	0.20	0.49	0.44	0.32
DBA-Fusion [37]	F+I	0.07	0.27	0.10	0.56	0.11	0.13	0.38	0.19	0.23
DEVO [19]	E	0.06	<b>0.06</b>	<b>0.71</b>	0.09	0.08	0.06	0.14	0.44	0.21
<b>DEIO</b>	E+I	0.07	0.09	<b>0.05</b>	<b>0.06</b>	<b>0.04</b>	<b>0.04</b>	<b>0.06</b>	<b>0.08</b>	<b>0.06</b>

Table 2. Accuracy comparison [MPE(%)] of our DEIO in the Mono-HKU dataset [13]. The estimated trajectory is aligned with the ground truth over the first 5 seconds.

Resolution	Methods	Modality	vicon_hdr1	vicon_hdr2	vicon_hdr3	vicon_hdr4	vicon_darktolight1	vicon_darktolight2	vicon_lighttodark1	vicon_lighttodark2	vicon_dark1	vicon_dark2	Average
DAVIS346 (346×260)	ORB-SLAM3 [2]	F	0.32	0.75	0.60	0.70	0.75	0.76	0.41	0.58	<i>failed</i>	0.60	0.61
	VINS-MONO [26]	F+I	0.96	1.60	2.28	1.40	0.51	0.98	0.55	0.55	0.88	0.52	1.02
	DBA-Fusion [37]	F+I	0.32	0.41	<i>failed</i>	<i>failed</i>	0.72	0.55	<i>failed</i>	2.65	3.32	<i>failed</i>	1.33
	Ultimate-SLAM [33]	E+I	1.49	1.28	0.66	1.84	1.33	1.48	1.79	1.32	1.75	1.10	1.40
	Ultimate-SLAM [33]	E+F+I	2.44	1.11	0.83	1.49	1.00	0.79	0.84	1.49	3.45	0.63	1.41
	Mono-EIO [13]	E+I	0.59	0.74	0.72	0.37	0.81	0.42	0.29	0.79	1.02	0.49	0.62
	PL-EIO [14]	E+I	0.57	0.54	0.69	0.32	0.66	0.51	0.33	0.53	0.35	0.38	0.49
	PL-EVIO [14]	F+E+I	0.17	0.12	0.19	0.11	0.14	0.12	0.13	0.16	0.43	0.47	0.20
	DEVO [19]	E	<b>0.11</b>	<b>0.07</b>	<b>0.12</b>	0.07	0.97	0.12	0.15	<b>0.12</b>	<b>0.07</b>	<b>0.07</b>	0.19
	<b>DEIO</b>	E+I	0.14	0.09	0.16	<b>0.07</b>	<b>0.11</b>	<b>0.10</b>	<b>0.11</b>	0.13	<b>0.05</b>	0.08	<b>0.10</b>

## 4. Experiments

We conduct quantitative and qualitative evaluations of our DEIO across *ten* challenging real-world datasets with varying camera resolution and diversity scenarios on different platforms. Specifically, in Section 4.1, we compare DEIO with baseline methods across multiple challenging event datasets, showcasing its superior performance and exceptional generalization capabilities. Section 4.2 provides a time efficiency evaluation of DEIO. Finally, the project website provides video demos and the qualitative results as supplementary material.

### 4.1. Comparisons with SOTA Methods in Challenge Benchmarks

To ensure a fair comparison, a consistent trajectory alignment protocol is required. Therefore, we employ different alignment ways and evaluation criteria according to the compared baseline methods, including the mean position error (MPE), and the root mean squared error (RMSE) / Absolute Trajectory Errors (ATE), using the publicly available trajectory evaluation tool [12]. The notations E, F, and I in each table represent the use of event, frame, and IMU, respectively.

**DAVIS240C [22]:** As shown in Table 1, EVI-SAM achieves the best performance among the non-learning methods. In contrast, learning-based methods (such as DEVO) can achieve performance comparable to EVI-SAM (which combines both direct and feature-based methods) using purely event sensors, highlighting the effectiveness and strength of learning-based approaches. Meanwhile, our learning-optimization combined method exhibits significantly superior performance compared to other learning-based methods (DPVO, DBA-Fusion, and DEVO). Compared to DEVO, our proposed DEIO reduces the pose tracking error by up to 71%, owing to the effective integration of learning-based and traditional optimization methods.

**Monocular HKU-dataset [13]:** Table 2 demonstrates that DEIO outperforms all the event-based methods and decreases the average pose tracking error by at least 47%. As illustrated in the project website, the estimated trajectory of DEVO suffers from significant scale loss because the absolute scale cannot be observed in monocular event-only odometry. In contrast, our DEIO, despite also being based on a monocular setup, effectively overcomes scale ambiguity and aligns closely with the ground truth trajectory. This improvement is attributed to the effective compensation provided by the IMU.

Table 3. Accuracy comparison [MPE(%)] of our DEIO in the Stereo-HKU dataset [4]. The entire sequence of estimated poses is aligned with the ground truth trajectory. The DPVO, DEVO are taken from [19], and the results of Kai et al. are taken from [31].

Methods	Modality	agg_translation	agg_rotation	agg_flip	agg_walk	hdr_circle	hdr_slow	hdr_tran_rota	hdr_agg	dark_normal	Average
ORB-SLAM3 [2]	Stereo F+I	0.15	0.35	0.36	<i>failed</i>	0.17	0.16	0.30	0.29	<i>failed</i>	0.25
VINS-Fusion [27]	Stereo F+I	0.11	1.34	1.16	<i>failed</i>	5.03	0.13	0.11	1.21	0.86	1.24
DPVO [32]	F	0.07	<b>0.04</b>	0.99	1.17	0.31	0.23	0.67	0.29	<i>failed</i>	0.47
DBA-Fusion [37]	F+I	0.13	0.16	0.83	0.37	0.18	<i>failed</i>	<i>failed</i>	0.10	0.27	0.29
Kai et al. [31]	E+I	0.21	0.28	0.81	0.35	0.71	0.43	0.50	0.27	0.52	0.45
PL-EVIO [14]	E+F+I	0.07	0.23	0.39	0.42	0.14	0.13	0.10	0.14	1.35	0.33
EVI-SAM [15]	E+F+I	0.17	0.24	0.32	<b>0.26</b>	<b>0.13</b>	0.11	0.11	0.10	0.85	0.25
ESIO [4]	Stereo E+I	0.55	0.78	3.17	1.30	0.46	0.31	0.91	1.41	0.35	1.03
ESVIO [4]	Stereo E+F+I	0.10	0.17	0.36	0.31	0.16	0.11	0.10	0.10	0.42	0.20
DEVO [19]	E	0.06	0.05	0.71	0.90	0.39	0.08	<b>0.08</b>	0.26	<b>0.06</b>	0.29
<b>DEIO</b>	E+I	<b>0.06</b>	0.09	<b>0.20</b>	0.48	0.14	<b>0.07</b>	0.09	<b>0.06</b>	0.11	<b>0.14</b>

Table 4. Accuracy comparison [MPE(%)] of our DEIO in VECtor dataset [8]. The entire sequence of estimated poses is aligned with the ground truth trajectory.

Methods	Modality	corner-slow	desk-normal	sofa-fast	mountain-fast	corridors-dolly	corridors-walk	units-dolly	units-scooter	average
ORB-SLAM3 [2]	Stereo F+I	1.49	0.46	0.21	2.11	1.03	1.32	7.64	6.22	2.56
VINS-Fusion [27]	Stereo F+I	1.61	0.47	0.57	<i>failed</i>	1.88	0.50	4.39	4.92	2.05
DPVO [32]	F	<b>0.30</b>	<b>0.09</b>	<b>0.07</b>	<b>0.11</b>	0.56	0.54	1.52	1.67	0.61
DBA-Fusion [37]	F+I	1.72	0.48	0.43	<i>failed</i>	1.37	0.59	1.23	0.48	0.90
EVO [28]	E	4.33	<i>failed</i>	<i>failed</i>	<i>failed</i>	<i>failed</i>	<i>failed</i>	<i>failed</i>	<i>failed</i>	4.33
ESVO [36]	Stereo E	4.83	<i>failed</i>	<i>failed</i>	<i>failed</i>	<i>failed</i>	<i>failed</i>	<i>failed</i>	<i>failed</i>	4.83
Ultimate-SLAM [33]	E+F+I	4.83	2.24	2.54	4.13	<i>failed</i>	<i>failed</i>	<i>failed</i>	<i>failed</i>	3.44
PL-EVIO [14]	E+F+I	2.10	3.66	0.17	0.13	1.58	0.92	5.84	5.00	2.43
ESVIO [4]	Stereo E+F+I	1.49	0.61	0.17	0.16	1.13	<b>0.43</b>	3.43	2.85	1.28
EVI-SAM [15]	E+F+I	2.50	1.45	0.98	0.38	1.58	1.27	0.59	0.83	1.20
DEVO [19]	E	0.59	0.11	0.38	0.37	<b>0.51</b>	1.04	0.48	0.88	0.55
<b>DEIO</b>	E+I	0.50	0.13	0.44	0.24	0.78	0.74	<b>0.35</b>	<b>0.35</b>	<b>0.44</b>

**Stereo HKU-dataset [4]:** In Table 3, our method outperforms all previous works in terms of average positioning error. Note that event-only VO methods, such as EVO [28], ESVO [36], as well as stereo event and IMU-based methods like ESVO2 [23, 24], fail to perform successfully on any of the sequences in this dataset. Moreover, DEIO beats DEVO and increases the average accuracy of the sequences up to 48%.

**VECtor [8]:** As presented in Table 4, our proposed DEIO achieves remarkable results on average. It surpasses all image-based baselines with high-quality frames (image 1224×1024 vs event 640×480) and even outperforms Ultimate-SLAM, PL-EVIO, ESVIO, and EVI-SAM on over 75% of the sequences, which utilize event, image, and IMU. Our DEIO also outperforms DEVO on average in large-scale sequences, thanks to the complementary integration of the event and IMU sensors, while other monocular visual-only methods struggle with scale ambiguity and drift.

**TUM-VIE [18]:** The results in Table 5 demonstrate that DEIO outperforms all other methods on four out of five sequences, despite DH-PTAM [30] utilizing four cameras of the setup (stereo events and stereo images). Our DEIO significantly outperforms ESVO2 [4, 24] and improves average accuracy by up to 79%. Notably, ESVO2 relies on stereo

Table 5. Accuracy comparison [ATE/RMSE (cm)] of our DEIO in TUM-VIE dataset [18]. The entire sequence of estimated poses is aligned with the ground truth trajectory. The EVO, ESVO, and ESPTAM are taken from [11], DH-PTAM and Ultimate-SLAM are sourced from [30], DEVO is from [19], and ESVIO\_AA, ESVO2 are from [24].

Methods	Modality	mocap-					Average
		1d-trans	3d-trans	6dof	desk	desk2	
EVO [28]	E	7.5	12.5	85.5	54.1	75.2	47.0
ESVO [36]	Stereo E	12.3	17.2	13.0	12.4	4.6	11.9
ESVIO_AA [23]	Stereo E+I	3.9	18.9	<i>failed</i>	9.00	9.5	10.3
ESVO2 [24]	Stereo E+I	3.3	7.3	3.2	6.2	4.0	4.8
ES-PTAM [11]	Stereo E	1.05	8.53	10.25	2.5	7.2	5.9
DH-PTAM [30]	Stereo E+F	10.3	<b>0.7</b>	2.4	1.6	1.5	3.3
Ultimate-SLAM [33]	E+F+I	3.9	4.7	35.3	19.5	34.1	19.5
DEVO [19]	E	0.5	1.1	1.6	1.7	1.0	1.2
<b>DEIO</b>	E+I	<b>0.4</b>	1.1	<b>1.4</b>	<b>1.4</b>	<b>0.7</b>	<b>1.0</b>

event and IMU setup, while our DEIO achieves superior results using only a monocular event camera and IMU. This highlights that our approach, using only a monocular event camera and IMU, can recover scale comparable to that of a stereo event setup.

**EDS [16]:** As shown in Table 6, our DEIO outperforms the image-based baselines, including ORB-SLAM3, DPVO, and DBA-Fusion. Moreover, DEIO achieves an average improvement of 30% over DEVO and demonstrates

Table 6. Accuracy comparison [ATE/RMSE (cm)] of our DEIO in EDS dataset [16]. The entire sequence of estimated poses is aligned with the ground truth trajectory. The ORB-SLAM3, DPVO, and DEVO are taken from [19], while RAMP-VO is sourced from [25].

Methods	Modality	peanuts_dark	peanuts_light	peanuts_run	rocket_dark	rocket_light	ziggy	ziggy_hdr	ziggy_flying	all_chars	Average
ORB-SLAM3 [2]	Stereo F+I	6.15	27.26	16.83	10.12	32.53	26.92	81.98	20.57	21.37	27.08
DPVO [32]	F	1.26	12.99	25.48	27.41	63.11	14.86	66.17	10.85	95.87	35.33
DBA-Fusion [37]	F+I	7.26	149.36	134.92	114.24	117.09	173.50	140.51	11.81	126.36	108.34
DEVO [19]	E	4.78	21.07	38.10	8.78	59.83	11.84	<b>22.82</b>	10.92	<b>10.76</b>	20.99
RAMP-VO [25]	E+F	<b>1.20</b>	<b>9.03</b>	<b>13.19</b>	<b>7.20</b>	17.53	19.05	28.78	6.35	28.61	<b>14.55</b>
<b>DEIO</b>	E+I	1.77	16.27	19.96	8.91	<b>15.41</b>	<b>10.39</b>	23.82	<b>3.84</b>	31.55	<b>14.66</b>

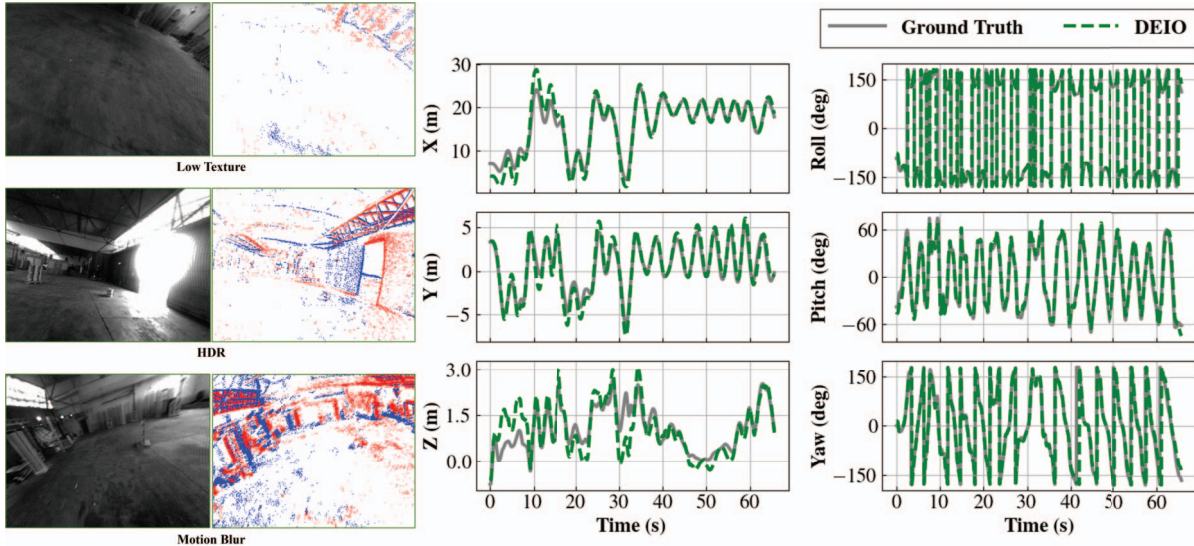


Figure 2. The estimated trajectories (X, Y, Z, Roll, Pitch, Yaw) of our DEIO against the GT in the sequence of indoor\_forward\_7 from the UZH-FPV [6] dataset. The image view (visualization-only) demonstrates the condition under low texture, HDR, and motion blur.

performance comparable to RAMP-VO, a learning-based VO system that leverages both event and image modalities. In the case of DBA-Fusion [37], despite operating within a dense learning-based VIO framework, DEIO demonstrates superior performance, showing the distinct advantages of event cameras in challenging scenarios.

**UZH-FPV [6]:** As shown in Table 7 and Fig. 2, this dataset poses significant challenges for existing methods, with even advanced learning-based VO approaches like DPVO failing to maintain reliable tracking across all sequences. Additionally, incorporating IMU measurements does not resolve these challenges, such as learning-based VIO methods (DBA-Fusion), which also fail to complete any sequences. This is due to the motion blur caused by rapid movement, which makes it difficult to effectively establish data association for the image sensor, even if these methods are equipped with a powerful learning network. In contrast, our DEIO achieves higher average performance than all baseline methods, demonstrating greater resilience to the fast flight conditions.

**MVSEC [38]:** In Table 8, DEIO surpasses the event-based baseline, especially for the *Flying\_4*, where it attains an RMSE of 40% lower than DEVO. Although ESvio [4] employs a setup integrating stereo images, stereo events, and IMU data, DEIO, which relies solely on monocular

Table 7. Accuracy comparison [MPE (%)] of our DEIO in UZH-FPV dataset [6]. The entire sequence of estimated poses is aligned with the ground truth trajectory. The DPVO, DEVO are taken from [19].

Methods	Modality	Indoor_forward						Average
		3	5	6	7	9	10	
VINS-Fusion [27]	Stereo F+I	0.84	failed	1.45	0.61	2.87	4.48	2.05
ORB-SLAM3 [2]	Stereo F+I	0.55	1.19	failed	0.36	0.77	1.02	0.78
DPVO [32]	F	failed	failed	failed	failed	failed	failed	—
VINS-MONO [26]	F+I	0.65	1.07	<b>0.25</b>	0.37	0.51	0.92	0.70
DBA-Fusion [37]	F+I	failed	failed	failed	failed	failed	failed	—
Ultimate SLAM [33]	E+F+I	failed	failed	failed	failed	failed	failed	—
PL-EVIO [14]	E+F+I	0.38	0.90	0.30	0.55	<b>0.44</b>	1.06	0.61
DEVO [19]	E	<b>0.37</b>	0.40	0.31	0.50	0.61	<b>0.52</b>	0.45
<b>DEIO</b>	E+I	0.39	<b>0.36</b>	0.33	<b>0.32</b>	0.59	0.55	<b>0.42</b>

event data and IMU, demonstrates an average accuracy improvement of over 80% compared to ESvio. As for the learning-based VIO method (DBA-Fusion) that relies on dense data association, it failed on three out of the four sequences. This indicates that even though the deep learning methods can provide strong data association capabilities, the degradation of images in challenging environments limits their performance compared to the event modality.

**DSEC [9]:** As presented in Table 9, our DEIO outperforms the stereo event methods (ES-PTAM, ESVO, ESvio-AA, and ESIO) by large margins on all sequences (at

Table 8. Accuracy comparison [MPE (%)] of our DEIO in MVSEC dataset [38]. The entire sequence of estimated poses is aligned with the ground truth trajectory. The DEVO result is taken from [19].

Methods	Modality	Flying_1	Flying_2	Flying_3	Flying_4	Average
ORB-SLAM3 [2]	Stereo F+I	5.31	5.65	2.90	6.99	5.21
VINS-Fusion [27]	Stereo F+I	1.50	6.98	0.73	3.62	3.21
EVO [28]	E	5.09	failed	2.58	failed	3.84
ESVO [36]	Stereo E	4.00	3.66	1.71	failed	3.12
Ultimate-SLAM [33]	E+F+I	failed	failed	failed	2.77	2.77
PL-EVIO [14]	E+F+I	1.35	1.00	0.64	5.31	2.08
ESVIO [4]	Stereo E+F+I	0.94	1.00	0.47	5.55	1.99
DBA-Fusion [37]	F+I	2.20	failed	failed	failed	2.20
DEVO [19]	E	0.26	0.32	0.19	1.08	0.46
<b>DEIO</b>	<b>E+I</b>	<b>0.24</b>	<b>0.21</b>	<b>0.12</b>	<b>0.78</b>	<b>0.34</b>

Table 9. Accuracy comparison [ATE/RMSE (cm)] of our DEIO in DSEC dataset [9]. The entire sequence of estimated poses is aligned with the ground truth trajectory. The ESVO and ES-VIO\_AA are taken from [23], and ES-PTAM is sourced from [11].

Methods	Modality	dsec_zurich_city_04					Average
		a	b	c	d	e	
ESVO [36]	Stereo E	371.1	116.6	1357.1	2676.6	794.9	1063.3
ESVIO_AA [23]	Stereo E+I	105.0	66.7	637.9	699.8	130.3	327.9
ES-PTAM [11]	Stereo E	131.62	<b>29.02</b>	1184.37	1053.87	<b>75.9</b>	495.0
ESIO [4]	Stereo E+I	543.5	295.1	896.2	2977.0	2326.4	1407.6
ESVIO [4]	Stereo E+F+I	371.2	445.8	1892.7	921.7	352.0	796.7
<b>DEIO</b>	<b>E+I</b>	<b>80.6</b>	<b>35.4</b>	<b>413.8</b>	<b>207.6</b>	<b>86.1</b>	<b>164.7</b>

least 66.7% lower RMSE). The results from our DEIO align more closely with the ground truth, despite using a monocular setup, while the other methods employ a stereo setup. This demonstrates that DEIO can achieve comparable scale estimation to these stereo setups while providing superior state estimation results.

**ECMD [3]** We select the *Dense\_street\_night\_easy\_a* sequences of the ECMD dataset [3], which feature numerous flashing lights from vehicles, street signs, buildings, and moving vehicles (details on the website). Our DEIO runs on the event from the DAVIS346 and the IMU sensor, while the image frame output from the DAVIS346 is only used for illustration purposes. Fig. 3 shows a small drift with a 4.7 m error of our estimated trajectory on the 620 m drive.



Figure 3. The estimated trajectory of our DEIO in the night driving scenarios [3] and its comparison against the GNSS-INS-RTK as ground truth. The image view is for visualization only.

## 4.2. Ablation Study and Runtime Analysis

Fig. 4 illustrates the real-time performance of DEIO variants under various patch configurations on an Nvidia RTX 3090 GPU. Our DEIO, configured with 96 patches per event

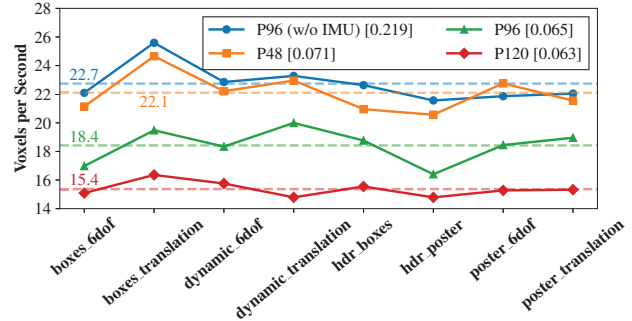


Figure 4. Runtime performance (voxels per second) of our DEIO using 48 (P48), 96 (P96), and 120 (P120) event patches per voxel, as well as a 96-patch version without IMU input (P96 w/o IMU). Values in the brackets indicate the average MPE (%) over all sequences.

voxel (P96), achieves an average processing speed of 18.4 voxels per second (VPS). Compared to the variant that excludes IMU data (P96 w/o IMU), our DEIO demonstrates a significant accuracy improvement of 69.0%, with only a minor runtime overhead of 4.3 VPS. The P48 variant achieves an average MPE of 0.071 while maintaining a runtime of 22.1 VPS, which is comparable to that of the P96 w/o IMU configuration (22.7 VPS). However, increasing the number of event patches further (P120) leads to diminishing returns in accuracy improvement while significantly increasing computational demands.

## 5. Conclusion

In this paper, we propose DEIO, a deep learning-based event-inertial odometry method. An event-based deep neural network is utilized to provide accurate and sparse associations of event patches over time, and DEIO further tightly integrates it with the IMU during the graph-based optimization process to provide robust 6 DoF pose tracking. Evaluation on *ten* challenging event-based benchmarks demonstrates that DEIO outperforms both image-based and event-based baselines. We have shown that the learning-optimization combined framework for SLAM is a promising direction. To further enhance the robustness and efficiency of the system, future work will focus on exploring IMU-bias online learning, event-image complementarity, and loop closure mechanisms for learning-based event-SLAM.

## References

- [1] Ignacio Alzugaray and Margarita Chli. Asynchronous multi-hypothesis tracking of features with event cameras. In *2019 International Conference on 3D Vision (3DV)*, pages 269–278. IEEE, 2019. 5
- [2] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual-inertial, and mul-

- timap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. 5, 6, 7, 8
- [3] Peiyu Chen, Weipeng Guan, Feng Huang, Yihan Zhong, Weisong Wen, Li-Ta Hsu, and Peng Lu. Ecmd: An event-centric multisensory driving dataset for slam. *IEEE Transactions on Intelligent Vehicles*, 2023. 8
- [4] Peiyu Chen, Weipeng Guan, and Peng Lu. Esvio: Event-based stereo visual inertial odometry. *IEEE Robotics and Automation Letters*, 8(6):3661–3668, 2023. 2, 6, 7, 8
- [5] Benny Dai, Cedric Le Gentil, and Teresa Vidal-Calleja. A tightly-coupled event-inertial odometry using exponential decay and linear preintegrated measurements. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9475–9482. IEEE, 2022. 5
- [6] Jeffrey Delmerico, Titus Cieslewski, Henri Rebecq, Matthias Faessler, and Davide Scaramuzza. Are we ready for autonomous drone racing? the uzh-fpv drone racing dataset. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6713–6719. IEEE, 2019. 7
- [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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3867–3876, 2018. 2
- [8] Ling Gao, Yuxuan Liang, Jiaqi Yang, Shaoxun Wu, Chenyu Wang, Jiaben Chen, and Laurent Kneip. Vector: A versatile event-centric benchmark for multi-sensor slam. *IEEE Robotics and Automation Letters*, 2022. 6
- [9] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 7, 8
- [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. 2
- [11] Suman Ghosh, Valentina Cavinato, and Guillermo Gallego. ES-PTAM: Event-based stereo parallel tracking and mapping. In *European Conference on Computer Vision (ECCV) Workshops*, 2024. 6, 8
- [12] Michael Grupp. evo: Python package for the evaluation of odometry and slam. Note: <https://github.com/MichaelGrupp/evo> Cited by: Table, 7, 2017. 5
- [13] Weipeng Guan and Peng Lu. Monocular event visual inertial odometry based on event-corner using sliding windows graph-based optimization. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2438–2445. IEEE, 2022. 2, 5
- [14] Weipeng Guan, Peiyu Chen, Yuhan Xie, and Peng Lu. Plevio: Robust monocular event-based visual inertial odometry with point and line features. *IEEE Transactions on Automation Science and Engineering*, 2023. 2, 4, 5, 6, 7, 8
- [15] Weipeng Guan, Peiyu Chen, Huibin Zhao, Yu Wang, and Peng Lu. Evi-sam: Robust, real-time, tightly-coupled event-visual-inertial state estimation and 3d dense mapping. *Advanced Intelligent Systems*, page 2400243, 2024. 2, 5, 6
- [16] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2022. 6, 7
- [17] Jae Hyung Jung and Chan Gook Park. Constrained filtering-based fusion of images, events, and inertial measurements for pose estimation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 644–650. IEEE, 2020. 5
- [18] Simon Klenk, Jason Chui, Nikolaus Demmel, and Daniel Cremers. Tum-vie: The tum stereo visual-inertial event dataset. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8601–8608. IEEE, 2021. 6
- [19] Simon Klenk, Marvin Motzet, Lukas Koestler, and Daniel Cremers. Deep event visual odometry. In *2024 International Conference on 3D Vision (3DV)*, pages 739–749. IEEE, 2024. 1, 2, 3, 5, 6, 7, 8
- [20] Min Seok Lee, Jae Hyung Jung, Ye Jun Kim, and Chan Gook Park. Event-and frame-based visual-inertial odometry with adaptive filtering based on 8-dof warping uncertainty. *IEEE Robotics and Automation Letters*, 2023. 5
- [21] Florian Mählknecht, Daniel Gehrig, Jeremy Nash, Friedrich M. Rockenbauer, Benjamin Morrell, Jeff De-laune, and Davide Scaramuzza. Exploring event camera-based odometry for planetary robots. *IEEE Robotics and Automation Letters (RA-L)*, 2022. 5
- [22] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 5
- [23] Junkai Niu, Sheng Zhong, and Yi Zhou. Imu-aided event-based stereo visual odometry. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11977–11983. IEEE, 2024. 2, 6, 8
- [24] Junkai Niu, Sheng Zhong, Xiuyuan Lu, Shaojie Shen, Guillermo Gallego, and Yi Zhou. Esvo2: Direct visual-inertial odometry with stereo event cameras. *IEEE Transactions on Robotics*, 2025. 2, 6
- [25] Roberto Pellerito, Marco Cannici, Daniel Gehrig, Joris Belhadj, Olivier Dubois-Matra, Massimo Casasco, and Davide Scaramuzza. Deep visual odometry with events and frames. In *IEEE/RSJ International Conference on Intelligent Robots (IROS)*, 2024. 1, 2, 7
- [26] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 5, 7
- [27] Tong Qin, Jie Pan, Shaozu Cao, and Shaojie Shen. A general optimization-based framework for local odometry estimation with multiple sensors. *arXiv preprint arXiv:1901.03638*, 2019. 6, 7, 8
- [28] Henri Rebecq, Timo Horstschäfer, Guillermo Gallego, and Davide Scaramuzza. Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time. *IEEE Robotics and Automation Letters*, 2(2):593–600, 2016. 1, 6, 8

- [29] Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. In *British Machine Vision Conference (BMVC)*, 2017. [2](#), [5](#)
- [30] Abanob Soliman, Fabien Bonardi, Désiré Sidibé, and Samia Bouchafa. Dh-ptam: a deep hybrid stereo events-frames parallel tracking and mapping system. *IEEE Transactions on Intelligent Vehicles*, 2024. [6](#)
- [31] Kai Tang, Xiaolei Lang, Yukai Ma, Yuehao Huang, Laijian Li, Yong Liu, and Jiajun Lv. Monocular event-inertial odometry with adaptive decay-based time surface and polarity-aware tracking. In *IEEE/RSJ International Conference on Intelligent Robots (IROS)*, 2024. [5](#), [6](#)
- [32] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024. [2](#), [4](#), [5](#), [6](#), [7](#)
- [33] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001, 2018. [2](#), [5](#), [6](#), [7](#), [8](#)
- [34] Chengxi Ye, Anton Mitrokhin, Cornelia Fermüller, James A Yorke, and Yiannis Aloimonos. Unsupervised learning of dense optical flow, depth and egomotion with event-based sensors. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5831–5838. IEEE, 2020. [2](#)
- [35] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. [2](#)
- [36] Yi Zhou, Guillermo Gallego, and Shaojie Shen. Event-based stereo visual odometry. *IEEE Transactions on Robotics*, 37(5):1433–1450, 2021. [1](#), [2](#), [6](#), [8](#)
- [37] Yuxuan Zhou, Xingxing Li, Shengyu Li, Xuanbin Wang, Shaoquan Feng, and Yuxuan Tan. Dba-fusion: Tightly integrating deep dense visual bundle adjustment with multiple sensors for large-scale localization and mapping. *IEEE Robotics and Automation Letters*, 2024. [5](#), [6](#), [7](#), [8](#)
- [38] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multi-vehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. [7](#), [8](#)
- [39] 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](#)
- [40] Alex Zihao Zhu, Nikolay Atanasov, and Kostas Daniilidis. Event-based visual inertial odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5391–5399, 2017. [2](#), [5](#)