

EVIT: Event-based Visual-Inertial Tracking in Semi-Dense Maps Using Windowed Nonlinear Optimization

Runze Yuan, Tao Liu, Zijia Dai, Yi-Fan Zuo, and Laurent Kneip

Abstract—Event cameras are an interesting visual exteroceptive sensor that reacts to brightness changes rather than integrating absolute image intensities. Owing to this design, the sensor exhibits strong performance in situations of challenging dynamics and illumination conditions. While event-based simultaneous tracking and mapping remains a challenging problem, a number of recent works have pointed out the sensor’s suitability for prior map-based tracking. By making use of cross-modal registration paradigms, the camera’s ego-motion can be tracked across a large spectrum of illumination and dynamics conditions on top of accurate maps that have been created a priori by more traditional sensors. The present paper follows up on a recently introduced event-based geometric semi-dense tracking paradigm, and proposes the addition of inertial signals in order to robustify the estimation. More specifically, the added signals provide strong cues for pose initialization as well as regularization during windowed, multi-frame tracking. As a result, the proposed framework achieves increased performance under challenging illumination conditions as well as a reduction of the rate at which intermediate event representations need to be registered in order to maintain stable tracking across highly dynamic sequences. Our evaluation focuses on a diverse set of real world sequences and comprises a comparison of our proposed method against a purely event-based alternative running at different rates.

I. INTRODUCTION

Event cameras have recently become popular in the vision and robotics community but remain less explored than their traditional camera correspondents. Unlike the latter, event cameras—also called Dynamic Vision Sensors (DVS)—react to brightness changes rather than integrating absolute image intensities. Each pixel acts asynchronously and fires an either positive or negative event whenever the perceived brightness level augments or decrements by a certain threshold amount. Each event comes with a high resolution time-stamp in the order of micro-seconds, and the camera has high dynamic range (120dB compared to 60dB for regular cameras). Owing to these properties, event cameras do not suffer from traditional blur effects and lend themselves to strong performance in challenging illumination and dynamics conditions. Often referred to as a silicon retina [1], this power-efficient bio-inspired sensor therefore has evolved into a promising choice for motion sensing in a variety of contexts, including eye tracking [2], hand tracking [3], human body tracking [4], [5], and—as targeted in this work—ego-motion tracking.

The traditional solution to vision-based ego-motion tracking is given by the Simultaneous Localization And Mapping (SLAM) paradigm [6], and has already been proposed for event cameras [7], [8]. However, unlike for regular cameras,

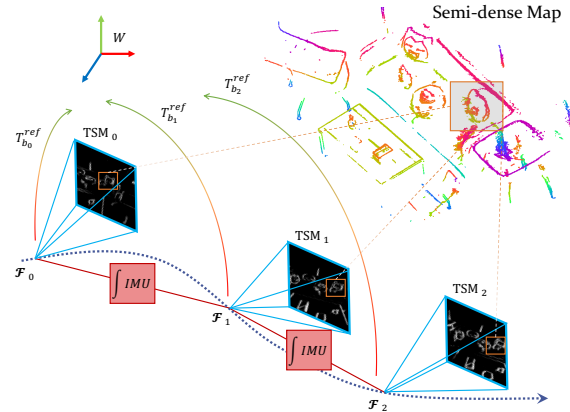


Fig. 1. Illustration of EVIT. Rather than registering only single time-surface maps with respect to a semi-dense point cloud, we propose to do windowed joint registration of multiple adjacent TSMs, which improves registration stability. The added IMU integration terms form the connections between adjacent keyframes, thereby creating a virtual multi-camera rig with elastic connections.

the measurements returned by an event camera are no longer just a function of pose and scene geometry or appearance, but also the instantaneous relative dynamics of the camera. Combined with the unusual and noisy nature of events, the solution to the full SLAM problem with DVS sensors remains an ongoing challenge.

In this paper, we address the question how we can solve the ego-motion estimation problem under the often practical assumption of an existing prior map generated by a different sensor? If the mapping problem is taken out of the equation, the target simply becomes prior map-based tracking. We follow up on the recently proposed strategy of abstracting the map of the environment in the form of a semi-dense point cloud [9]. The semi-dense map can be generated from regular imagery, and each point in the representation corresponds to a location in space around which we commonly observe high appearance gradient (i.e. structural edges and discontinuities, or textural appearance boundaries). Based on the assumption that the generated events pre-dominantly react to such high-gradient regions, the current location of the event camera can be optimized by aligning edges in the surface of active events with the reprojected semi-dense point cloud. While this semi-dense cross-modal registration paradigm has already been proposed in prior art [10], [9], challenges remain in case of high dynamics or ambiguous event distributions.

We propose to tackle this issue by fusion with an Inertial Measurement Unit (IMU). This achieves the following contributions and advancements:

- Rather than doing single frame registration, the addition of an IMU permits the joint registration of multiple adjacent frames in the form of a windowed tracking problem. The states corresponding to the individual times of each registered frame are lifted to include first-order translational velocity and IMU bias terms, thereby enabling the use of IMU pre-integration terms for a pair-wise regularization of adjacent frames. Sets of multiple adjacent frames are thereby considered as a virtual multi-camera rig with elastic connections. The joint registration of multiple frames helps in avoiding unobservable directions of displacement caused by ambiguous event distributions in individual frames (e.g. events distributed along multiple parallel lines).
- IMU pre-integration terms may again be used in order to create improved initialization of new frames. Especially in situations of jerky motion with sudden abrupt changes in the direction of motion, the inertial signal-based prediction strongly outperforms vision-only alternatives such as constant-velocity motion models.
- Owing to poor prediction abilities, handling situations of high acceleration in a vision only setting typically requires a high temporal density of tracked frames. In turn, the superior prediction ability offered by IMU signals enables a substantial reduction of the frame density during windowed optimization. In fact, we propose an adaptive mechanism that places frames in dependence of the number of occurred events.

We test our new framework—Event-based Visual-Inertial Tracking (EVIT)—on publicly available benchmark sequences covering a variety of challenges such as normal motion, changing illumination, and high dynamics. As demonstrated, the addition of an IMU strongly supports tracking performance with respect to the vision-only alternative. We also plan to release the complete framework code later to enable re-usability.

II. RELATED WORK

A good survey paper on event-based vision is given by Gallego et al. [11]. The addressed topics in the literature cover many aspects such as human motion tracking [4], [5], hand tracking [3], and eye ball tracking [2]. The current literature review focusses on SLAM and motion tracking with only event cameras or with added inertial readings.

Pure approaches to monocular event camera SLAM have been proposed by Kim et al. [7] and Rebecq et al. [8]. The former is a filter-based method reconstructing photometric depth maps, while the latter is a geometric approach that aims at the reconstruction of a volumetric event density field [12]. In an effort to increase the performance of event camera-based SLAM, more recent methods have investigated the use of line features [13], [14], [15]. The latter two methods are limited to local dynamics estimation, and all methods are restricted in that they require a sufficient number of line features. The inclusion of an IMU into the incremental tracking and motion framework is proposed by Zhu et al. [16], Rebecq et al. [17], Le Gentil et al. [18], and Xu et al. [19].

In particular, the latter two methods are again based on line features. In general, a truly robust continuous tracking and mapping framework that employs only a single event camera as an exteroceptive sensor remains a challenging problem. Vidal et al. [20] therefore investigate the combination with a normal camera, and Zhou et al. [21] present a stereo event camera alternative.

A different line of research that is more related to our work aims at pure tracking based on known structure priors. Early approaches make use of fiducial markers [22], [23] or known targets with distinctive texture [24], [25], [26], [27]. Of particular interest are the works of Gallego et al. [28] and Bryner et al. [29], who propose direct tracking of event cameras from photometric depth maps. However, their approaches are computationally demanding and fail to employ a geometrically optimal objective.

The most highly related works to ours are all based on the idea of using a semi-dense map representation. The latter have originally been introduced by Engel et al. [30] through their SDVO framework. Later on, Kneip et al. [31], Kuse and Shaojie [32], and Zhou et al. [33] have introduced improved geometric semi-dense registration paradigms for normal cameras. The latter work in particular has formed the foundation for the first geometric semi-dense tracking paradigms for event cameras. Zuo et al. [10] propose DEVO, an event-based incremental motion estimation framework that makes use of an additional depth channel. Finally, Zuo et al. [9] propose Canny-EVT, which applies a similar strategy to global tracking of a single event camera based on a semi-dense map. The latter work in particular forms the reference implementation against which we compare our contribution.

III. METHOD

In this section, we present our newly proposed event-inertial, semi-dense point cloud-based tracking system EVIT. The framework comprises an efficient adaptive measurement preprocessing module, a loosely coupled bootstrapping scheme, and a tightly coupled back-end optimization for the event camera-IMU setup. The structure of the proposed framework is shown in Fig. 2, in which we highlight the essential modules by dashed lines.

Let us define notations used in this paper. The ultimate goal for this framework is to estimate the body frame pose $\mathbf{T}_b^w = [\mathbf{R}_b^w, \mathbf{p}_b^w] \in \text{SE}(3)$ (transformation from the body frame to the world frame) of the event camera and IMU rig. For practical reasons, we set the IMU frame equal to the body frame, denoted by b . The constant extrinsic transformation \mathbf{T}_b^c between IMU and event camera is assumed to be well calibrated. Visual-inertial tracking introduces additional variables to the state estimation, leading to the expanded state vector

$$\mathcal{S}_i = [\boldsymbol{\theta}_i, \mathbf{p}_i, \mathbf{v}_i, \mathbf{b}_{\alpha_i}, \mathbf{b}_{\omega_i}], \quad (1)$$

where $\boldsymbol{\theta}_i$ is a minimal rotation representation, \mathbf{p}_i is the position of the body frame, \mathbf{v}_i is the velocity in the world reference frame, and \mathbf{b}_{α_i} and \mathbf{b}_{ω_i} are the accelerometer and gyroscope biases, respectively.

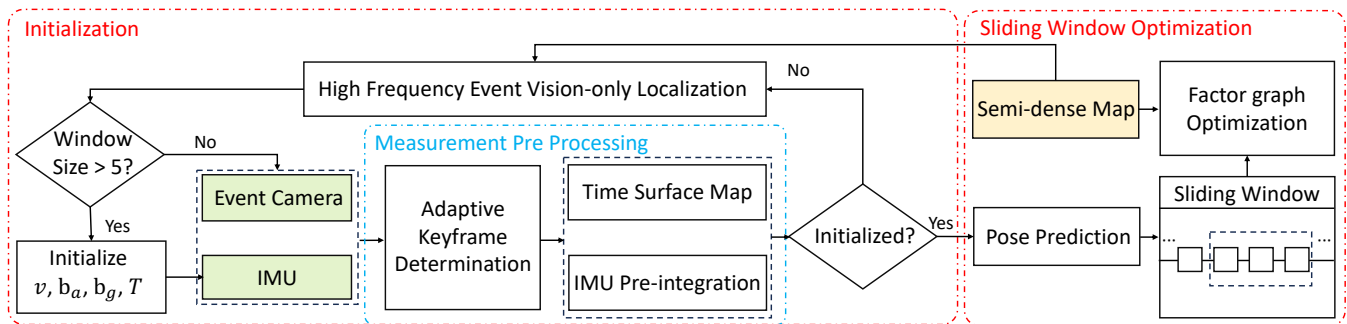


Fig. 2. Block diagram of the full event-based visual-inertial tracking pipeline. The system takes stream of events and IMU measurements (colored block) as input and tracks against the reconstructed semi-dense map. The measurement processing module (Section III-B) dynamically choose keyframes and process raw data stream into usable single frame observations. The initialization (Section III-C) module utilizes high frequency event localization results to provide bootstrapping states for subsequent prediction and optimization. The optimization module (Section III-D) tightly fuses IMU pre-integration measurements and TSM representations to achieve accurate state estimation.

A. System overview

We start by introducing the entire pipeline and describing the functionality of each core module in our framework. The framework starts with an offline-constructed, inertial-aligned semi-dense point cloud and takes as input a stream of events and IMU measurements from a calibrated event-inertial sensor setup. A specially designed data collection module is responsible for the asynchronous collection of these two types of data thereby achieving maximum effectiveness. The data collection module is followed by a measurement processing module that takes the quasi-continuous input stream of data and converts it into compact single frame observations (i.e. time surface maps and IMU pre-integration terms). The generated frames are henceforth referred to as keyframes. Details on the intermediate, frame-based event representation are provided in Section III-B. Note that the rate at which keyframes are generated is adaptive and depending on the actual number of events that have elapsed since the last keyframe generation. The generation of IMU pre-integration terms adjusts to the asynchronous keyframe rate. The proposed dynamic, adaptive frame generation strategy simultaneously satisfies two objectives. First, it avoids the generation of redundant keyframes when no actual change or camera motion has happened. Second, it avoids a lack of sufficient observations under fast displacements.

During the initialization phase (cf. Section III-C), we perform high-frequency event-only single keyframe localization to retrieve initial body frame poses. When the displacement is sufficient, a closed-form visual-inertial solver is executed in order to give coarse initial values for IMU states to be estimated, meaning the camera velocity and IMU biases. After successful bootstrapping, the tracking module (see Section III-D) starts to predict poses for new keyframes using a second-order motion model (i.e. using IMU forward integration) and activates a tightly coupled event-inertial back-end optimizer employing a sliding window over a fixed number of recent keyframes. Rather than only optimizing keyframe poses, this factor graph optimization module optimizes the full above-mentioned inertial states including camera velocity and biases for each keyframe. The addition

of inter-frame pre-integration terms adds regularization constraints between adjacent keyframes, and thereby transforms the registered set of keyframes into a virtual multi-camera rig with elastic extrinsic connections.

B. Measurement preprocessing

In traditional visual-inertial tracking frameworks, it is common to select keyframes based on image rate given that conventional cameras are synchronously triggered and usually operate at frequencies lower than IMUs. Considering a single event provides limited information, for event cameras we are equally inclined to preprocess events and somehow group them in batches that are triggered within a specific time interval so they can be jointly aligned against the map. In analogy to the work of Zuo et al. [9], we propose to summarize the contribution of entire intervals of events in snap-shots of the Time Surface Map (TSM), which can then be used to extract edges and perform individual frame alignment against the reprojected semi-dense map. The question is at what rate such TSM-based keyframes should be generated.

1) *Adaptive keyframe determination*: An event camera measures motion with respect to the scene, and the rate at which events are generated is related to the actual amount of displacement. The fixed frequency keyframe selection scheme proposed in previous works [21] therefore is sub-optimal and would lead to inferior optimization performance. It would either lead to redundant frames if little motion occurs, or insufficient frame density and large inter-keyframe baselines under fast displacements. Considering the IMU, the pre-integration also requires sufficient readings in order to ensure noise cancellation effects and good Signal-to-Noise Ratio (SNR). We therefore propose an adaptive keyframe generation scheme that buffers events and inertial readings and—only when a minimum number of both events and IMU readings has been surpassed—outputs the accumulated series of observations to the frame builder. The latter then processes the raw data into TSMs and IMU pre-integration terms. The thresholds for the minimum number of events and IMU observations are denoted n_{event} and n_{imu} , respectively.

The time stamp of the keyframe is determined by the last IMU measurement.

2) *Event representation*: As mentioned, we rely on snapshots of the Time Surface Map (TSM) for semi-dense registration. The latter is efficient to update and does not involve costly feature extractors. It is defined as follows. Let us assume a stream of N events denoted as $e_i = \{\mathbf{x}_i, t_i, p_i\}$, $i \in \mathbb{N}$, which includes location $\mathbf{x}_i = \{u_i, v_i\}$, timestamp t_i and polarity $p_i \in \{-1, 1\}$. A time surface map is a 2D map in which the value at a pixel \mathbf{x} is defined by the most recent event occurring at that location. It is notably depending on an exponential decay kernel defined as

$$\mathcal{T}(\mathbf{x}, t) \doteq \exp\left(-\frac{t - t_{\text{last}}(\mathbf{x})}{\delta}\right) \quad (2)$$

where δ is a constant decay rate parameter typically set to a small value. $t_{\text{last}}(\mathbf{x})$ is the timestamp corresponding to the most recent triggered event at \mathbf{x} . t is a given time stamp for which the TSM is constructed. Note that in the context of our work, the TSM can be regarded as the historical trajectory of scene edges moving in the image plane, and sharp contours remain identifiable at the reprojection location of the currently observed edges in the scene. It is this phenomenon that encourages a direct use of TSMs towards semi-dense registration.

For practical use, we usually use the inverse (or negated) TSM given by

$$\overline{\mathcal{T}}(\mathbf{x}, t) = 1 - \mathcal{T}(\mathbf{x}, t).$$

It has the advantage that the locations corresponding to the reprojection of the currently observed edges now present the lowest values in the map. As a result, the negated TSM can be directly used as a cost field against which the reprojected semi-dense map can be registered in nonlinear optimization. The relationship between an actual distance field and the negated TSM is introduced in [21]. We may readily perform nonlinear optimization on this geometric field.

For the sake of efficient measurement preprocessing, we maintain a global surface of active events (SAE) [34] which stores the most recent event time stamp at each pixel. We also map the pixel values in the TSM from $[0, 1]$ to $[0, 255]$ for easy visualization and stable numerical computation. We make use of the truncated TSM in which all values below a given threshold δ are set to 0. Finally, we apply a Gaussian kernel on the TSM to further smoothify the field, reduce ripple effects, and reduce the influence of outlier events.

3) *Imu pre-integration*: The IMU provides measurements for acceleration α_b and angular velocity ω_b of the sensor at a high frequent rate. We follow the IMU sensor model and continuous-time quaternion-based derivation of IMU pre-integration terms introduced in previous work [35] [36]. The raw IMU measurements $\hat{\alpha}_t, \hat{\omega}_t$ are given by

$$\begin{aligned} \hat{\alpha}_t &= \alpha_t + \mathbf{b}_{\alpha_t} + \mathbf{R}_w^t \mathbf{g}^w + \mathbf{n}_\alpha \\ \hat{\omega}_t &= \omega_t + \mathbf{b}_{\omega_t} + \mathbf{n}_\omega \end{aligned}$$

where noise is modeled as Gaussian white noise, i.e. $\mathbf{n}_\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$, $\mathbf{n}_\omega \sim \mathcal{N}(0, \sigma_\omega^2)$, and biases are modeled as a random walk.

The pre-integration terms between consecutive frames are given by

$$\begin{aligned} \alpha_{b_{k+1}}^{b_k} &= \iint_{t \in [t_k, t_{k+1}]} \mathbf{R}_t^{b_k} (\hat{\alpha}_t - \mathbf{b}_{\alpha_t} - \mathbf{n}_\alpha) dt^2 \\ \beta_{b_{k+1}}^{b_k} &= \int_{t \in [t_k, t_{k+1}]} \mathbf{R}_t^{b_k} (\hat{\alpha}_t - \mathbf{b}_{\alpha_t} - \mathbf{n}_\alpha) dt \\ \gamma_{b_{k+1}}^{b_k} &= \int_{t \in [t_k, t_{k+1}]} \frac{1}{2} \Omega(\hat{\omega}_t - \mathbf{b}_{\omega_t} - \mathbf{n}_\omega) \gamma_t^{b_k} dt \end{aligned}$$

where

$$\Omega(\omega) = \begin{bmatrix} -[\omega]_\times & \omega \\ \omega^T & 0 \end{bmatrix}, [\omega]_\times = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_x & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}$$

$\alpha_{b_{k+1}}^{b_k}, \beta_{b_{k+1}}^{b_k}$, and $\gamma_{b_{k+1}}^{b_k}$ represent the pre-integrated position, velocity and rotation measurements. A covariance matrix $\Sigma_{b_k, b_{k+1}}$ is also calculated during the pre-integration process. In our practical implementation, we use first-order integration for processing discrete time IMU measurements.

C. Initialization

Tightly coupled visual-inertial optimization-based tracking usually needs good initial values for states to achieve convergence. The goal of the initialization module is to align vision-based and IMU measurements to calibrate the bias parameters of the IMU and provide initial states. Inspired by [35], we employ a light-weight loosely coupled initialization method in our localization framework. In contrast to traditional VIO frameworks, the states to be bootstrapped in our localization system are fewer and only include velocity, biases and pose.

The initialization module proceeds by first constructing many intermediate TSMs between adjacent frames and then utilizing this information for high-frequency event-based vision-only localization. This procedure will give initial poses for each frame, and requires the solution of individual optimization problems that minimize the values of the negated TSM field at the reprojected semi-dense pixel locations for each frame. The problem formulation for an individual frame is given by

$$\arg \min_{\theta, \mathbf{p}} \sum_{P_j \in \mathcal{D}} \rho(\overline{\mathcal{T}}(\pi_e(\mathbf{R}_b^c \mathbf{R}^T(\theta)(\mathbf{P}_j - \mathbf{p}) + \mathbf{t}_e^c))), \quad (3)$$

where ρ is a robust loss function (e.g. Huber loss), π_e is a function which projects the 3D points in event camera frame onto the image plane, $\mathbf{R}(\cdot)$ is the function that maps a minimal rotation representation to a 3-by-3 rotation matrix, and $\mathbf{P}_j \in \mathcal{D}$ is a 3D world point within the current field of view of the event camera.

When sufficiently many frames for bootstrapping have been collected, we adopt the IMU measurement residual function from [35] and loosely align the IMU pre-integration

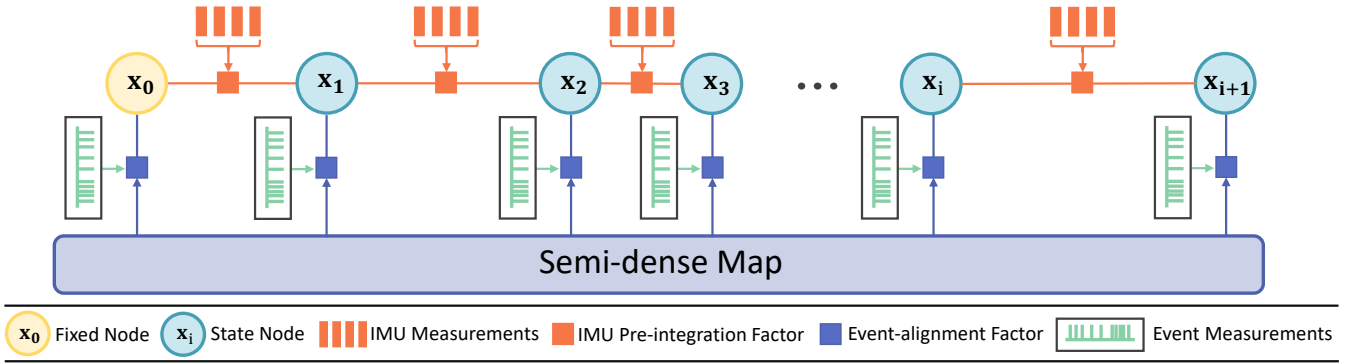


Fig. 3. Factor graph representation for our sliding window optimization. The graph fuses observations from the event camera, the IMU, and the semi-dense map. Two types of factors are introduced to construct the factor graph: (a) IMU pre-integration factors, (b) event alignment factors using TSMs. The formulation of these factors are discussed in Sections III-C and III-D. We fix the first node to maintain consistency with matured nodes that left the window.

terms with the event-based vision-only pose estimations. The IMU measurement residual function is given by

$$\mathbf{r}(\mathcal{S}_i, \mathcal{S}_{i+1}) = \begin{bmatrix} \mathbf{R}^T(\boldsymbol{\theta}_i)(\mathbf{p}_{i+1} - \mathbf{p}_i + \frac{1}{2}\mathbf{g}^w \Delta t_i^2 - \mathbf{v}_i \Delta t_i) - \hat{\boldsymbol{\alpha}}_{i+1}^i \\ \mathbf{R}^T(\boldsymbol{\theta}_i)(\mathbf{v}_{i+1} + \mathbf{g}^w \Delta t_i - \mathbf{v}_i) - \hat{\boldsymbol{\beta}}_{i+1}^i \\ 2[\mathbf{Q}(\boldsymbol{\theta}_i)^{-1} \otimes \mathbf{Q}(\boldsymbol{\theta}_{i+1}) \otimes (\hat{\boldsymbol{\gamma}}_{i+1}^i)^{-1}]_{xyz} \\ \mathbf{b}_{\alpha_{i+1}} - \mathbf{b}_{\alpha_i} \\ \mathbf{b}_{\omega_{i+1}} - \mathbf{b}_{\omega_i} \end{bmatrix}, \quad (4)$$

where $\mathbf{Q}(\cdot)$ transforms the minimal rotation representation to a quaternion and $\hat{\boldsymbol{\alpha}}_{i+1}^i, \hat{\boldsymbol{\beta}}_{i+1}^i, \hat{\boldsymbol{\gamma}}_{i+1}^i$ represent the IMU pre-integration vector using raw IMU measurements. Given N frames, the loosely coupled visual-inertial alignment objective is given by

$$\arg \min_{\mathbf{v}_i, \mathbf{b}_{\alpha_i}, \mathbf{b}_{\omega_i}} \sum_{i=1}^{N-1} \|\mathbf{r}(\mathcal{S}_i, \mathcal{S}_{i+1})\|_{\Sigma_{i,i+1}}^2. \quad (5)$$

Note that we fix the keyframe poses and optimize the remaining states only. Since the existing IMU pre-integrations are based on the old states, they should be re-propagated using the optimized state once the initialization is finished.

D. Sliding window optimization

The initialization module constructs a window of frames with initial states. Once given, we use a second-order motion model to obtain predictions for each new keyframe state. In practice, the field of view of frames within the sliding window typically only covers part of the semi-dense map. In order to reduce the complexity of the optimization, we maintain a set of activate points. The set will keep the points visible in the last sliding window optimization. We also construct a wider frustum for new keyframes using their predicted pose to select potentially visible new points and merge them into the activate points set. We randomly sample a fixed-size subset \mathcal{P} from this set of points to ensure the optimization maintain approximately constant time complexity.

Fig. 3 illustrates the factor graph of the concluding tightly-coupled sliding window optimization framework. The optimization is executed each time a new frame is added to the

window and the oldest frame is removed. It is furthermore important to remain consistent with mature nodes that already left the sliding window such that a smoother trajectory can be maintained. Here we choose a simple solution that consists of fixing the first node in the graph. Combining the residuals in (4) and (3), the objective of the event-based visual-inertial sliding window optimizer finally becomes

$$\arg \min_{\mathcal{S}_1, \dots, \mathcal{S}_N} \sum_{\mathbf{P}_j \in \mathcal{P}} \sum_{i=1}^N \rho(\bar{\mathcal{T}}(\pi_e(\mathbf{R}_b^c \mathbf{R}^T(\boldsymbol{\theta}_i)(\mathbf{P}_j - \mathbf{p}_i) + \mathbf{t}_b^c))) + \sum_{i=1}^{N-1} \|\mathbf{r}(\mathcal{S}_i, \mathcal{S}_{i+1})\|_{\Sigma_{i,i+1}}^2. \quad (6)$$

where N is the sliding window size and \mathcal{P} is the maintained activate point set. We use Ceres solver [37] to solve all above nonlinear problems.

IV. EXPERIMENTS

We now proceed to the experimental evaluation of our event-based visual-inertial tracking framework. All experiments are conducted on the *small scale* sequences of the publicly available **VECT**or dataset [38]. The dataset contains event stereo cameras (Prophesee Gen3 CD) with VGA resolution (640×480), only the left one of which we use, and a nine-axis IMU (XSens MTi-30 AHRS) at 200Hz. The ground truth is generated by the OptiTrack motion capture system. All sensors are well synchronised though a micro-controller unit at hardware level. We noticed that at certain moments in some sequences, there is only a small overlap between the event camera's field of view and the reconstructed semi-dense map, which leads to tracking failures for all methods. Therefore, we apply clipping as necessary on certain sequences. We first compare the proposed **EVIT** framework against several state-of-art event-based localization approaches on both normal and challenging sequences. To conclude, we conduct an ablation study using different motion models to analyze the impact of using IMU predictions on tracking performance. We provide both qualitative and quantitative results to show the robustness and effectiveness of our method.

We use the *evo* odometry evaluation tool [39] to align trajectories and conduct comparisons based on the RMSE of the average trajectory error (ATE) metric. The trajectories are aligned by an SE(3) transformation derived from the first pair of the synchronized camera and ground truth poses.

A. Evaluation on VECTOR

In the first experiment, we compare the performance of the proposed system with existing event-based localization frameworks. Note that the present work is an extension of the event vision tracking module Canny-EVT presented by Zuo et al. [9]. As mentioned in there, the **EVT** module relies on a high-frequency localization method which trades efficiency for robustness, thereby resulting in an offline method with no real-time capabilities. We compare our proposed **EVIT** against **EVT** running at different rates and on the same semi-dense map.

The global semi-dense map is constructed from the open-source framework [40], which uses the probabilistic semi-dense mapping technique proposed by Mur-Artal and Tardos [41]. Note that the constructed edges do not necessarily need to be geometric edges in 3D and may still be appearance edges. In order to avoid the scale ambiguity issue caused by monocular slam, we replace each keyframe pose by the corresponding ground truth pose. We also manually transform the reconstructed semi-dense map to a global world reference frame where the negative z -axis direction aligns with the direction of gravity to ensure consistency with IMU data.

TABLE I
ABSOLUTE TRAJECTORY ERROR (ATE) ON NORMAL SEQUENCES.
POSITION: [CM], ORIENTATION: [°]

Sequence	Method	Frequency	30% sequence		50% sequence		100% sequence	
			Pos.	Orient.	Pos.	Orient.	Pos.	Orient.
sofa_normal	EVT	300hz	3.26	1.34	3.66	1.75	3.37	1.65
		100hz	3.16	1.33	6.98	2.70	5.43	2.28
		30hz	3.54	1.58	-	-	-	-
	EVPT	(dyn)	15.64	5.08	15.27	5.50	-	-
	EVIT	(dyn)	3.23	1.40	3.23	1.56	3.15	1.53
robot_normal	EVT	300hz	0.91	0.89	1.04	1.08	1.09	1.31
		100hz	0.97	0.91	1.09	1.09	1.16	1.31
	EVPT	(dyn)	15.36	9.927	13.40	8.39	14.17	9.03
	EVIT	(dyn)	0.84	0.89	1.04	1.10	1.00	1.14
desk_normal	EVT	300hz	1.21	0.74	1.87	0.81	2.25	0.88
		100hz	1.34	0.77	1.87	0.84	2.14	0.92
	EVPT	(dyn)	7.71	1.83	11.02	2.61	16.08	3.33
	EVIT	(dyn)	1.32	0.80	1.79	0.87	2.22	0.94

We also compare the **EVIT** with the photometric event-based camera localization method proposed by Bryner et al. [29]. We refer to it here as **EVPT**. The photometric map is directly constructed using the RGBD data and ground truth poses contained in the dataset. We directly re-project all valid pixels back into the 3D space and carefully polish the map by hand to reduce edge effects caused by the RGBD observations.

Note that not all methods can track the entire sequence so we define three milestones at 30%, 50% and 100% of the sequence. The corresponding error is simply not displayed if

a certain milestone is not reached. Since **EVIT** and **EVPT** all adopt dynamic frame determination schemes, the rate depends on the motion and the scene. The actual rate is lower for mild motion and higher during aggressive motion, and typically varies between 30 and 60 Hz. As shown in Table. I, though the improvement brought by **EVIT** is not significant on normal sequences, both **EVT** and **EVIT** outperform **EVPT**. This is sensible given that photometric localization methods strongly rely on the consistency of the actual light conditions when the map was built, which is not only hard to achieve for varying times of the day, but also hard to achieve across different sensors. Furthermore, photometric methods involve error metrics with no direct geometric meaning, and it is thus hard to ensure optimality of the motion estimation. In terms of EVIT, it can be observed that the addition of the IMU arguably offers only a limited improvement over the event-only alternative when the motion is normal and not exhibiting strong dynamics. However, EVIT generally produces small errors, often leading to the absolutely best tracking result.

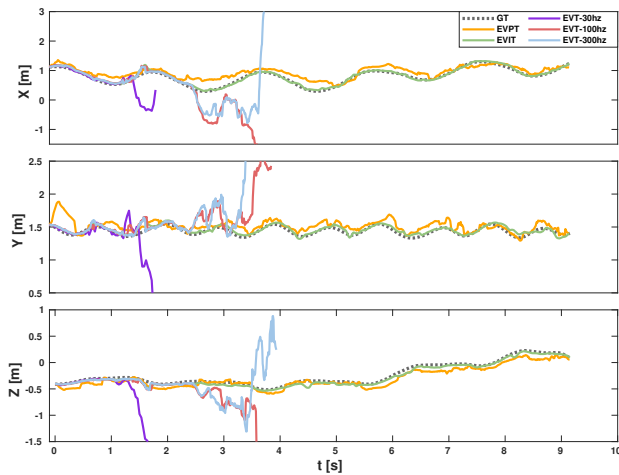
TABLE II
ABSOLUTE TRAJECTORY ERROR (ATE) ON CHALLENGING SEQUENCES.
POSITION: [CM], ORIENTATION: [°]

Sequence	Method	Frequency	30% sequence		50% sequence		100% sequence	
			Pos.	Orient.	Pos.	Orient.	Pos.	Orient.
sofa_fast	EVT	300hz	20.45	7.84	-	-	-	-
		100hz	22.66	8.56	-	-	-	-
		30hz	-	-	-	-	-	-
	EVPT	(dyn)	25.29	6.53	26.32	6.77	23.24	6.18
EVIT	(dyn)	9.02	4.02	8.11	3.60	7.08	3.22	
robot_fast	EVT	300hz	3.76	2.83	4.12	2.92	-	-
		100hz	5.47	3.67	9.15	4.49	-	-
		30hz	-	-	-	-	-	-
	EVPT	(dyn)	-	-	-	-	-	-
EVIT	(dyn)	3.22	2.67	3.24	2.80	3.85	2.98	
desk_fast	EVT	300hz	4.03	2.47	4.52	2.35	14.65	7.95
		100hz	2.24	2.19	2.48	2.52	-	-
		30hz	5.13	3.35	15.4	7.23	-	-
	EVPT	(dyn)	5.80	2.91	5.63	3.99	7.13	4.27
	EVIT	(dyn)	1.53	1.69	1.61	2.23	3.59	3.01

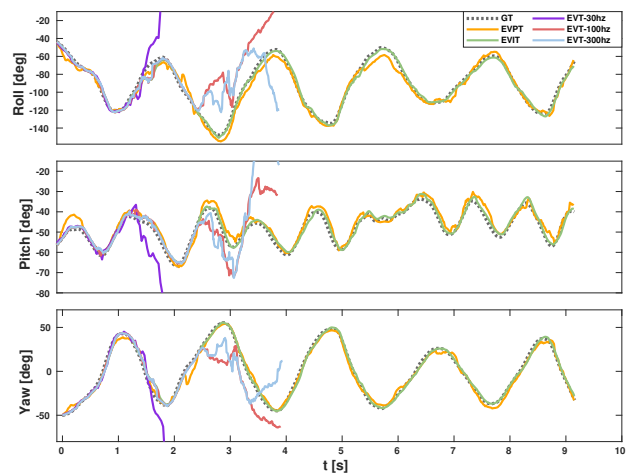
In contrast, the numerical results shown in Table. II illustrate a substantial improvement in accuracy and robustness in challenging scenarios. The majority of the alternative methods either fail to complete the entire sequence or result in large localization errors. Fig. 4 visualizes a comparison of actual trajectories from one of the more challenging sequences, demonstrating how all semi-dense comparison methods that do not employ inertial readings fail after at most 4s.

B. Ablation study on motion model

The tracking process in this work is based on the nonlinear optimization technique which is often sensitive to initial state given by motion model. Therefore, we also conduct an ablation study to evaluate the effect of different motion models. The latter can be adopted based on the system's available sensors. When there is no internal motion information, a simple strategy would be the initialize the pose of a newly incoming keyframe with the pose of the previous one, a paradigm we may refer to as a zeroth-order motion model.



(a) translation estimates



(b) rotation estimates

Fig. 4. Comparison of trajectories generated by various pure and inertial-supported semi-dense event-based tracking solutions, as well as a photometric alternative. The sequence is *sofa_fast* from the **VECTor** benchmark [38]. The method proposed in this work is called **EVIT**.

A more advanced model could be the first-order motion model (i.e. constant velocity model), which assumes that the velocity remains constant over small periods of time, and thereby permits the initialization of a new keyframe with a relative pose to the previous one that is similar to the relative pose between the previous keyframe and its preceding one. Equipped with an IMU, we can introduce the second-order motion model, which uses acceleration and angular velocity readings to propagate the previous pose to the current one.



(a) zeroth-order model (b) first-order model (c) second-order model

Fig. 5. Initial projection of semi-dense cloud on **TSM** using different motion models.

The three motion models are tested with the same tightly coupled optimization back end. Fig. 5 shows qualitative results for the different models, demonstrating how the second-order motion model provides better initial poses for the optimization back end. The full ATE results are listed in Table. III. Although the differences are not significant on mild sequences, it shows that even under the same optimization scheme, the motion model has a crucial implication in fast-motion scenarios.

V. CONCLUSION

In this work, we have introduced a novel event-inertial tracking framework that aligns with previous approaches in that it proposes the registration of a semi-dense point cloud with contours observed in the time surface map. However, rather than simply aligning individual frames, the introduction of inertial signals permits the joint alignment of multiple adjacent keyframes through the addition of pair-wise

TABLE III
ABSOLUTE TRAJECTORY ERROR (ATE) ON DIFFERENT MOTION MODELS. POSITION: [CM], ORIENTATION: [°]

Sequence	zeroth-order		first-order		second-order	
	Pos.	Orient.	Pos.	Orient.	Pos.	Orient.
sofa_fast	-	-	-	-	7.08	3.22
sofa_normal	-	-	2.84	1.60	3.15	1.53
robot_fast	-	-	-	-	3.85	2.98
robot_normal	1.08	1.32	1.46	1.23	1.00	1.14
desk_fast	-	-	-	-	3.59	3.01
desk_normal	3.75	1.59	2.77	1.16	2.22	0.94

regularization terms. The proposed sliding window alignment shows benefits in tracking robustness as the contours in a single TSM are often orthogonal to the projected instantaneous direction of motion, and thus do not fully constrain an individual pose. Furthermore, the addition of the IMU leads to better initializations and thereby improves tracking performance in highly dynamic situations without a substantial increase in frame rate. Interestingly, the proposed windowed tracking method could be equally applied to regular cameras.

VI. ACKNOWLEDGEMENT

We would like to acknowledge the funding support provided by project 62250610225 by the Natural Science Foundation of China, as well as projects 22DZ1201900, 22ZR1441300, and dfycbj-1 by the Natural Science Foundation of Shanghai.

REFERENCES

- [1] M. A. Mahowald and C. Mead, "The silicon retina," *Scientific American*, vol. 264, no. 5, pp. 76–83, 1991.

- [2] A. N. Angelopoulos, J. Martel, A. Kohli, J. Conradt, and G. Wetzstein, "Event-based near-eye gaze tracking beyond 10,000 Hz," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2577–2586, 2021.
- [3] V. Rudnev, V. Golyanik, J. Wang, H.-P. Seidel, F. Mueller, M. Elgharib, and C. Theobalt, "EventHands: Real-time neural 3d hand reconstruction from an event stream," 2021.
- [4] L. Xu, W. Xu, V. Golyanik, M. Habermann, L. Fang, and C. Theobalt, "Eventcap: Monocular 3d capture of high-speed human motions using an event camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [5] Z. Zhang, K. Chai, H. Yu, R. Majaj, F. Walsh, E. Wang, U. Mahbub, H. Siegelmann, D. Kim, and T. Rahman, "Neuromorphic high-frequency 3d dancing pose estimation in dynamic environment," *Neurocomputing*, vol. 547, p. 126388, 2023.
- [6] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics (T-RO)*, vol. 32, no. 6, pp. 1309–1332, 2016.
- [7] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3d reconstruction and 6-dof tracking with an event camera," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- [8] H. Rebecq, T. Horstschäfer, G. Gallego, and D. Scaramuzza, "Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 593–600, 2016.
- [9] Y. Zuo, W. Xu, X. Wang, Y. Wang, and L. Kneip, "Cross-Modal Semi-Dense 6-DoF Tracking of an Event Camera in Challenging Conditions," *IEEE Transactions on Robotics (T-RO)*, vol. 40, pp. 1600–1616, 2024.
- [10] Y.-F. Zuo, J. Yang, J. Chen, X. Wang, Y. Wang, and L. Kneip, "Devo: Depth-event camera visual odometry in challenging conditions," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2022.
- [11] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, and K. Daniilidis, "Event-based vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 44, no. 1, pp. 154–180, 2020.
- [12] H. Rebecq, G. Gallego, and D. Scaramuzza, "EMVS: Event-based multi-view stereo," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.
- [13] W. Chamorro, J. Sola, and J. Andrade-Cetto, "Event-based line slam in real-time," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8146–8153, 2022.
- [14] X. Peng, W. Xu, J. Yang, and L. Kneip, "Continuous Event-Line Constraint for Closed-Form Velocity Initialization," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.
- [15] L. Gao, H. Su, D. Gehrig, M. Cannici, D. Scaramuzza, and L. Kneip, "A 5-Point Minimal Solver for Event Camera Relative Motion Estimation," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.
- [16] A. Z. Zhu, N. Atanasov, and K. Daniilidis, "Event-based visual inertial odometry," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5391–5399.
- [17] H. Rebecq, T. Horstschäfer, and D. Scaramuzza, "Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.
- [18] C. Le Gentil, F. Tschopp, I. Alzugaray, T. Vidal-Calleja, R. Siegwart, and J. Nieto, "IDOL: A framework for imu-dvs odometry using lines," in *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5863–5870.
- [19] W. Xu, X. Peng, and L. Kneip, "Tight Fusion of Events and Inertial Measurements for Direct Velocity Estimation," *IEEE Transactions on Robotics (T-RO)*, vol. 40, pp. 240–256, 2023.
- [20] A. R. Vidal, H. Rebecq, T. Horstschäfer, and D. Scaramuzza, "Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios," vol. 3, no. 2, pp. 994–1001, 2018.
- [21] Y. Zhou, G. Gallego, and S. Shen, "Event-based stereo visual odometry," *IEEE Transactions on Robotics (T-RO)*, vol. 37, no. 5, pp. 1433–1450, 2021.
- [22] A. Censi, J. Strubel, C. Brandli, T. Delbruck, and D. Scaramuzza, "Low-latency localization by Active LED Markers tracking using a Dynamic Vision Sensor," in *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2013.
- [23] G. Chen, W. Chen, Q. Yang, Z. Xu, L. Yang, J. Conradt, and A. Knoll, "A novel visible light positioning system with event-based neuromorphic vision sensor," *IEEE Sensors*, vol. 20, no. 17, pp. 10 211–10 219, 2020.
- [24] E. Mueggler, B. Huber, and D. Scaramuzza, "Event-based, 6-dof pose tracking for high-speed maneuvers," in *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2014.
- [25] J. Bertrand, A. Yigit, and S. Durand, "Embedded event-based visual odometry," in *IEEE International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP)*, 2020.
- [26] W. Chamorro, J. Andrade-Cetto, and J. Solà, "High-speed event camera tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.
- [27] D. Reverter Valeiras, G. Orchard, S.-H. Ieng, and R. B. Benosman, "Neuromorphic event-based 3d pose estimation," *Front. Neurosci.*, vol. 9, p. 522, 2016.
- [28] G. Gallego, J. E. A. Lund, E. Mueggler, H. Rebecq, T. Delbruck, and D. Scaramuzza, "Event-based, 6-dof camera tracking from photometric depth maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 40, no. 10, pp. 2402–2412, 2018.
- [29] S. Bryner, G. Gallego, H. Rebecq, and D. Scaramuzza, "Event-based, direct camera tracking from a photometric 3d map using nonlinear optimization," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
- [30] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013.
- [31] L. Kneip, Z. Yi, and H. Li, "SDICP: Semi-dense tracking based on iterative closest points," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.
- [32] M. Kuse and S. Shen, "Robust camera motion estimation using direct edge alignment and sub-gradient method," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2016.
- [33] Y. Zhou, H. Li, and L. Kneip, "Canny-vo: Visual odometry with rgb-d cameras based on geometric 3-d-2-d edge alignment," *IEEE Transactions on Robotics (T-RO)*, vol. 35, no. 1, pp. 184–199, 2018.
- [34] R. Benosman, C. Clercq, X. Lagorce, S.-H. Ieng, and C. Bartolozzi, "Event-based visual flow," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 2, pp. 407–417, 2013.
- [35] T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics (T-RO)*, vol. 34, no. 4, pp. 1004–1020, 2018.
- [36] S. Shen, N. Michael, and V. Kumar, "Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft MAVs," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015.
- [37] S. Agarwal, K. Mierle, and T. C. S. Team, "Ceres Solver," 10 2023. [Online]. Available: <https://github.com/ceres-solver/ceres-solver>
- [38] L. Gao, Y. Liang, J. Yang, S. Wu, C. Wang, J. Chen, and L. Kneip, "Vector: A versatile event-centric benchmark for multi-sensor slam," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8217–8224, 2022.
- [39] M. Grupp, "evo: Python package for the evaluation of odometry and SLAM," <https://github.com/MichaelGrupp/evo>, 2017.
- [40] S. He, X. Qin, Z. Zhang, and M. Jagersand, "Incremental 3d line segment extraction from semi-dense slam," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1658–1663.
- [41] R. Mur-Artal and J. D. Tardós, "Probabilistic semi-dense mapping from highly accurate feature-based monocular slam," in *Proceedings of Robotics: Science and Systems (RSS)*, 2015.