Event6D: Event-based Novel Object 6D Pose Tracking

Jae-Young Kang[1*] Hoonehee Cho[1*] Taeyeop Lee[1*]
Minjun Kang[1] Bowen Wen[2] Youngho Kim[1] Kuk-Jin Yoon[1]
[1]KAIST [2]NVIDIA

Abstract

Event cameras provide microsecond latency, making them suitable for 6D object pose tracking in fast, dynamic scenes where conventional RGB and depth pipelines suffer from motion blur and large pixel displacements. We introduce EventTrack6D, an event-depth tracking framework that generalizes to novel objects without object-specific training by reconstructing both intensity and depth at arbitrary timestamps between depth frames. Conditioned on the most recent depth measurement, our dual reconstruction recovers dense photometric and geometric cues from sparse event streams. Our EventTrack6D operates at over 120 FPS and maintains temporal consistency under rapid motion. To support training and evaluation, we introduce a comprehensive benchmark suite: a large-scale synthetic dataset for training and two complementary evaluation sets, including real and simulated event datasets. Trained exclusively on synthetic data, EventTrack6D generalizes effectively to real-world scenarios without fine-tuning, maintaining accurate tracking across diverse objects and motion patterns. Our method and datasets validate the effectiveness of event cameras for event-based 6D pose tracking of novel objects. Code and datasets are publicly available at https://chohoonhee.github.io/Event6D.

1. Introduction

Estimating 6D object pose is a fundamental problem in computer vision. Early 6D object pose estimation methods focused on instance-level approaches [117, 132, 134], where models are trained and evaluated on specific object instances. Research then progressed to category-level pose estimation [59, 110, 119], enabling generalization across object categories. Recent studies have advanced toward novel object generalization [66, 83, 84] in the context of robotic applications [61], developing models that handle unseen objects. However, accurate pose estimation from a single frame is insufficient for real-world applications requiring temporal consistency and continuous tracking. This motivates 6D object pose tracking [91, 128, 131], which maintains consistent pose estimates across video sequences.

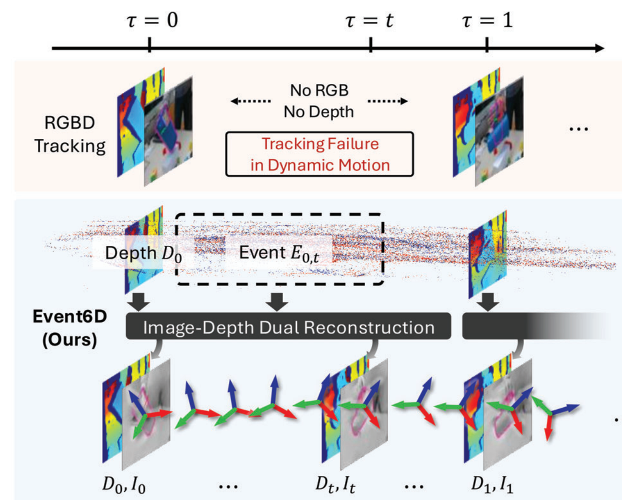

Figure 1. Conventional RGB-D based methods often fail under highly dynamic scenes due to limited frame rate from common RGB-D cameras. Our EventTrack6D addresses this issue by reconstructing dual modalities, image and depth, between consecutive depth frames to bridge the gap with event data. This enables inference at finer temporal intervals and yields robust tracking over highly dynamic motion.

Despite recent advances in novel object 6D pose tracking, existing methods [58, 81, 85, 128] and datasets [3, 8, 9, 39, 47, 120] rely on RGB or depth modalities at conventional frame rates (up to 30 FPS), limiting their applicability to dynamic scenes with fast motions where motion blur and large pixel displacements degrade performance. This limitation motivates the need for robust 6D pose tracking methods in high-speed scenarios.

Event cameras [31] emerged as a promising sensor for high-speed visual perception. Unlike conventional frame-based cameras [40, 54] that capture full images at fixed intervals, event cameras asynchronously record per-pixel brightness changes with microsecond temporal resolution, offering negligible motion blur, low latency, and high dynamic range. Several works [27, 44, 96] have investigated event-based 6D pose tracking tasks. RGB-DE [27] pioneered this area by introducing a single-object track-

*Equal contribution

Table 1. Dataset Comparison. Overview of publicly available event-based 6D object pose datasets. All datasets include RGB, Event, and depth data. n/a indicates that the paper does not provide the corresponding information.

Dataset	Events	# Samples	# Objects	Event Resolution	Annotation Frequency (Hz)	Motion	6D Pose Annotation
YCB-Ev [96]	Real	13,851	21	1280×720	30	Static	CosyPose [57]+ICG [107]
E-POSE [44]	Real	333,357	13	346×260	100	Moderate	Registration [88]+ICP [7]
RGB-DE [27]	Real	2,500	1	346×260	30	Dynamic	Manual+ICP [7]
EventBlender6D (Ours)	Synthetic	495,840	1033	640×480	60	Dynamic	BlenderProc [23]
EventHO3D (Ours)	Synthetic	103,462	5	640×480	n/a	Moderate	Multi-view Opt. [42]
Event6D (Ours)	Real	54,556	14	1280×720	120	Dynamic	Motion Capture [30]

ing dataset with an RGB-Depth-Event fusion method for instance-level 6D object pose tracking. However, their 6D pose annotations are limited to 30 Hz due to the RGB-D conventional frame rate. They show promising results by introducing RGB-Depth-Event fusion method for instance-level 6D object pose tracking. As summarized in Table 1, recent datasets, such as YCB-Ev [96] and E-Pose [44], introduce multi-object event-based 6D datasets but still remain constrained to controlled settings with moderate or static object motion. Moreover, their annotation pipelines are limited in handling dynamic objects, as they rely on existing RGB-D 6D pose methods (CosyPose [57] and ICG [107], or point cloud registration [88] with refinements [7]), which struggle with highly dynamic motions. Existing 6D event datasets are still limited in both scale and motion diversity, which are crucial for developing generalizable tracking methods.

To address these limitations, we propose EventTrack6D, an event-based 6D pose tracking framework that achieves robust tracking in high-speed scenarios and generalizes to novel unseen object instances without retraining. As shown in Fig. 1, our key idea is to reconstruct both intensity and depth at arbitrary timestamps between depth frames by leveraging event data conditioned on the most recent depth measurement. The reconstructed intensity and depth enable matching against CAD renderings, recovering photometric and geometric cues from sparse event data. This dual reconstruction provides dense geometric and photometric information for render-and-compare objectives [58, 128], allowing pose estimation at temporal resolutions beyond the native depth frame rate. EventTrack6D runs at over 120 FPS with a lightweight architecture and maintains robust tracking in highly dynamic scenes.

Alongside our method, we introduce a comprehensive dataset suite: EventBlender6D for large-scale synthetic training (495k samples, 1k objects), Event6D (real-world, motion-captured) and EventHO3D (simulated event) for evaluation. By training exclusively on synthetic data, we demonstrate strong cross-domain generalization to real-world scenarios without fine-tuning.

The core contributions can be summarized as follows:

- We propose EventTrack6D, an event camera-based 6D object pose tracking framework that generalizes to novel objects without retraining.
- We introduce a dual reconstruction approach that leverages event streams to reconstruct both intensity and depth between consecutive depth frames, which can be seamlessly integrated with the downstream module in a render-and-compare paradigm.
- We present large-scale synthetic and real-world event-based datasets (EventBlender6D, EventHO3D, and Event6D) for both training and evaluating event camera-based 6D object pose tracking (see Table 1).

2. Related Works

6D Object Pose Estimation and Refinement. 6D object pose estimation has been divided into instance-level [89, 117, 132], category-level [59, 60, 119], and novel object pose estimation [62, 84, 128]. To improve initial pose predictions, pose refinement methods [58, 64, 80, 81, 135] have been developed for each setting. However, most existing methods use RGB or depth as inputs and are limited by the sensing speed of common cameras. Under large and fast motion, these cameras produce severe motion blur in RGB frames and large inter-frame displacements, significantly degrading the performance.

6D Object Pose Tracking. Classical 6DoF tracking, including keypoint based [86, 97, 98, 105, 114], edge based [18, 25, 43, 102, 106], and direct optimization methods [6, 20, 76, 101, 111], struggles with textureless objects, clutter, and generalization. This has motivated learning based approaches [21, 32, 79, 124, 126], which often require large object specific datasets [2, 124]. Category-level [67, 118], model-based [49, 107, 131], and model-free trackers [125, 127] still depend on instance-level supervision, while recent work explores unseen object tracking [26] and iterative refinement [58, 64, 80]. In this context [58, 85, 128], our EventTrack6D targets tracking of novel objects given CAD models and leverages event cameras for robustness under rapid and large motions.

Event-based Image Reconstruction. Event-based image reconstruction is a well-established topic in previous works [5, 19, 34, 55, 56, 82, 99, 122]. Recent work uses deep models to produce high-quality reconstructions [94, 95, 100, 130, 133], though supervised approaches typically

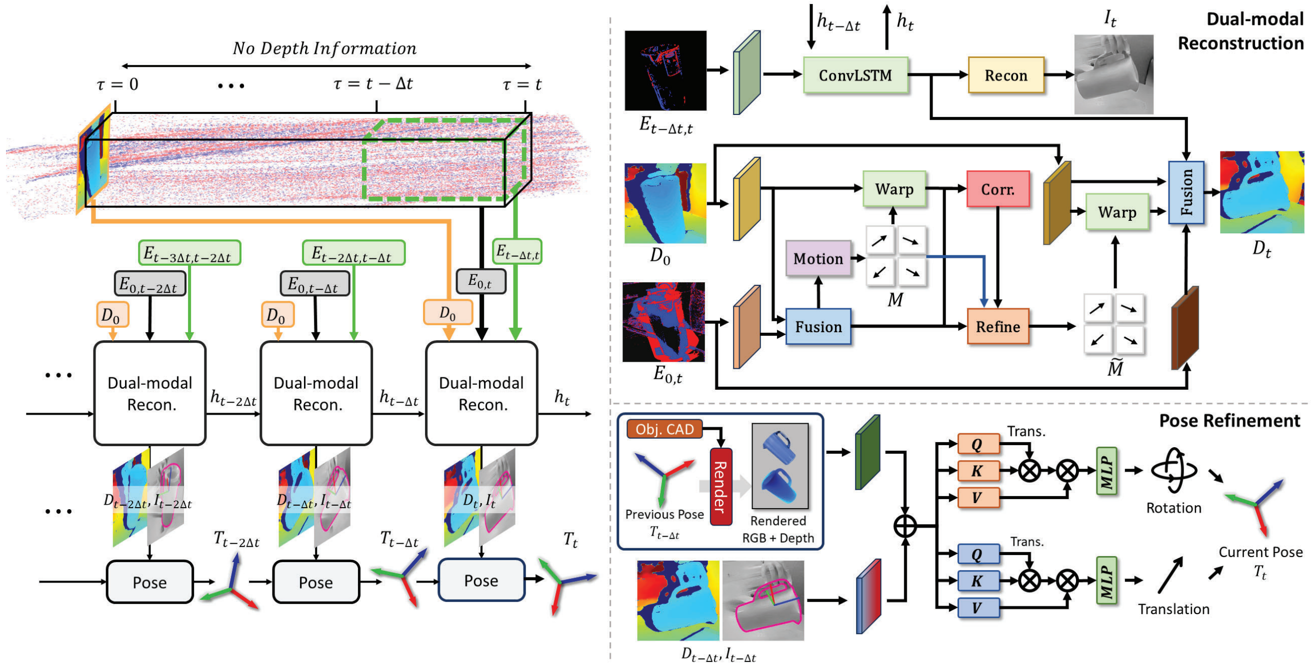

Figure 2. Overview of our EventTrack6D. EventTrack6D consists of a dual-modal reconstruction module and a pose refinement module. It can perform 6D pose tracking over high-frequency event stream, despite the limited frame rate of depth images which results in missing depth information between time intervals $\tau = 0$ and $\tau = t$. To achieve this, the dual-modal reconstruction module takes as input the most recent depth frame D_0, the event stream $E_{0,t}$ accumulated from that frame to the current time t where depth frame is missing, as well as the event stream $E_{t-\Delta t,t}$ from the most recent dual-modal reconstruction to the current time. From these inputs, it reconstructs the current intensity image I_t, and depth D_t. These reconstructed modalities are then used in a pose refinement module to estimate the 6D pose transformation from time $t-\Delta t$ to t.

require precisely aligned image–event pairs. Alternative paradigms have also been explored [29, 87, 90, 121], each introducing its own constraints. Notably, reconstructed images have been used for downstream tasks by leveraging dense photometric cues that sparse events lack [122, 123].

Event-based Depth Reconstruction. Depth estimation using event cameras has progressed rapidly [13, 45, 48, 52, 69, 140], yet many methods are too slow for real-time or struggle with absolute scale. Rather than regressing depth directly from events, we exploit motion cues in the event stream [36, 77, 103, 104, 116] and the high frame rate of event sensing [16, 33, 53]. Inspired by design principles of event-based video frame interpolation methods [10, 15, 78, 109, 112, 113, 137, 139], we propose an event-driven depth extrapolation that predicts the current depth from incoming events and the latest depth frame. Unlike interpolation, our method does not use future frames.

Event-based 6D Pose Estimation. Event cameras provide high-temporal-resolution and robustness to illumination [14, 50], enabling event-only ego-motion [1, 56, 92, 115], stereo and visual–inertial fusion [11, 68, 70, 138], and globally optimized SLAM [12, 38, 75]. These strengths have motivated 6DoF pose estimation via geometric line tracking [37, 71, 72], hybrid event–RGB pipelines [65], marker-based LEDs [28, 74], and stereo for spacecraft [73]. Despite this progress, event-based 6D tracking remains constrained by small, object-specific datasets collected in controlled settings [27, 44, 96] and a predominant focus on geometric formulations with few data-driven methods.

3. Approach

Problem Formulation. Given a CAD model of a rigid object and known camera intrinsics, our goal is to estimate the current object pose $\mathbf{T}_t = [\mathbf{R}_t \mid \mathbf{t}_t]$ in camera frame at an arbitrary time step t, where t is a normalized time in $[0, 1)$: $\tau = 0$ corresponds to the timestamp of the most recent depth frame, and $\tau = 1$ denotes the next depth frame in the future. $\mathbf{R}_t \in SO(3)$ and $\mathbf{t}_t \in \mathbb{R}^3$ represent its rotation and translation.

We assume access to an initial pose estimate $\mathbf{T}_0$, asynchronous event data, $E_{0,t} = \{e_i = (\mathbf{x}_i, p_i, \tau_i) \mid 0 \leq \tau_i < t\}$, and depth measurement D_0. Each event e_i is defined by its pixel location $\mathbf{x}_i = (x_i, y_i)^\top$, timestamp τ_i, and polarity $p_i \in \{-1, 1\}$, indicating the sign of the brightness change.
Method Overview. To achieve robust tracking in dynamic scenarios, we increase the temporal frequency of pose updates. While prior methods [58, 125, 127, 128] typically infer poses only at timestamps where a depth frame is available and thereby constrained by the limited frame rate from a depth camera, EventTrack6D performs tracking updates at finer temporal intervals. As shown in Fig. 2, our framework enables inference at an arbitrary time t using

two modules. The dual-modal reconstruction module predicts the current depth from the most recent depth frame and intervening events, while simultaneously reconstructing a dense intensity image. The pose-refinement module then uses the CAD model and the previous pose to estimate the pose at time t by matching against the reconstructed image and depth following a render-and-compare paradigm [57, 64, 85, 126, 128].

3.1. Dual-modal Reconstruction

Given the most recent depth frame D_0, the dual-modal reconstruction module processes two separate event streams with distinct roles. The long-range stream $E_{0,t}$, accumulated since D_0, provides motion cues primarily for depth reconstruction. The short-range stream $E_{t-\Delta t,t}$, collected over the most recent interval, focuses on fine temporal details for intensity image reconstruction. Using these inputs, the module reconstructs the intensity image I_t and depth map D_t for the current timestamp when these images are missing due to their limited frame rate. The process is illustrated in Fig. 2 (top-right).

Intensity Reconstruction. From the short-range event stream $E_{t-\Delta t,t}$, we extract spatio–temporal features with an event encoder for image reconstruction ϕ_E^I:

$$F^E_{t-\Delta t,t} = \phi_E^I(E_{t-\Delta t,t})\,.$$

Then, we employ a ConvLSTM to aggregate temporal context and produce temporally integrated event features:

$$\tilde{F}^E_{t-\Delta t,t},\, h_t = \mathrm{ConvLSTM}\big(F^E_{t-\Delta t,t},\, h_{t-\Delta t}\big)\,,$$

where $h_{t-\Delta t}$ is the hidden state propagated from the previous time step. Finally, the image decoder ψ_E^I fuses the aggregated event features to reconstruct the current intensity image:

$$I_t = \psi_E^I(\tilde{F}^E_{t-\Delta t,t}).$$

The recurrent hidden state h_t is propagated to the next time step, serving as temporal memory for the subsequent dual-modal reconstruction.

Depth Reconstruction. Given the most recent depth frame D_0 and the long-range event stream $E_{0,t}$, we first extract motion-related features using the event and depth encoders ϕ_E^M and ϕ_D^M: $F_{0,t}^{E,M} = \phi_E^M(E_{0,t})$ from the long-range events and $F_0^{D,M} = \phi_D^M(D_0)$ from the depth frame.

Then, we concatenate these motion-related features and apply a convolutional layer to generate the fused motion feature $\tilde{F}^M_{0,t}$. Next, the initial motion predictor, ψ^M, generates a coarse motion field from the fused features as $M^D_{0,t} = \psi^M(\tilde{F}^M_{0,t})$. Using the initial motion field, we warp the previous depth features to the current time:

$$F^{D,M}_{0\to t} = \mathrm{Warp}(F_0^{D,M},\, M^D_{0,t}). \tag{1}$$

We then compute a cost volume via a correlation layer [51, 108, 129] between the warped depth features and the fused motion feature:

$$\mathcal{C}_{0,t} = \mathrm{Corr}\Big(F^{D,M}_{0\to t},\, \tilde{F}^M_{0,t}\Big)\,. \tag{2}$$

Finally, we estimate a residual motion vector by concatenating the cost volume, the initial motion field, and the fused motion feature, followed by a convolutional layer, to obtain the refined motion field, $\tilde{M}^D_{0,t}$, by residual addition:

$$\begin{aligned} \Delta M^D_{0,t} &= \mathrm{Conv}(\mathrm{Concat}(\mathcal{C}_{0,t},\, M^D_{0,t},\, \tilde{F}^M_{0,t})),\\ \tilde{M}^D_{0,t} &= M^D_{0,t} + \Delta M^D_{0,t}. \end{aligned} \tag{3}$$

While the refined motion field enables temporal propagation of past depth information, geometric correction remains necessary to handle changes in 3D structure over time. To this end, we extract geometry-related features from long-range events and the depth frame using the encoders ϕ_E^G and ϕ_D^G, producing $F_{0,t}^{E,G} = \phi_E^G(E_{0,t})$ and $F_0^{D,G} = \phi_D^G(D_0)$. Using the refined motion field $\tilde{M}^D_{0,t}$, we warp the geometry-related depth features to the current time:

$$F^{D,G}_{0\to t} = \mathrm{Warp}(F_0^{D,G},\, \tilde{M}^D_{0,t}). \tag{4}$$

Event cues $F^{E,G}_{0,t}$ capture long-range motion but are inherently sparse and edge-dominated, which causes ambiguity for depth reconstruction in textureless regions or under large motion. To supply dense photometric context that encodes correspondences and regularizes geometry, we also use the temporally integrated event features from the image reconstruction stage $\tilde{F}^E_{t-\Delta t,t}$. We then combine $F^{E,G}_{0,t}$ with $\tilde{F}^E_{t-\Delta t,t}$ to reconcile changes in 3D structure over time. These representations are concatenated and refined, and the geometry module ψ^G produces the depth at time t:

$$D_t = \psi^G(\mathrm{Concat}(\tilde{F}^{D,G}_{0\to t},\, F_0^{D,G},\, F^{E,G}_{0,t},\, \tilde{F}^E_{t-\Delta t,t})). \tag{5}$$

3.2. 6D Pose Refinement

Dual-modal reconstruction predicts both the intensity image I_t and the depth map D_t at arbitrary timestamps. This allows pose tracking under large motion to be decomposed into a sequence of simpler subproblems with smaller motion.

$$\mathbf{T}_t = \left(\prod_{k=1}^{N} \mathbf{T}_{k-1\to k}\right)\mathbf{T}_0, \qquad \text{where } N = \frac{t}{\Delta t}. \tag{6}$$

Inspired by recent work in pose refinements [58, 128], the reconstructed intensity and depth images can be seamlessly integrated into a pose-refinement module that adopts a CAD-based render-and-compare paradigm to generalize to novel objects without retraining, as long as the object CAD model is given at test time.

We restrict pose updates to a region of interest (ROI) around the object. The crop is adapted from the previous pose estimate: its center is obtained by projecting the object origin onto the image plane, and its size is chosen to cover the object and its local context. Dual reconstruction is performed only within this ROI, reducing the computational cost of modality alignment.

The refinement module iteratively predicts a pose update that aligns rendered object views with the observed input. At each iteration, the current estimate is initialized from the previous pose, $\mathbf{T}_t = [R_t \mid t_t] \leftarrow [R_{t-\Delta t} \mid t_{t-\Delta t}]$, and independently updated as

$$\begin{aligned} R_t^+ &= \Delta R\, R_t, \\ t_t^+ &= t_t + \Delta t, \end{aligned} \quad \text{where } (\Delta R, \Delta t) = \mathcal{R}\big(\mathbf{T}_t, I_t, D_t\big). \tag{7}$$

I_t and D_t denote the predicted intensity image and depth map at time t, and $\mathcal{R}(\cdot)$ is the refinement operator.

3.3. Objective Function

Dual-modal Reconstruction Loss. The reconstructed intensity image I_t and depth map D_t are supervised using ground-truth data. For the image reconstruction, we apply a perceptual loss, LPIPS [136], that encourages photometric consistency with the ground-truth image I_t^{gt}:

$$\mathcal{L}_{\text{img}} = \text{LPIPS}(I_t, I_t^{gt}). \tag{8}$$

For depth reconstruction, we applied L_1 loss as:

$$\mathcal{L}_{\text{depth}} = \|D_t - D_t^{gt}\|_1. \tag{9}$$

The overall reconstruction objective is a weighted combination of both terms:

$$\mathcal{L}_{\text{recon}} = \lambda_I \mathcal{L}_{\text{img}} + \lambda_D \mathcal{L}_{\text{depth}}, \tag{10}$$

where λ_I and λ_D balance the contributions of photometric and geometric supervision.

6D Pose Refinement Loss. Pose refinement is optimized using an $L2$ loss:

$$\mathcal{L}_{pose} = \lambda_r \|\Delta R - \Delta R^\star\|_2 + \lambda_t \|\Delta t - \Delta t^\star\|_2 \tag{11}$$

where ΔR^t, Δt^* are ground truth and λ_r, λ_t are weights.

4. Dataset Generation and Acquisition

4.1. EventBlender6D Dataset

We present **EventBlender6D**, the first large-scale synthetic dataset for event-based object pose estimation and tracking. We build our pipeline with BlenderProc [22], enabling high-frame-rate RGB rendering and accurate annotations using Google Scanned Objects [24]. An event simulator [93] is then applied to the rendered sequences to produce synthetic event streams. The objects synthesized in the EventBlender6D dataset are disjoint from those in the evaluation datasets, Event6D and EventHO3D.

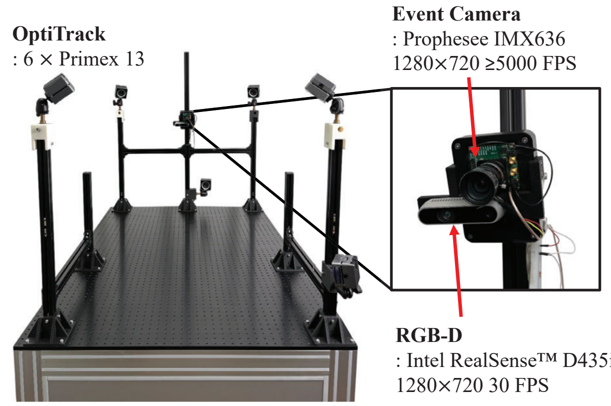

Figure 3. System designed for acquiring the Event6D dataset. The event camera, RGB-D camera, and motion capture system are all hardware-triggered, temporally synchronized and calibrated.

4.2. Event6D Dataset

We propose **Event6D** dataset, a real-world event-based 6D object pose dataset. As shown in Fig. 3, motion capture system is employed to obtain accurate annotations. The RGB-D camera, event camera, and motion capture system are hardware-triggered, ensuring precise temporal synchronization. Event6D provides high-quality 6D pose ground truth at 120 FPS, enabling reliable benchmarking of event-based high-frame-rate 6D object pose tracking methods. The RGB and depth streams were recorded at 30 FPS, following existing datasets [4, 9, 96]. Our Event6D dataset includes a subset of objects from YCB [8, 132] and HOGraspNet [17] datasets with existing CAD models. To incorporate novel objects, we captured real objects using a 3D scanner and generated corresponding CAD meshes. The dataset includes diverse and challenging scenarios. Further details are provided in the supplementary material.

4.3. EventHO3D Dataset

The HO3D dataset [42] provides video sequences commonly used for pose tracking evaluation. To assess the generalization of our method, we generate an event-based counterpart, EventHO3D, by simulating events using ESIM [93].

5. Experiments

Experimental Setup. We train our model on the synthetic EventBlender6D dataset and evaluate it (without any additional training or fine-tuning) on the real Event6D and synthetic EventHO3D datasets under the novel object setting [47, 81, 128], where all evaluation objects are unseen during training. Following [128], we assume that only the first frame of 6D pose is provided and evaluate long-term tracking robustness without re-initialization.

Evaluation is conducted for two different settings: 120 FPS and 30 FPS. In the 120 FPS setting, where RGB-D data (captured at 30 FPS) are unavailable, we compare our

Table 2. Comparison of event-based 6D object pose tracking methods on Event6D dataset, evaluated against ground-truth poses at 120 FPS. The event camera operates asynchronously at high frame rates, while RGB or depth is only available at 30 FPS. Therefore, conventional RGB or RGB-D-based methods cannot be evaluated under the 120 FPS ground-truths. The runtime is measured on preprocessed data, with a patch size of 160 × 160 corresponding to the region of interest in FP. † denotes that the model is trained for event inputs and bracket (·) in FPS indicates the runtime when applied TensorRT.

Approach	E2VID [95] + MG [58]			E2VID [95] + FP [128]			ETAP [41]			FP† [128]			**EventTrack6D (Ours)**		
Modality	Event + Depth			Event + Depth			Event + Depth			Event + Depth			Event + Depth		
Metric	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR
banana	0.00	0.00	0.17	24.92	9.65	25.38	6.09	1.54	7.11	13.99	1.70	9.43	43.46	16.26	41.83
bowl	4.43	0.00	17.41	46.20	0.57	77.11	48.20	1.25	75.21	2.70	0.38	8.98	54.86	0.34	85.65
cracker	0.00	0.00	0.03	36.65	28.42	42.05	1.17	1.04	1.31	45.53	27.28	60.42	74.65	62.44	89.19
drill	0.16	0.06	1.56	4.24	2.67	5.10	1.02	0.53	1.20	39.04	6.83	21.66	66.13	38.58	64.94
hammer	35.17	16.16	35.55	57.66	40.34	62.46	39.61	27.64	42.82	23.32	3.20	12.47	53.84	39.68	57.66
marker	12.04	0.83	17.32	0.17	0.17	0.17	0.17	0.17	0.17	12.33	1.02	8.44	62.36	28.01	65.92
mouse	2.10	0.43	10.74	25.48	3.31	63.32	0.13	0.13	0.36	8.60	0.81	9.89	34.54	4.93	70.48
mug	0.00	0.00	0.24	21.43	6.34	27.79	1.96	0.80	4.16	3.46	0.29	2.58	38.74	8.08	39.30
mustard	0.59	0.31	1.10	70.79	49.66	81.07	5.20	3.62	5.42	26.38	4.42	21.56	82.21	63.26	89.79
pitcher	12.83	4.52	17.70	72.35	43.29	74.51	8.44	5.12	8.19	58.15	17.66	47.99	71.47	42.86	81.02
scrub	3.93	1.26	3.98	76.67	54.53	82.57	1.21	1.01	1.44	61.01	22.19	46.54	84.55	65.97	91.63
spam	1.28	0.32	18.60	49.74	23.66	71.36	8.47	4.01	12.59	27.44	8.03	48.64	49.03	24.40	77.01
spatula	2.07	0.99	4.23	47.97	33.39	50.07	2.68	2.02	2.98	5.07	2.23	4.22	0.10	0.10	0.10
wine	0.18	0.18	0.27	0.18	0.18	1.23	0.18	0.18	0.19	16.66	0.94	19.87	17.87	0.48	26.21
MEAN	6.78	2.12	9.02	37.24	16.97	48.72	7.77	2.22	11.42	22.93	4.31	25.72	**52.79**	**25.26**	**64.38**
FPS (Hz)	10.92			79.37			1.56			**108.70**			50.19 (128.04)		

method against event-based baselines under high-frame-rate inference. For fair comparison with conventional RGB-D methods, we evaluate our approach at 30 FPS.

Evaluation Metrics. We report standard object pose metrics including the area under the curve (AUC) of ADD and ADD-S [46, 132], as well as the Average Recall (AR) [83] of Visible Surface Discrepancy (VSD), Maximum Symmetry-Aware Surface Distance (MSSD), and Maximum Symmetry-Aware Projection Distance (MSPD). Following previous studies [83, 128], we use a threshold of 0.1 times the object diameter for ADD and ADD-S, and varying thresholds for AR metrics. We measure inference time on an NVIDIA RTX A6000 GPU, enforcing CPU–GPU synchronization following prior work [35].

Event-based Baselines. To the best of our knowledge, there are no existing learning–based event-driven methods for novel 6D object pose tracking. We therefore compare two categories of baselines: (1) frame-based methods using event-to-image conversion and (2) event-based methods that operate directly on events. For the first category, we adapt state-of-the-art (SOTA) RGB-D-based 6D pose tracking methods, MegaPose (MG) [58] and FoundationPose (FP) [128], using E2VID [94] to convert event streams into images compatible with their pipelines. For the second category, we use event-based baselines that operate directly on event and depth. ETAP [41] tracks query points sampled from the CAD surface and estimates the 6D pose by fitting a rigid transformation [7, 63]. In addition, we fine-tune FP on the EventBlender6D dataset to support event inputs. All event-based baselines support both event-only and event-depth input. Details are in the supplementary material.

RGB-based Baselines. We also compare against the RGB-D SOTA pose-tracking methods MegaPose and FoundationPose, under 30 FPS settings only.

5.1. Comparison on Event6D dataset

Evaluation under the 120 FPS setting. In Table 2, we evaluate baselines for event-based methods at 120 FPS. Since depth is captured at 30 FPS (every fourth frame), baselines use depth and events when available, and events only at intermediate frames.

E2VID [94] + MG [58] and E2VID + FP [128] struggle with temporally sparse depth, leading to unstable pose tracking under rapid motion, as shown in Fig. 4. ETAP [41]-based point tracking with ICP [7] and PnP [63] also degrades under dynamic motion and occlusion, which limits overall performance. For fine-tuned FP using event inputs, the modality mismatch between CAD renderings and event streams, together with intermittent depth measurements, results in suboptimal performance.

Our method reconstructs intensity and depth at arbitrary timestamps, producing CAD-aligned observations that enable 6D pose tracking and robust performance across objects. The model runs at 50 FPS without optimization and exceeds 120 FPS with TensorRT.

Evaluation under the 30 FPS setting. In this section, we compare our method against RGB-D pose tracking baselines FP [128], and MG [58] at 30 FPS (Table 3), where RGB-D data are available. The event-based methods perform pose tracking at 120 FPS with the same configuration as in Table 2, but the pose estimates are evaluated at 30 FPS. Consistent with 120 FPS results (Table 2), EventTrack6D outperforms other event-depth baselines. Moreover, our method surpasses strong RGB-D baselines, MG [58] and FP [128]. Both strong RGB-D baselines often fail when faced with large inter-frame motion or severe motion blur,

Table 3. Comparison of event-based 6D pose tracking methods on Event6D dataset at 30 FPS. All methods are evaluated at RGB/depth frame intervals for fair comparison with RGB-based methods. † denotes models trained on event inputs.

Approach	MG [58]			FP [128]			E2VID [95] + MG [58]			E2VID [95] + FP [128]			FP† [128]			**EventTrack6D (Ours)**		
Modality	RGB + Depth			RGB + Depth			Event + Depth			Event + Depth			Event + Depth			Event + Depth		
Metric	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR
banana	24.87	11.78	22.56	20.52	10.23	17.61	0.00	0.00	0.26	16.32	5.64	14.91	24.96	2.11	15.39	35.68	13.58	33.21
bowl	1.13	0.47	2.08	36.14	0.97	21.18	0.78	0.00	11.04	48.43	1.07	73.95	3.80	0.38	10.31	63.19	1.41	88.18
cracker	5.45	2.66	5.68	32.05	24.65	40.10	0.00	0.00	0.15	69.29	52.49	79.59	53.20	34.33	64.55	76.40	65.88	90.06
drill	27.75	9.78	25.88	14.33	9.89	13.58	0.53	0.23	0.90	8.00	4.42	7.60	39.91	4.80	20.97	64.24	42.68	64.90
hammer	11.56	5.38	13.24	3.19	1.99	2.60	18.10	2.66	27.13	43.11	28.81	46.17	27.85	3.56	13.52	57.96	44.20	59.81
marker	7.55	1.24	5.23	6.31	2.86	6.84	2.57	0.68	6.51	51.38	19.04	50.19	14.05	1.37	8.41	49.94	23.68	49.28
mouse	0.00	0.00	0.72	1.04	0.52	0.92	0.81	0.00	6.13	26.25	3.85	57.37	10.82	1.23	10.58	34.30	4.06	70.95
mug	3.57	1.58	3.82	1.11	0.58	1.12	0.00	0.00	0.19	4.28	1.70	6.53	4.57	0.29	2.94	43.35	9.58	39.08
mustard	0.94	0.51	2.00	5.74	4.38	6.72	0.00	0.00	0.22	63.62	38.70	72.22	27.89	5.71	20.56	85.10	68.15	91.81
pitcher	20.39	12.38	20.69	55.66	40.23	59.25	0.93	0.00	8.28	69.00	37.25	69.72	63.28	23.83	51.28	77.81	54.42	85.17
scrub	4.60	2.30	5.32	5.76	4.40	5.22	2.25	0.49	2.24	76.20	53.88	82.02	67.73	31.41	53.54	89.65	77.70	94.10
spam	2.86	1.74	5.48	14.85	11.04	17.50	0.00	0.00	9.08	46.05	22.21	68.32	37.48	13.23	55.20	60.11	36.37	81.15
spatula	1.10	0.45	0.66	14.74	10.13	13.37	1.15	0.39	2.99	0.10	0.10	0.18	6.24	3.50	4.48	0.38	0.38	0.38
wine	1.07	0.00	0.80	0.73	0.73	0.73	0.73	0.73	0.47	0.36	0.18	2.43	14.68	1.51	19.12	1.65	0.73	1.80
MEAN	11.08	5.59	10.49	16.63	7.98	19.25	2.05	0.19	5.24	35.45	16.03	44.92	26.21	5.87	27.53	**56.08**	**29.40**	**63.71**

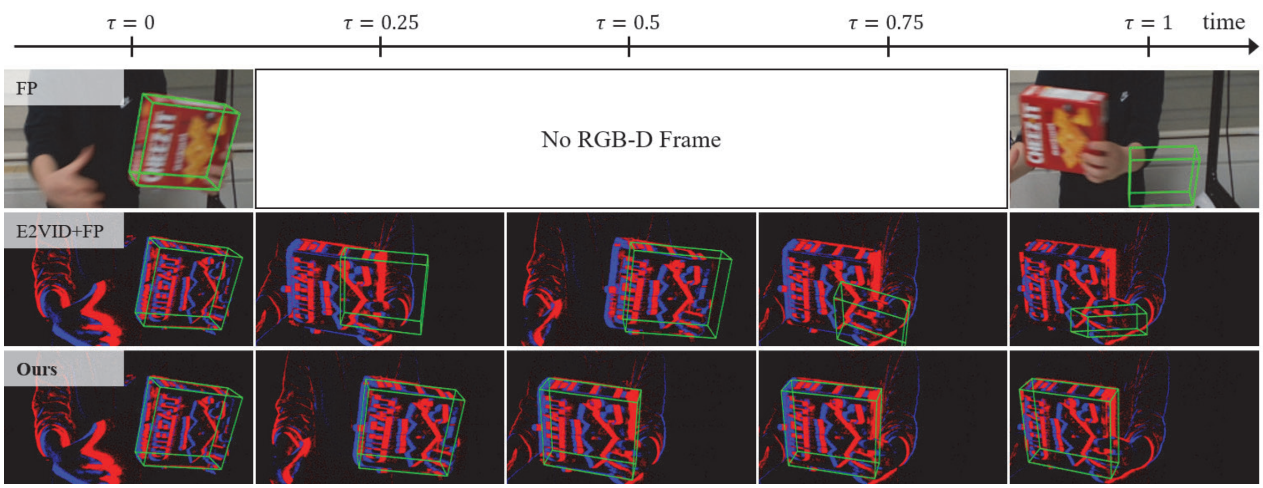

Figure 4. Qualitative comparison of 6D object tracking at 120 FPS on the Event6D dataset. Original FoundationPose(FP) [128] assumes RGB-D input and thus cannot be applied to a high frame rate setting. Note that for $\tau = 0.25, 0.5, 0.75$, ours utilizes its reconstructed depth rather than sensor-captured depth.

leading to unstable pose tracking.

Qualitative Results. Figure 4 shows 120 FPS pose-tracking results. The RGB-D baseline fails under large motions because it is unable to access depth frames for pose tracking due to limited frame rates. The hybrid E2VID [94]+FP [128] approach enables more frequent updates via image reconstruction but still degrades when depth is unavailable. In contrast, EventTrack6D reconstructs observations at arbitrary timestamps, enabling robust tracking across diverse motion conditions.

5.2. Comparison on EventHO3D Dataset

We further evaluate the baselines on a different domain, the EventHO3D dataset, which differs from Event6D in both motion distributions and sensor settings, providing an additional test of domain generalization.

As shown in Table 4, EventTrack6D generalizes across datasets, even when trained solely on the synthetic Event-Blender6D dataset. For comparison, we also report results for other event-based baselines, evaluated under a 30-FPS protocol following Table 3. Because EventHO3D exhibits more moderate motion than Event6D, tracking failures are less frequent even when depth is missing at intermediate timesteps. Nevertheless, by jointly reconstructing intensity and depth, EventTrack6D outperforms all baselines.

5.3. Ablation Study and Analysis

Dual-modal Reconstruction. We conduct an ablation study on the 120 FPS Event6D dataset, as summarized in Table 5. In the baseline configuration, both depth and image reconstruction are removed: depth is used only when available at 30 FPS, and the refinement module is trained on EventBlenderProc to handle event inputs directly without image reconstruction. This setup suffers from limited

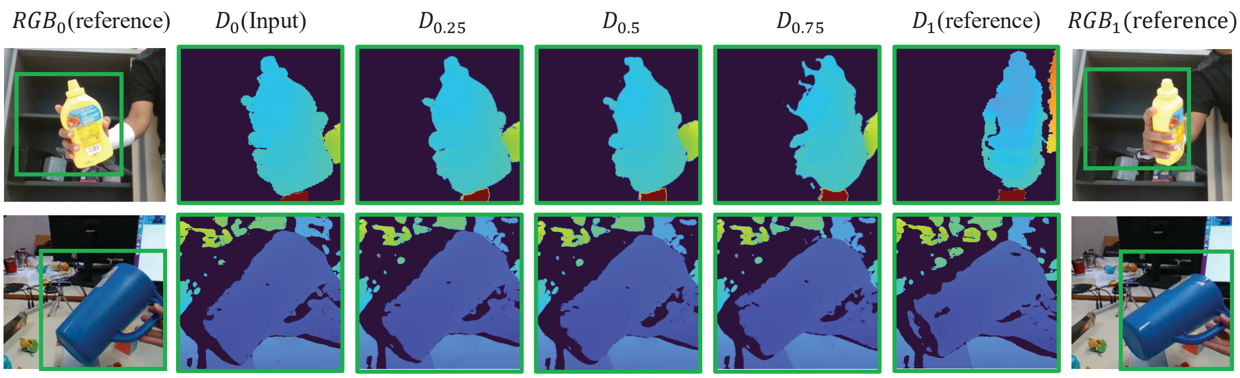

Figure 5. Qualitative depth-reconstruction results on depth-absent intervals. The future depth D_1 is provided solely for reference and is not used by the method. Despite dynamic motion, our approach reconstructs depth images that preserve coherent object structure and align with the object motion, providing geometric guidance for downstream pose tracking.

Table 4. Comparison on the EventHO3D dataset. † denotes that the model is trained for event inputs.

Method	E2VID [95] + FP [128]			FP† [128]			**EventTrack6D (Ours)**		
Modality	Event + Depth			Event + Depth			Event + Depth		
Metric	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR
AP10	52.60	34.86	46.75	10.41	0.63	4.76	88.66	77.06	90.08
AP11	82.55	57.65	83.16	49.62	13.89	38.62	42.51	29.09	40.52
AP12	70.74	26.00	50.48	64.96	24.05	55.36	85.29	64.90	82.54
AP13	80.93	51.59	77.83	59.57	17.37	41.34	78.17	42.13	66.02
AP14	53.52	42.28	51.67	12.06	6.44	10.75	85.09	67.78	83.06
SM1	41.95	33.78	45.31	0.64	0.64	0.64	85.75	74.99	85.79
SB11	45.08	36.44	47.24	43.21	5.93	23.50	88.17	75.54	90.82
SB13	44.08	37.54	46.37	38.30	7.62	36.82	92.00	83.96	95.42
MPM10	65.58	55.11	70.13	36.48	9.92	23.80	40.31	33.21	43.03
MPM11	53.76	29.30	45.18	2.05	0.64	1.11	43.03	36.60	46.90
MPM12	32.73	25.99	32.98	58.35	12.83	43.96	43.94	38.27	46.78
MPM13	91.30	84.18	95.18	64.50	34.79	57.07	48.42	39.77	49.59
MPM14	35.33	29.31	34.63	36.14	2.27	11.25	43.08	34.79	45.46
MEAN	57.10	41.13	56.77	35.31	8.83	27.64	**64.75**	**50.95**	**66.27**

Table 5. Ablation study of the dual-modal reconstruction.

Depth Recon.	Image Recon.	ADD-S↑	ADD↑	AR↑
		18.45	3.29	20.07
✓		28.67	4.75	29.08
	✓	30.53	13.79	44.99
✓	✓	52.79	25.26	64.38

geometric information between sparse depth frames and a modality mismatch between CAD renderings and event observations, ultimately leading to degraded performance.

When only depth reconstruction is added, tracking at arbitrary timestamps becomes feasible by injecting geometric cues, yet training remains challenging due to the persistent mismatch between event inputs and CAD renderings. When only image reconstruction is considered, photometric alignment improves, but the lack of continuous geometric information limits robustness under dynamic motion.

In contrast, our dual-modal reconstruction produces observations that align well with CAD renderings at arbitrary timestamps, providing both geometric and photometric cues, thereby achieving consistently superior performance across diverse motion conditions.

Table 6. Ablation study of depth-reconstruction components.

Motion Vector	Image Feature	ADD-S↑	ADD↑	AR↑
		36.03	17.52	43.60
	✓	42.26	19.32	48.99
✓		41.88	17.61	50.54
✓	✓	52.79	25.26	64.38

Depth Reconstruction. Table 6 summarizes ablations on the depth reconstruction module. Incorporating image features provides dense visual context that complements sparse event features, improving reconstruction by adding foreground and texture information. Motion vectors capture object-centric dynamics and enable accurate estimation of geometric changes under diverse motion patterns. Fusing image features with motion vectors improves depth predictions that more accurately represent real-world geometry.

Figure 5 shows reconstruction results at intervals between depth observations($\tau = 0$, $\tau = 1$). The module produces depth maps that are consistent with available observations and pixel-aligned with RGB images. Such realistic depth is input to the pose refinement stage and significantly enhances event-based 6D pose tracking.

6. Conclusion

In this paper, we explore the problem of event-based 6D object pose tracking. Due to the lack of large-scale datasets for training and evaluation, we introduce three datasets: EventBlender6D, EventHO3D, and Event6D. Moreover, we propose the EventTrack6D framework for novel 6D object pose tracking. Our efficient event-aware design processes 6D pose tracking at 128 FPS. Our experiments demonstrate strong generalization capability in 6D object pose tracking tasks, effectively handling the unique characteristics of event cameras. We believe this work will foster further research on event-based perception and high-speed 6D pose tracking.

Acknowledgments This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25473963), the InnoCORE program of the Ministry of Science and ICT(26-InnoCORE-01), and the InnoCORE program of the Ministry of Science and ICT(N10250156).

References

[1] Shrutarv Awasthi, Anas Gouda, Sven Franke, Jérôme Rutinowski, Frank Hoffmann, and Moritz Roidl. Mtevent: A multi-task event camera dataset for 6d pose estimation and moving object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5102–5110, 2025. 3

[2] Seunghyeok Back, Joosoon Lee, Kangmin Kim, Heeseon Rho, Geonhyup Lee, Raeyoung Kang, Sangbeom Lee, Sangjun Noh, Youngjin Lee, Taeyeop Lee, et al. Graspclutter6d: A large-scale real-world dataset for robust perception and grasping in cluttered scenes. *IEEE Robotics and Automation Letters*, 2025. 2

[3] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 1

[4] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7071, 2025. 5

[5] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 884–892, 2016. 2

[6] Selim Benhimane and Ezio Malis. Real-time image-based tracking of planes using efficient second-order minimization. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, pages 943–948. IEEE, 2004. 2

[7] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 2, 6

[8] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015. 1, 5

[9] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021. 1, 5

[10] Jingxi Chen, Brandon Y Feng, Haoming Cai, Tianfu Wang, Levi Burner, Dehao Yuan, Cornelia Fermuller, Christopher A Metzler, and Yiannis Aloimonos. Repurposing pretrained video diffusion models for event-based video interpolation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12456–12466, 2025. 3

[11] Peiyu Chen, Weipeng Guan, and Peng Lu. Esvio: Event-based stereo visual inertial odometry. *IEEE Robotics and Automation Letters*, 8(6):3661–3668, 2023. 3

[12] Yang Chen and Lin Zhang. Gre-slam: 6-dof pure event-based slam with semi-dense depth recovery assisted bundle adjustment. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 90–98, 2025. 3

[13] Hoonhee Cho, Jae-Young Kang, and Kuk-Jin Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *European Conference on Computer Vision*, pages 294–314. Springer, 2024. 3

[14] Hoonhee Cho, Taewoo Kim, Yuhwan Jeong, and Kuk-Jin Yoon. A benchmark dataset for event-guided human pose estimation and tracking in extreme conditions. *Advances in Neural Information Processing Systems*, 37:134826–134840, 2024. 3

[15] Hoonhee Cho, Taewoo Kim, Yuhwan Jeong, and Kuk-Jin Yoon. Tta-evf: Test-time adaptation for event-based video frame interpolation via reliable pixel and sample estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25701–25711, 2024. 3

[16] Hoonhee Cho, Jae-young Kang, Youngho Kim, and Kuk-Jin Yoon. Ev-3dod: Pushing the temporal boundaries of 3d object detection with event cameras. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27197–27210, 2025. 3

[17] Woojin Cho, Jihyun Lee, Minjae Yi, Minje Kim, Taeyun Woo, Donghwan Kim, Taewook Ha, Hyokeun Lee, Je-Hwan Ryu, Woontack Woo, et al. Dense hand-object (ho) graspnet with full grasping taxonomy and dynamics. In *European Conference on Computer Vision*, pages 284–303. Springer, 2024. 5

[18] Andrew I Comport, Eric Marchand, Muriel Pressigout, and Francois Chaumette. Real-time markerless tracking for augmented reality: the virtual visual servoing framework. *IEEE Transactions on visualization and computer graphics*, 12(4):615–628, 2006. 2

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

[20] Alberto Crivellaro and Vincent Lepetit. Robust 3d tracking with descriptor fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3414–3421, 2014. 2

[21] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao–blackwellized

particle filter for 6-d object pose tracking. *IEEE Transactions on Robotics*, 37(5):1328–1342, 2021. 2

[22] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Dmitry Olefir, Tomas Hodan, Youssef Zidan, Mohamad Elbadrawy, Markus Knauer, Harinandan Katam, and Ahsan Lodhi. Blenderproc: Reducing the reality gap with photorealistic rendering. In *16th Robotics: Science and Systems, RSS 2020, Workshops*, 2020. 5

[23] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Wendelin Knauer, Klaus H Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 2

[24] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. Ieee, 2022. 5

[25] Tom Drummond and Roberto Cipolla. Real-time visual tracking of complex structures. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24(07):932–946, 2002. 2

[26] Yuming Du, Yang Xiao, Michael Ramamonjisoa, Vincent Lepetit, et al. Pizza: A powerful image-only zero-shot zero-cad approach to 6 dof tracking. In *2022 International Conference on 3D Vision (3DV)*, pages 515–525. IEEE, 2022. 2

[27] Etienne Dubeau, Mathieu Garon, Benoit Debaque, Raoul de Charette, and Jean-François Lalonde. Rgb-de: Event camera calibration for fast 6-dof object tracking. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–135. IEEE, 2020. 1, 2, 3

[28] Gerald Ebmer, Adam Loch, Minh Nhat Vu, Roberto Mecca, Germain Haessig, Christian Hartl-Nesic, Markus Vincze, and Andreas Kugi. Real-time 6-dof pose estimation by an event-based camera using active led markers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8137–8146, 2024. 3

[29] Gereon Fox, Xingang Pan, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. Unsupervised event-based video reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4179–4188, 2024. 3

[30] Joshua S Furtado, Hugh HT Liu, Gilbert Lai, Herve Lacheray, and Jason Desouza-Coelho. Comparative analysis of optitrack motion capture systems. In *Advances in Motion Sensing and Control for Robotic Applications: Selected Papers from the Symposium on Mechatronics, Robotics, and Control (SMRC'18)-CSME International Congress 2018, May 27-30, 2018 Toronto, Canada*, pages 15–31. Springer, 2019. 2

[31] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (1):154–180, 2020. 1

[32] Mathieu Garon and Jean-François Lalonde. Deep 6-dof tracking. *IEEE transactions on visualization and computer graphics*, 23(11):2410–2418, 2017. 2

[33] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014): 1034–1040, 2024. 3

[34] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–765, 2018. 2

[35] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13884–13893, 2023. 6

[36] 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. 3

[37] Arren Glover, Luna Gava, Zhichao Li, and Chiara Bartolozzi. Edopt: Event-camera 6-dof dynamic object pose tracking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18200–18206. IEEE, 2024. 3

[38] Weipeng Guan, Fuling Lin, Peiyu Chen, and Peng Lu. Deio: Deep event inertial odometry. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4606–4615, 2025. 3

[39] Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11428–11435. IEEE, 2023. 1

[40] Hussein Haggag, Mohammed Hossny, Despina Filippidis, Douglas Creighton, Saeid Nahavandi, and Vinod Puri. Measuring depth accuracy in rgbd cameras. In *2013, 7th international conference on signal processing and communication systems (ICSPCS)*, pages 1–7. IEEE, 2013. 1

[41] Friedhelm Hamann, Daniel Gehrig, Filbert Febryanto, Kostas Daniilidis, and Guillermo Gallego. Etap: Event-based tracking of any point. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27186–27196, 2025. 6

[42] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 2, 5

[43] Chris Harris and Carl Stennett. Rapid-a video rate object tracker. In *BMVC*, page 3, 1990. 2

[44] Oussama Abdul Hay, Xiaoqian Huang, Abdulla Ayyad, Eslam Sherif, Randa Almadhoun, Yusra Abdulrahman, Lakmal Seneviratne, Abdulqader Abusafieh, and Yahya Zweiri.

E-pose: A large scale event camera dataset for object pose estimation. *Scientific data*, 12(1):245, 2025. 1, 2, 3

[45] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *2020 International Conference on 3D Vision (3DV)*, pages 534–542. IEEE, 2020. 3

[46] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012. 6

[47] Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2024. 1, 5

[48] Chenming Hu, Junjie Jiang, Yaohui Li, Mingyuan Sun, and Zheng Fang. Ede-distill: Boosting event-based monocular depth estimation performance via knowledge distillation. *IEEE Robotics and Automation Letters*, 2025. 3

[49] Jan Issac, Manuel Wüthrich, Cristina Garcia Cifuentes, Jeannette Bohg, Sebastian Trimpe, and Stefan Schaal. Depth-based object tracking using a robust gaussian filter. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 608–615. IEEE, 2016. 2

[50] Yuhwan Jeong, Hoonhee Cho, and Kuk-Jin Yoon. Towards robust event-based networks for nighttime via unpaired day-to-night event translation. In *European Conference on Computer Vision*, pages 286–306. Springer, 2024. 3

[51] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *European conference on computer vision*, pages 557–572. Springer, 2020. 4

[52] Jae-Young Kang, Hoonhee Cho, and Kuk-Jin Yoon. Temporal stereo matching from event cameras via joint learning with stereoscopic flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[53] Jae-Young Kang, Hoonhee Cho, and Kuk-Jin Yoon. Unleashing the temporal potential of stereo event cameras for continuous-time 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6869–6881, 2025. 3

[54] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–10, 2017. 1

[55] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J Davison. Simultaneous mosaicing and tracking with an event camera. *J. Solid State Circ*, 43:566–576, 2008. 2

[56] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European conference on computer vision*, pages 349–364. Springer, 2016. 2, 3

[57] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European conference on computer vision*, pages 574–591. Springer, 2020. 2, 4

[58] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022. 1, 2, 3, 4, 6, 7

[59] Taeyeop Lee, Byeong-Uk Lee, Myungchul Kim, and In So Kweon. Category-level metric scale object shape and pose estimation. *IEEE Robotics and Automation Letters*, 6(4): 8575–8582, 2021. 1, 2

[60] Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon. Tta-cope: Test-time adaptation for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21285–21295, 2023. 2

[61] Taeyeop Lee, Gyuree Kang, Bowen Wen, Youngho Kim, Seunghyeok Back, In So Kweon, David Hyunchul Shim, and Kuk-Jin Yoon. Delta: Demonstration and language-guided novel transparent object manipulation. *arXiv preprint arXiv:2510.05662*, 2025. 1

[62] Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6d: Model-free 6d pose estimation of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11633–11643, 2025. 2

[63] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International Journal of Computer Vision*, 81(2):155–166, 2009. 6

[64] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 683–698, 2018. 2, 4

[65] Zhichao Li, Arren Glover, Chiara Bartolozzi, and Lorenzo Natale. 6-dof object tracking with event-based optical flow and frames. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 18880–18887. IEEE, 2025. 3

[66] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024. 1

[67] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Keypoint-based category-level object pose tracking from an rgb sequence with uncertainty estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1258–1264. IEEE, 2022. 2

[68] Daqi Liu, Alvaro Parra, and Tat-Jun Chin. Spatiotemporal registration for event-based visual odometry. In *Proceed-*

ings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4937–4946, 2021. 3

[69] Xu Liu, Xiaopeng Fan, Jianing Li, Dianze Li, Wei Zhang, Zhengyu Ma, and Yonghong Tian. High-rate monocular depth estimation via cross frame-rate collaboration of frames and events. *International Journal of Computer Vision*, 133(10):7332–7351, 2025. 3

[70] Zhe Liu, Dianxi Shi, Ruihao Li, Yi Zhang, and Shaowu Yang. T-esvo: improved event-based stereo visual odometry via adaptive time-surface and truncated signed distance function. *Advanced Intelligent Systems*, 5(9):2300027, 2023. 3

[71] Zibin Liu, Banglei Guan, Yang Shang, Shunkun Liang, Zhenbao Yu, and Qifeng Yu. Optical flow-guided 6dof object pose tracking with an event camera. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6501–6509, 2024. 3

[72] Zibin Liu, Banglei Guan, Yang Shang, Qifeng Yu, and Laurent Kneip. Line-based 6-dof object pose estimation and tracking with an event camera. *IEEE Transactions on Image Processing*, 33:4765–4780, 2024. 3

[73] Zibin Liu, Banglei Guan, Yang Shang, Yifei Bian, Pengju Sun, and Qifeng Yu. Stereo event-based, 6-dof pose tracking for uncooperative spacecraft. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025. 3

[74] Adam Loch, Germain Haessig, and Markus Vincze. Event-based high-speed low-latency fiducial marker tracking. In *Fifteenth International Conference on Machine Vision (ICMV 2022)*, pages 323–330. SPIE, 2023. 3

[75] Xiuyuan Lu, Yi Zhou, Junkai Niu, Sheng Zhong, and Shaojie Shen. Event-based visual inertial velometer. *arXiv preprint arXiv:2311.18189*, 2023. 3

[76] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 2

[77] Xinglong Luo, Ao Luo, Zhengning Wang, Chunyu Lin, Bing Zeng, and Shuaicheng Liu. Efficient meshflow and optical flow estimation from event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19198–19207, 2024. 3

[78] Yongrui Ma, Shi Guo, Yutian Chen, Tianfan Xue, and Jinwei Gu. Timelens-xl: Real-time event-based video frame interpolation with large motion. In *European Conference on Computer Vision*, pages 178–194. Springer, 2024. 3

[79] Mateusz Majcher and Bogdan Kwolek. 3d model-based 6d object pose tracking on rgb images using particle filtering and heuristic optimization. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 690—-697, 2020. 2

[80] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2024. 2

[81] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Co-op: Correspondence-based novel object pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11622–11632, 2025. 1, 2, 5

[82] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018. 2

[83] Nguyen Van Nguyen, Tyree Stephen, Guo Andrew, Fourmy Mederic, Gouda Anas, Lee Taeyeop, Moon Sungphill, Son Hyeontae, Ranftl Lukas, Tremblay Jonathan, et al. Bop challenge 2024 on model-based and model-free 6d object pose estimation. *arXiv preprint arXiv:2504.02812*, 2025. 1, 6

[84] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9903–9913, 2024. 1, 2

[85] Van Nguyen Nguyen, Christian Forster, Sindi Shkodrani, Vincent Lepetit, Bugra Tekin, Cem Keskin, and Tomas Hodan. Gotrack: Generic 6dof object pose refinement and tracking. *arXiv preprint arXiv:2506.07155*, 2025. 1, 2, 4

[86] Mustafa Özuysal, Vincent Lepetit, François Fleuret, and Pascal Fua. Feature harvesting for tracking-by-detection. In *European conference on computer vision*, pages 592–605. Springer, 2006. 2

[87] Federico Paredes-Vallés and Guido CHE De Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 3

[88] Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 143–152, 2017. 2

[89] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4561–4570, 2019. 2

[90] Stefano Pini, Guido Borghi, and Roberto Vezzani. Learn to see by events: Color frame synthesis from event and rgb cameras. *arXiv preprint arXiv:1812.02041*, 2018. 3

[91] Georgy Ponimatkin, Martin Cífka, Tomáš Souček, Médéric Fourmy, Yann Labbé, Vladimir Petrik, and Josef Sivic. 6D Object Pose Tracking in Internet Videos for Robotic Manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[92] 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. 3

[93] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pages 969–982. PMLR, 2018. 5

[94] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer

vision to event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3857–3866, 2019. 2, 6, 7

[95] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2019. 2, 6, 7, 8

[96] Pavel Rojtberg and Thomas Pöllabauer. Ycb-ev 1.1: Event-vision dataset for 6dof object pose estimation. In *European Conference on Computer Vision*, pages 1–13. Springer, 2024. 1, 2, 3, 5

[97] Edward Rosten and Tom Drummond. Fusing points and lines for high performance tracking. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 1508–1515. Ieee, 2005. 2

[98] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[99] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pages 308–324. Springer, 2018. 2

[100] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 156–163, 2020. 2

[101] Byung-Kuk Seo and Harald Wuest. A direct method for robust model-based 3d object tracking from a monocular rgb image. In *European conference on computer vision*, pages 551–562. Springer, 2016. 2

[102] Byung-Kuk Seo, Hanhoon Park, Jong-Il Park, Stefan Hinterstoisser, and Slobodan Ilic. Optimal local searching for fast and robust textureless 3d object tracking in highly cluttered backgrounds. *IEEE transactions on visualization and computer graphics*, 20(1):99–110, 2013. 2

[103] Yichen Shen, Yijin Li, Shuo Chen, Guanglin Li, Zhaoyang Huang, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Blinktrack: Feature tracking over 80 fps via events and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9308, 2025. 3

[104] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *European Conference on Computer Vision*, pages 628–645. Springer, 2022. 3

[105] Iryna Skrypnyk and David G Lowe. Scene modelling, recognition and tracking with invariant image features. In *Third IEEE and ACM international symposium on mixed and augmented reality*, pages 110–119. IEEE, 2004. 2

[106] Manuel Stoiber, Martin Pfanne, Klaus H Strobl, Rudolph Triebel, and Alin Albu-Schäffer. Srt3d: A sparse region-based 3d object tracking approach for the real world. *International Journal of Computer Vision*, 130(4):1008–1030, 2022. 2

[107] Manuel Stoiber, Martin Sundermeyer, and Rudolph Triebel. Iterative corresponding geometry: Fusing region and depth for highly efficient 3d tracking of textureless objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6855–6865, 2022. 2

[108] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 4

[109] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 3

[110] Meng Tian, Marcelo H Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision*, pages 530–546. Springer, 2020. 1

[111] Xuhui Tian, Xinran Lin, Fan Zhong, and Xueying Qin. Large-displacement 3d object tracking with hybrid non-local optimization. In *European Conference on Computer Vision*, pages 627–643. Springer, 2022. 2

[112] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021. 3

[113] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 3

[114] Luca Vacchetti, Vincent Lepetit, and Pascal Fua. Stable real-time 3d tracking using online and offline information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1385–1391, 2004. 2

[115] 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. 3

[116] Zhexiong Wan, Jianqin Luo, Yuchao Dai, and Gim Hee Lee. Event-aided dense and continuous point tracking: Everywhere and anytime. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7936–7946, 2025. 3

[117] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 1, 2

[118] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In *2020 IEEE International Conference on*

Robotics and Automation (ICRA), pages 10059–10066. IEEE, 2020. 2

[119] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2642–2651, 2019. 1, 2

[120] Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. Ho-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. *arXiv preprint arXiv:2406.06843*, 2024. 1

[121] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019. 3

[122] Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2135–2145, 2021. 2, 3

[123] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. Joint framework for single image reconstruction and super-resolution with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7657–7673, 2021. 3

[124] Long Wang, Shen Yan, Jianan Zhen, Yu Liu, Maojun Zhang, Guofeng Zhang, and Xiaowei Zhou. Deep active contours for real-time 6-dof object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14034–14044, 2023. 2

[125] Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074. IEEE, 2021. 2, 3

[126] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E Bekris. se (3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10367–10373. IEEE, 2020. 2, 4

[127] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023. 2, 3

[128] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17868—-17879, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[129] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5249–5260, 2025. 4

[130] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 2

[131] Manuel Wuthrich, Peter Pastor, Mrinal Kalakrishnan, Jeannette Bohg, and Stefan Schaal. Probabilistic object tracking using a range camera. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3195—-3202, 2013. 1, 2

[132] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1, 2, 5, 6

[133] Zeyu Xiao and Xinchao Wang. Event-based video super-resolution via state space models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12564–12574, 2025. 2

[134] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019. 1

[135] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019. 2

[136] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[137] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022. 3

[138] Yi Zhou, Guillermo Gallego, and Shaojie Shen. Event-based stereo visual odometry. *IEEE Transactions on Robotics*, 37(5):1433–1450, 2021. 3

[139] Juanjuan Zhu, Zhexiong Wan, and Yuchao Dai. Video frame prediction from a single image and events. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, number 7, pages 7748–7756, 2024. 3

[140] Jinjing Zhu, Tianbo Pan, Zidong Cao, Yexin Liu, James T Kwok, and Hui Xiong. Depth any event stream: Enhancing event-based monocular depth estimation via dense-to-sparse distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5146–5155, 2025. 3