

Received 24 August 2026, accepted 13 September 2026, date of publication 18 September 2026, date of current version 1 October 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3735008

RESEARCH ARTICLE

EventEgoHands++: Event-Based Egocentric 3D Hand Mesh Reconstruction With Real Dataset

RYOSEI HARA¹, (Graduate Student Member, IEEE),
WATARU IKEDA¹, (Graduate Student Member, IEEE),
MASASHI HATANO¹, (Graduate Student Member, IEEE),
AND MARIKO ISOGAWA^{1,2}, (Member, IEEE)

¹Graduate School of Science and Technology, Keio University, Yokohama, Kanagawa 223-8522, Japan

²JST PRESTO, Tokyo 102-0076, Japan

Corresponding author: Ryosei Hara (ryosei_hara@keio.jp)

This work was supported in part by Japan Science and Technology Agency (JST) Precursory Research for Embryonic Science and Technology (PRESTO) under Grant JPMJPR22C1, in part by Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 24K22296 and Grant 25H01159, and in part by the JST Fusion Oriented Research for Disruptive Science and Technology (FOREST) Program under Grant JPMJFR242I. The work of Wataru Ikeda was supported by JST Broadening Opportunities for Outstanding Young Researchers and Doctoral Students in Strategic Areas (BOOST), Japan, under Grant JPMJBS2409. The work of Masashi Hatano was supported in part by JST BOOST, Japan, under Grant JPMJBS2409, and in part by the Amano Institute of Technology.

ABSTRACT 3D hand mesh reconstruction is a challenging yet essential task for downstream applications, including human-robot interaction and AR/VR. Although conventional cameras (*e.g.*, RGB or depth cameras) have been widely adopted for this task, methods that rely on them struggle in low-light environments and under severe motion blur. To address these limitations, event-based cameras have recently attracted attention for their high dynamic range and high temporal resolution. However, applying event cameras to egocentric hand reconstruction remains challenging because camera wearer's motion produces dense background events that obscure hand-specific signals. Although the first egocentric event-based approach mitigates this issue using hand segmentation, its binary hand mask does not distinguish between left and right hands. As a result, the model lacks instance-level hand information and predicts both hands even when only one or neither hand is present. This limitation leads to incorrect inter-hand relationships and degraded reconstruction accuracy. In this paper, we propose EventEgoHands++, a framework for event-based 3D hand mesh reconstruction from an egocentric viewpoint that overcomes these limitations. The proposed method incorporates a Hand Detector that estimates instance-level bounding boxes and masks for both the left and right hands. Moreover, we introduce Adaptive Attention, which dynamically gates the attention based on these detection results to accurately learn the spatial relationship and mutual interactions between the hands. To train and evaluate our framework, we extend the synthetic N-HOT3D dataset with bounding-box annotations and refined masks, and newly construct EEH-R, the largest real-world event-based egocentric hand dataset to date, comprising approximately 1M annotated frames captured in environments including low-light conditions. Extensive experiments on both synthetic and real datasets demonstrate that our method consistently outperforms the baselines. Our code and datasets are available at <https://ryhara.github.io/EventEgoHandsV2/>

INDEX TERMS Egocentric vision, event-based vision, 3D hand mesh reconstruction, 3D hand pose estimation.

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen¹.

I. INTRODUCTION

Hands play a fundamental role in human interaction with the physical world. For example, daily activities such as grasping objects, manipulating tools, and performing gestures rely heavily on hand motions. Therefore, reconstructing a 3D

hand mesh from egocentric vision is important for applications such as human–computer interaction, AR/VR, and robotics. In particular, capturing fine-grained hand motions and shapes from the camera wearer’s viewpoint is essential for enabling immersive experiences and safe interactions.

However, conventional methods based on RGB(D) cameras [1], [2], [3], [4], [5], [6] may encounter challenging situations when applied to egocentric vision. For example, in addition to rapid hand movements, the head motion of the camera wearer can cause motion blur. Moreover, in environments with varying illumination conditions, particularly in low-light settings, recognizing the hand can become challenging.

Recently, the use of event-based cameras, hereafter referred to as event cameras, has attracted increasing attention [7]. Event cameras provide high temporal resolution and a high dynamic range, enabling the capture of fast motions even in low-light environments where RGB cameras struggle. They also offer low power consumption and memory-efficient sensing, making them well suited for deployment on edge devices and wearable platforms.

However, most existing event-based 3D hand mesh reconstruction methods [8], [9], [10], [11] focus on fixed third-person camera setups. In contrast, an egocentric perspective provides greater flexibility and mobility for capturing natural hand interactions, as illustrated in Fig. 1. One of the main challenges in using an event camera in a first-person setting is that camera motion generates a large number of events across the entire background, making it difficult to extract events corresponding to hand motion. In our earlier conference paper, we proposed EventEgoHands [12], the first framework for egocentric event-based hand reconstruction, which suppresses such background events with a hand segmentation module and established the feasibility of this task.

Nevertheless, EventEgoHands should be regarded as a first step, as it leaves three fundamental problems unresolved.

Lack of instance-level hand identity. The binary mask in [12] does not encode left/right hand identity, so the model always reconstructs both hands regardless of their actual visibility, which frequently breaks in egocentric video where hands often leave the field of view. This yields invalid outputs when only one or neither hand is present and degrades the estimated inter-hand relative positions.

Visibility agnostic interaction modeling. Its cross-attention is applied unconditionally between the two hand branches, forcing feature exchange with absent or misidentified hands. Such a fixed attention pathway cannot adapt to the constantly changing hand visibility inherent to egocentric interaction.

Unverified real-world applicability. Its validation was confined to synthetic data, although real event streams differ substantially from simulated ones in event density, temporal distribution, and sensor noise. Since no real-world egocentric event dataset existed, the effectiveness of this task

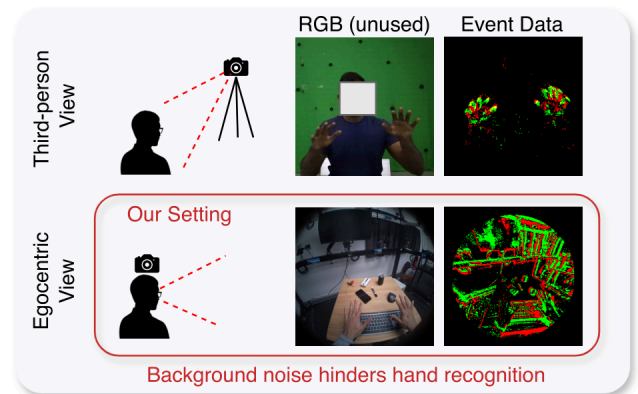


FIGURE 1. Challenges of egocentric event-based cameras. While fixed third-person camera setups are limited to specific environments, egocentric viewpoints provide greater mobility and flexibility for capturing hand interactions. However, in egocentric event camera scenarios, the wearer’s motion generates numerous background events, which can obscure hand-related signals and make accurate hand recognition difficult.

on real sensor data remained unverified. In particular, performance under low-light conditions, the very scenario that motivates event cameras, had never been evaluated on real data.

To address these problems, we propose **EventEgoHands++**, a robust framework for event-based egocentric 3D hand mesh reconstruction. Our proposed method consists of two main stages: hand extraction and hand reconstruction. First, to resolve the inability to distinguish between left and right hands, we incorporate a Hand Detector in the hand extraction stage. This detector simultaneously estimates instance-level bounding boxes and masks, ensuring that each hand is accurately localized and identified only when present. Second, to address the performance degradation caused by fixed attention mechanisms, we introduce Adaptive Attention in the hand reconstruction stage. Unlike existing methods that apply cross-attention regardless of visibility, our Adaptive Attention dynamically selects the attention operations according to the detection results. This allows the model to effectively learn spatial correlations and inter-hand interactions by adapting to the visibility of each hand.

In addition, to facilitate research in this field, we provide two datasets. Specifically, we extend **N-HOT3D**, the synthetic dataset introduced in [12], with refined segmentation masks and newly added bounding-box annotations, enlarging it to approximately 480K annotated frames. More importantly, we introduce **EEH-R**, a newly collected real-world dataset captured with an actual event camera in challenging egocentric scenarios, including low-light conditions. While event-based hand reconstruction on real event data has been explored only in third-person settings, EEH-R enables, for the first time, evaluation on real event data in egocentric settings. To the best of our knowledge, EEH-R is the largest real-world event-based egocentric hand dataset, containing

TABLE 1. Comparison of event-based hand datasets. The asterisk indicates that the corresponding annotation is available for a subset of the frames.

Type	Dataset	View	Hands	Subs.	Dur.	Seq.	#GT	Annotation
Synthetic	EventHands (Synthetic) [8]	Third	Single	-	100 h	-	360M	MANO
	Ev2Hands-S [10]	Third	Both	-	-	-	28K	MANO, Event seg.
	N-HOT3D (Ours)	Ego	Both	9	4.4 h	136	480K	MANO, Hand mask/bbox
Real	EventHands (Real) [8]	Third	Single	-	12.6 s	4	357	2D keypoints
	Ev2Hands-R [10]	Third	Both	5	20.1 m	8	70K	3D keypoints
	EvRealHands [9]	Third	Single	10	79 m	102	425K	MANO
	EEH-R (Ours)	Ego	Both	8	2.36 h	85	1M	MANO, (Hand mask/bbox)*

approximately 1M annotated frames. The details of the two datasets, N-HOT3D and EEH-R, are summarized in Table 1.

Using both datasets, we validate the effectiveness of EventEgoHands++. On the synthetic N-HOT3D dataset, our method reduces mean MPJPE by 21.82 mm (33.7%) and MPVPE by 20.57 mm (34.0%) compared with the best existing method. On the EEH-R dataset, it further reduces MPJPE by 7.90 mm (18.8%) and MPVPE by 7.13 mm (18.1%).

Our technical contributions are summarized as follows:

- We propose **EventEgoHands++**, a robust event-based egocentric 3D hand mesh reconstruction framework that explicitly handles hand visibility and left/right hand identity under severe egocentric camera motion.
- We introduce two key methodological components: a **Hand Detector** that jointly estimates instance-level bounding boxes and masks for the left and right hands, and **Adaptive Attention**, which dynamically activates inter-hand attention based on detection results to model spatial relationships only when the corresponding hands are visible.
- We newly construct **EEH-R**, the first and largest real-world event-based egocentric hand dataset, containing approximately 1M annotated frames captured under both well-lit and low-light conditions. We further extend the synthetic **N-HOT3D** dataset with refined masks and newly added bounding-box annotations, enlarging it to approximately 480K annotated frames. Experiments on these datasets demonstrate the effectiveness of EventEgoHands++ against existing baselines.

The paper is structured as follows. Section II reviews related work on egocentric hand analysis, 3D hand pose estimation, and event-based 3D hand mesh reconstruction. Section III describes the proposed EventEgoHands++ framework, including the hand extraction and hand reconstruction stages. Section IV introduces the synthetic N-HOT3D dataset and the real-world EEH-R dataset. Section V presents the experimental setup, including implementation details, baseline methods, and evaluation metrics. Section VI reports the experimental results, including evaluations on both datasets, hand segmentation performance, ablation studies, and failure case analysis. Finally, Section VII concludes the paper.

II. RELATED WORK

A. HANDS IN EGOCENTRIC VIDEOS

Hands are the primary interface through which humans interact with the physical world, making hand analysis central to egocentric vision [13], [14]. Compared with exocentric views, egocentric videos present unique challenges for hand analysis, including rapid hand motion, frequent occlusions, and large appearance changes caused by the camera wearer's motion.

Early studies in egocentric hand analysis focused on hand localization and segmentation [15], [16], [17], [18]. Bambach et al. [19] introduced the EgoHands dataset, establishing a benchmark for first-person hand detection. More recent egocentric datasets have extended this setting toward 3D hand pose, shape, and hand-object interaction understanding, including FPHA [20], H2O [21], ARCTIC [22], and HOT3D [23]. These datasets show that hands are a key target in egocentric vision and that accurate 3D hand estimation is essential for understanding natural interactions from a wearable viewpoint. This growing interest in egocentric hand understanding has motivated extensive research on 3D hand pose estimation and mesh reconstruction. The following subsections review general methods for 3D hand pose estimation and mesh reconstruction.

B. 3D HAND POSE ESTIMATION

Initial efforts in 3D hand pose estimation primarily relied on depth sensors [1], [2], [3] to capture the complex geometry of the hand. With the shift toward monocular RGB input, the field has largely converged on the use of parametric models to provide anatomical priors. The MANO model [24] has become the de facto standard in this field, offering a low-dimensional yet anatomically plausible representation of hand shape and pose. Boukhayma et al. [4] pioneered the first fully learnable pipeline to directly regress MANO parameters from a single image, followed by the use of intermediate representations like 2D heatmaps [25], iterative 2D alignment loops [26], and occlusion-robust methods [27].

On the other hand, non-parametric methods [28], [29], [30] have explored direct vertex regression using Graph Convolutional Networks (GCNs) [31], which have been widely used to exploit the inherent graph structure of the hand mesh. THOR-Net [32] further extended these to hand-object interaction, in which two hands and object poses are

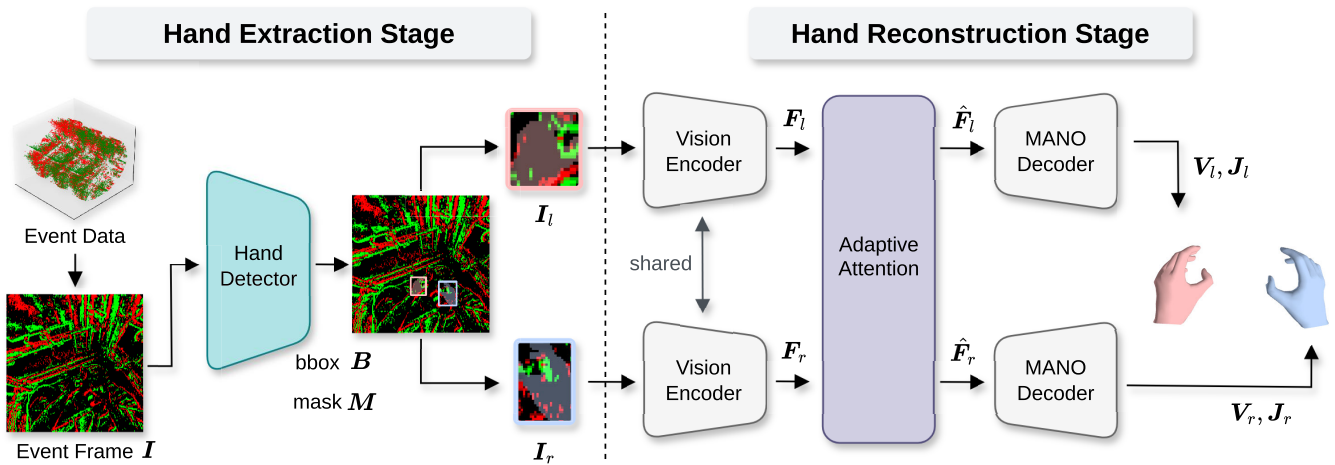


FIGURE 2. Architecture of EventEgoHands++. The pipeline consists of two primary stages: (1) Hand Extraction Stage, where a hand detector identifies and extracts hand regions from event data, and (2) Hand Reconstruction Stage, which utilizes a vision encoder for feature extraction and estimates 3D hand poses through Adaptive Attention based on the detection results.

estimated. More recently, Transformer-based architectures, such as METRO [33] and Mesh Graphormer [34], have set new state-of-the-art benchmarks by modeling global interactions between vertices and joints without relying on a fixed kinematic tree.

However, the parametric approach remains highly favored for its robustness to occlusions and its ability to maintain valid hand topology. Recently, Pavlakos et al [5] demonstrated that the performance of parametric reconstruction can be significantly enhanced by scaling up model capacity based on Vision Transformer (ViT) [35] backbones, following the success in body pose estimation [36], [37]. Recent image-based 3D hand pose estimation methods [5], [6] achieve state-of-the-art accuracy across diverse datasets. Several works [38], [39], [40] have extended their efforts to 4D hand mesh reconstruction that incorporates temporal information.

While monocular 3D hand mesh reconstruction has advanced significantly, conventional camera-based approaches often fail in low-light conditions and under motion blur caused by rapid hand or camera movement. To this end, this study explores the use of event cameras, which are robust in such extreme scenarios due to their high dynamic range and high temporal resolution.

C. EVENT-BASED 3D HAND MESH RECONSTRUCTION

Event cameras output asynchronous streams of events by recording changes in pixel intensity. In recent years, event cameras have gained increasing popularity [7], [41], [42], [43], [44] owing to their high dynamic range and high temporal resolution, which help address challenges such as low-light conditions and motion blur that conventional sensors, such as RGB or depth cameras, often fail to handle effectively.

In the context of 3D vision, event cameras are particularly advantageous as they provide nearly continuous trajectories

of moving edges, allowing for the recovery of precise geometric structures and motion cues that are typically lost during fast movements in frame-based captures. Recently, event cameras have been utilized for several 3D tasks [45], such as 3D geometric reconstruction [46], [47], [48], [49], non-rigid reconstruction [50], depth estimation [51], [52], human pose estimation [53], [54], [55], [56], and 3D hand tracking [57].

3D hand mesh reconstruction is a central task for various applications, such as robotics or AR/VR. Some studies [58], [59] have utilized both RGB and event data to complement modality-specific information. Although these methods are robust, the hardware setup for inference is relatively expensive and unrealistic for real-world applications.

Several studies [8], [9], [10], [11] tackle the 3D hand mesh reconstruction task using only event cameras during inference. While reconstruction methods in other domains often focus on task-specific network architectures [41], [46], event-based hand reconstruction has mainly emphasized event representations suitable for subsequent 3D hand estimation. EventHands [8] is a lightweight framework designed for fast hand motion reconstruction. It employs a frame-based 2D representation, termed Locally-Normalized Event Surfaces (LNES), which preserves relative temporal information and polarity within a short period of time. Ev2Hands [10] tackles the reconstruction of both hands via a point cloud-based approach. It preserves temporal information via a point cloud representation, termed Event Cloud, effectively leveraging the raw event data. EvHandPose [9] likewise adopts the LNES representation. In addition, it introduces motion representations based on shape flow and edges to effectively reduce motion ambiguity, and addresses the challenge of sparse annotations using a weakly supervised learning framework. Recently, RPEP [11] was proposed as a pre-training method that leverages labeled RGB images and

unlabeled event data during training to improve 3D hand pose estimation. It uses an iterative module to generate realistic pseudo-events from static images, effectively capturing non-rigid hand articulations and reducing the need for scarce event-based annotations.

These event-based 3D hand mesh reconstruction methods are restricted to fixed third-person camera views. While there are a lot of works on 3D hand pose estimation using conventional cameras from an egocentric viewpoint [60], [61], [62], [63], [64], capturing from egocentric perception using event cameras is relatively underexplored. Although an egocentric viewpoint offers greater flexibility and mobility than fixed third-person setups, it introduces a key challenge unique to event cameras: the wearer's motion changes the brightness across the entire scene and generates numerous background events that obscure hand-related signals. EventEgoHands [12] pioneered an egocentric approach by estimating coarse hand regions and filtering events within them to mitigate egocentric background noise. However, it cannot distinguish between left and right hands and consistently predicts both regardless of their actual presence, which limits its practical use. Moreover, its reconstruction module indiscriminately applies cross-attention, hindering the effective learning of correlations between hands. To address these issues, we propose **EventEgoHands++**, which introduces instance-level detection and Adaptive Attention for more robust and flexible reconstruction.

III. PROPOSED METHOD

We propose EventEgoHands++, a 3D hand mesh reconstruction method that relies solely on event data captured in dynamic egocentric scenes. As illustrated in Fig. 2, given an event frame $\mathbf{I} \in \mathbb{R}^{2 \times H \times W}$, EventEgoHands++ reconstructs 3D joint positions $\mathbf{J}_l, \mathbf{J}_r \in \mathbb{R}^{20 \times 3}$ and mesh vertex positions $\mathbf{V}_l, \mathbf{V}_r \in \mathbb{R}^{778 \times 3}$ of both hands. Our approach comprises two main stages: 1) the Hand Extraction Stage, which employs a Hand Detector to estimate bounding boxes and masks for the left and right hands, thereby extracting events occurring within the hand regions, and 2) the Hand Reconstruction Stage, which reconstructs the 3D hand mesh from the extracted hand events by applying Adaptive Attention. These are described in Section III-A and Section III-B, respectively, followed by the description of the loss function in Section III-C. The overall workflow of EventEgoHands++ during training is summarized in Algorithm 1.

A. HAND EXTRACTION STAGE

The Hand Detector takes an event frame $\mathbf{I} \in \mathbb{R}^{2 \times H \times W}$ as input and predicts instance-level hand bounding boxes $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^N$ with $\mathbf{b}_i = (x_i, y_i, w_i, h_i) \in \mathbb{R}^4$, where (x_i, y_i) denotes the center and (w_i, h_i) the width and height of the bounding box. All coordinates are normalized by the image width and height. The module also predicts a hand region mask $\mathbf{M} \in \mathbb{R}^{N \times H \times W}$, where N denotes the number of hand instances.

Algorithm 1 Training Procedure of EventEgoHands++

Require: Training set $\mathcal{D} = \{(\mathbf{I}, \mathbf{J}^{\text{gt}}, \mathbf{V}^{\text{gt}}, \boldsymbol{\theta}^{\text{gt}}, \boldsymbol{\beta}^{\text{gt}})\}$, pre-trained Hand Detector $\mathcal{D}_{\text{hand}}$, Hand Reconstructor $\mathcal{H}_{\phi}$, MANO model $\mathcal{M}$, loss weights $\lambda_{\gamma}, \lambda_{\delta}, \lambda_{\epsilon}, \lambda_{\zeta}$, learning rate η , number of epochs E

Ensure: Optimized parameters ϕ of $\mathcal{H}_{\phi}$

```

1: for  $e = 1, \dots, E$  do
2:   for each minibatch in  $\mathcal{D}$  do
3:     // Hand Extraction Stage
4:      $\mathbf{B}, \mathbf{M}, \mathbf{c}, \mathbf{s} \leftarrow \mathcal{D}_{\text{hand}}(\mathbf{I})$  { $\mathbf{c}$ : hand-side labels (l/r),  $\mathbf{s}$ : confidence scores}
5:     Select  $\mathbf{M}_l$  and  $\mathbf{M}_r$  using hand-side labels  $\mathbf{c}$  and confidence scores  $\mathbf{s}$ 
6:     Determine validity indicators  $m_l, m_r \in \{0, 1\}$ 
7:      $\mathcal{H}_{\text{vis}} \leftarrow \{h \in \{l, r\} \mid m_h = 1\}$ 
8:     // Hand Reconstruction Stage
9:     if  $\mathcal{H}_{\text{vis}} = \emptyset$  then
10:      {Condition 3: no hand detected}
11:      skip this sample
12:     end if
13:     for  $h \in \mathcal{H}_{\text{vis}}$  do
14:        $\mathbf{I}_h \leftarrow \mathbf{I} \odot \mathbf{M}_h$ 
15:        $\mathbf{F}_h \leftarrow \text{VisionEncoder}(\mathbf{I}_h)$ 
16:     end for
17:     // Adaptive Attention
18:     if  $m_l = 1$  and  $m_r = 1$  then
19:       {Condition 1: both hands detected}
20:        $\bar{\mathbf{F}}_h \leftarrow \text{SelfAttn}_h(\mathbf{F}_h), \quad h \in \{l, r\}$ 
21:        $\hat{\mathbf{F}}_l \leftarrow \text{CrossAttn}(\bar{\mathbf{F}}_l, \bar{\mathbf{F}}_r)$ 
22:        $\hat{\mathbf{F}}_r \leftarrow \text{CrossAttn}(\bar{\mathbf{F}}_r, \bar{\mathbf{F}}_l)$ 
23:     else
24:       {Condition 2: single hand detected}
25:        $\hat{\mathbf{F}}_h \leftarrow \text{SelfAttn}_h(\mathbf{F}_h), \quad h \in \mathcal{H}_{\text{vis}}$ 
26:     end if
27:     // MANO Decoder
28:     for  $h \in \mathcal{H}_{\text{vis}}$  do
29:        $(\boldsymbol{\theta}_h, \boldsymbol{\beta}_h, \mathbf{t}_h, \mathbf{R}_h) \leftarrow \text{Decoder}(\hat{\mathbf{F}}_h)$ 
30:        $\hat{\mathbf{J}}_h, \hat{\mathbf{V}}_h \leftarrow \mathcal{M}(\boldsymbol{\theta}_h, \boldsymbol{\beta}_h, \mathbf{t}_h, \mathbf{R}_h)$ 
31:     end for
32:     // Training Objective
33:      $\mathcal{L}_{\text{hand}} \leftarrow \lambda_{\gamma} \mathcal{L}_{\text{joints}} + \lambda_{\delta} \mathcal{L}_{\text{interhand}} + \lambda_{\epsilon} \mathcal{L}_{\text{vertices}} + \lambda_{\zeta} \mathcal{L}_{\text{MANO}}$ 
34:     Update  $\phi$  using AdamW with learning rate  $\eta$ 
35:   end for
36: end for
37: return  $\phi$ 

```

We adopt an instance segmentation framework rather than a mask-only estimation approach [12] because mask-only prediction tends to be unstable. By jointly learning bounding boxes and masks, the estimation becomes more robust. The predicted mask $\mathbf{M}$ is used to extract only the events within the hand region from the event frame $\mathbf{I}$, yielding separate masked event frames for each hand: $\mathbf{I}_l, \mathbf{I}_r \in \mathbb{R}^{2 \times H \times W}$.

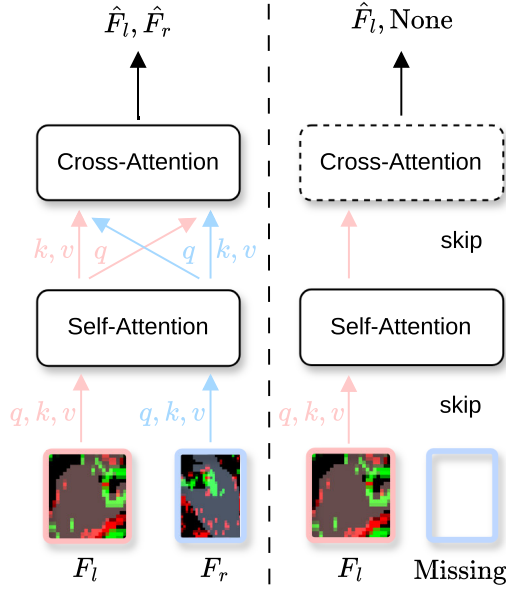


FIGURE 3. Adaptive Attention. When both hands are detected, self-attention first refines intra-hand features, followed by cross-attention to capture inter-hand dependencies. When only one hand is detected, cross-attention is skipped and only self-attention is applied.

We apply LNES [8], one of the event-frame representations, to generate I from raw events, as LNES is known to preserve temporal information through the use of temporal weighting. We adopt YOLO26 [65], the latest model in the widely used YOLO family for hand detection [6], [66], as our hand detector to jointly estimate bounding boxes and segmentation masks.

B. HAND RECONSTRUCTION STAGE

The hand reconstruction stage takes filtered event frames $I_l, I_r \in \mathbb{R}^{2 \times H \times W}$ from the hand extraction stage and estimates the 3D joint positions $J_l, J_r \in \mathbb{R}^{20 \times 3}$ and mesh vertex positions $V_l, V_r \in \mathbb{R}^{778 \times 3}$ for each hand.

1) FEATURE EXTRACTION

Given the masked event frames for the left and right hands, a shared backbone extracts spatial feature maps F_l and F_r , each of size $C \times H' \times W'$, where C denotes the channel dimension and H', W' denote the spatial resolution of the feature map. As the vision encoder, we employ EfficientNetV2-S [67], which has demonstrated strong performance in various hand pose estimation architectures [61], [63], [68].

2) ADAPTIVE ATTENTION

To dynamically handle the presence or absence of each hand, we introduce Adaptive Attention, which selects attention operations based on the detection results. Let $m_l, m_r \in \{0, 1\}$ denote validity indicators for the left and right hands, obtained from the hand extraction stage, where $m_h = 1$ indicates that the corresponding hand is detected.

(e.g., if $h = l$ and $m_l = 1$, then it indicates that the left hand is detected). We describe the behavior of Adaptive Attention under three hand-detection conditions: (1) both hands are detected, (2) a single hand is detected, and (3) no hands are detected.

Condition 1: Both hands detected ($m_l=1, m_r=1$). Each feature map is first refined independently by self-attention to model intra-hand spatial relationships, and then the two feature maps interact through bidirectional cross-attention to capture inter-hand dependencies:

$$\bar{F}_h = \text{SelfAttn}_h(F_h), \quad h \in \{l, r\}, \quad (1)$$

$$\hat{F}_l = \text{CrossAttn}(\bar{F}_l, \bar{F}_r), \quad (2)$$

$$\hat{F}_r = \text{CrossAttn}(\bar{F}_r, \bar{F}_l). \quad (3)$$

The self-attention layer (1) first refines each hand's feature map by allowing each spatial position to attend to all other positions within the same hand feature map. In the subsequent cross-attention layers (2) and (3), each hand's feature map serves as the query while the other hand's feature map provides the keys and values, enabling bidirectional information exchange. This allows the model to capture inter-hand context such as relative hand positioning and coordinated finger configurations, building upon already-refined single-hand representations.

Condition 2: Single hand detected ($m_l+m_r=1$). Cross-attention is disabled since the other hand's feature map is unavailable, and only self-attention is applied to the detected hand:

$$\hat{F}_h = \text{SelfAttn}_h(F_h). \quad (4)$$

Condition 3: Neither hand detected ($m_l=0, m_r=0$). If neither hand is detected, the subsequent processing is skipped.

Unlike masked attention, which applies masks to attention scores inside the attention operation, our Adaptive Attention switches the attention operations themselves. With masked attention, masking all tokens of an undetected hand corrupts the cross-attention output, which in turn destroys the features of the detected hand, while the masked attention computation itself is still executed. In contrast, our method skips the processing of the undetected hand entirely and processes the detected hand independently via self-attention, thereby avoiding both feature corruption and unnecessary computational overhead. We note that the self-attention and cross-attention operations themselves are standard, and the novelty of Adaptive Attention lies in using per-hand visibility as an explicit control signal that determines which operations to apply.

3) MANO DECODER

Finally, each refined spatial feature map is converted into a feature vector via attention pooling, which is then mapped by a linear layer to the MANO [24] parameter space. These parameters are decoded by the MANO model to obtain the 3D hand joint positions J and mesh vertices V . MANO

parameters for each hand consist of a pose vector $\theta \in \mathbb{R}^{45}$, a shape vector $\beta \in \mathbb{R}^{10}$, a translation $t \in \mathbb{R}^3$, and a rotation $R \in \mathbb{R}^3$. Using the MANO model $\mathcal{M}$, we obtain sparse hand joints and dense hand mesh vertices for each hand individually as $J, V = \mathcal{M}(\theta, \beta, t, R)$. The resulting joint locations $J \in \mathbb{R}^{20 \times 3}$ represent the 3D coordinates of the regressed hand joints, while the mesh vertices $V \in \mathbb{R}^{778 \times 3}$ correspond to the 3D coordinates of the hand surface. For simplicity, we use the same notation for the parameters and outputs of the left and right hands unless explicitly stated otherwise.

C. TRAINING OBJECTIVE

To train our model, we use the loss functions adopted in the prior work [10], [12].

1) 3D HAND JOINTS LOSS

The loss terms used to assess the accuracy of the estimated 3D hand joints are defined as follows. The 3D hand joints loss $\mathcal{L}_{\text{joints}}$ is given by the L1 distance between the predicted and ground-truth joint positions, whereas the interaction hand joints loss $\mathcal{L}_{\text{interhand}}$ is the L2 distance between the predicted and ground-truth relative joint offsets of the left and right hands, measuring their positional consistency:

$$\mathcal{L}_{\text{joints}} = \frac{1}{N_J} \sum_{i=1}^{N_J} \|\hat{J}_i - J_i\|_1, \quad (5)$$

$$\mathcal{L}_{\text{interhand}} = \frac{1}{N_J} \sum_{i=1}^{N_J} \|(\hat{J}_{\text{left},i} - \hat{J}_{\text{right},i}) - (J_{\text{left},i} - J_{\text{right},i})\|_2, \quad (6)$$

where $\hat{J}_i$ and J_i denote the predicted and ground-truth 3D coordinates of the i -th hand joint, and N_J is the total number of hand joints.

2) 3D HAND MESH VERTICES LOSS

The loss terms for assessing the accuracy of the estimated 3D hand mesh vertices are defined as follows. The 3D hand mesh vertices loss $\mathcal{L}_{\text{vertices}}$ is given by the L1 distance between the predicted and ground-truth vertex positions, encouraging accurate reconstruction of the hand mesh structure:

$$\mathcal{L}_{\text{vertices}} = \frac{1}{N_V} \sum_{i=1}^{N_V} \|\hat{V}_i - V_i\|_1, \quad (7)$$

where $\hat{V}_i$ and V_i represent the predicted and ground-truth 3D coordinates of the i -th hand mesh vertex, and N_V is the total number of vertices.

3) MANO LOSS

The MANO loss quantifies the discrepancy between the predicted and ground-truth MANO pose θ and shape parameters β . This term encourages the model to produce accurate hand pose and shape representations:

$$\mathcal{L}_{\text{MANO}} = \|\hat{\theta} - \theta\|_2 + \|\hat{\beta} - \beta\|_2. \quad (8)$$

4) TOTAL HAND LOSS

These four loss terms are linearly combined to form the final training objective $\mathcal{L}_{\text{hand}}$:

$$\mathcal{L}_{\text{hand}} = \lambda_\gamma \mathcal{L}_{\text{joints}} + \lambda_\delta \mathcal{L}_{\text{interhand}} + \lambda_\epsilon \mathcal{L}_{\text{vertices}} + \lambda_\zeta \mathcal{L}_{\text{MANO}}, \quad (9)$$

where the balancing hyperparameters $\lambda_\gamma, \lambda_\delta, \lambda_\epsilon$, and λ_ζ correspond to $\mathcal{L}_{\text{joints}}$, $\mathcal{L}_{\text{interhand}}$, $\mathcal{L}_{\text{vertices}}$, and $\mathcal{L}_{\text{MANO}}$, respectively.

IV. DATASET COLLECTION

A. SYNTHETIC DATASET: N-HOT3D

Training our model requires event data with ground-truth annotations of both hands from an egocentric viewpoint. However, no existing dataset provides egocentric event data with such annotations. To address this limitation, we first construct a synthetic event dataset, N-HOT3D, by applying the event simulator v2e [69] to the HOT3D dataset [23]. Specifically, we use a subset of the Aria glasses data within HOT3D, consisting of nine subjects. Sample data from N-HOT3D are illustrated in Fig. 4 (top).

For data generation, we utilize the MANO parameters and the camera extrinsic and intrinsic parameters provided by HOT3D. We first apply distortion correction to the videos and then convert them into event data using the simulator. The spatial resolution of the output events is set to 346×260 pixels, matching the resolution of the DAVIS346 event camera [70], which is also used in Ev2Hands [10]. Additionally, we project the provided 3D mesh ground-truth annotations onto 2D space to generate ground-truth hand segmentation masks. We also compute bounding box annotations from the projected meshes to enable detection training. The annotation quality of N-HOT3D depends on that of the original HOT3D annotations, since we directly use the MANO parameters and camera calibration provided by HOT3D. The released HOT3D annotations had already been manually and visually inspected for all frames to exclude lower-quality poses.

N-HOT3D is split by both subject identity and capture sequence. The training and validation sets share the same subjects but contain disjoint capture sequences, ensuring that no identical sequence appears in both splits. Specifically, we use subjects P0001, P0003, P0009, P0010, P0011, P0012, and P0015 for training and validation, while the evaluation set is composed of unseen subjects, P0002 and P0014. N-HOT3D contains a total of 480,120 frames, divided into 334,190 for training, 83,760 for validation, and 62,170 for evaluation. Compared to the preliminary version of N-HOT3D introduced in our earlier conference paper [12], this work extends the dataset in several aspects. First, we visually inspected all frames and regenerated the segmentation masks for frames where mask generation had failed in the previous version. As a result, the dataset composition has been updated from 447,704 frames in the previous version to 480,120 frames in this work. Second, we newly provide bounding box annotations computed from

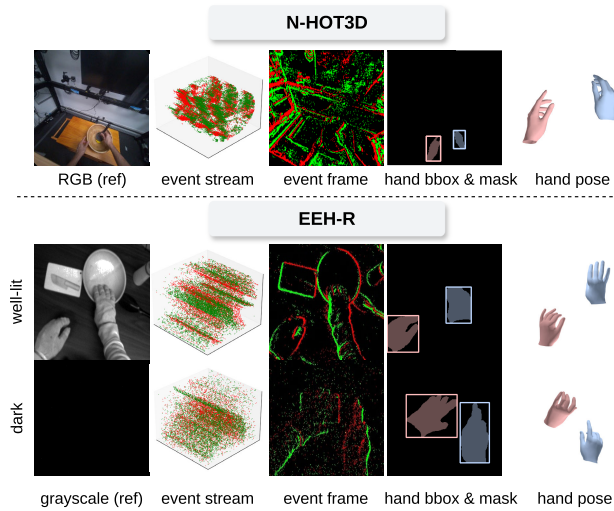


FIGURE 4. Dataset samples. The top shows the synthetic dataset N-HOT3D, and the bottom shows the real-world dataset EEH-R.

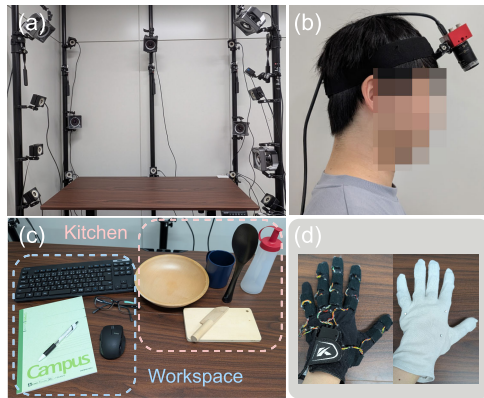


FIGURE 5. Capture system setup. (a) Arrangement of OptiTrack cameras, (b) head-mounted DAVIS346 event camera, (c) objects from kitchen and workspace categories, (d) MoCap gloves (left) and plain fabric gloves worn over MoCap gloves (right).

the projected meshes, enabling detection training in addition to segmentation. Finally, whereas the data split was not explicitly described in the previous version, we clearly define the splitting protocol in this work.

B. REAL DATASET: EEH-R

To evaluate the effectiveness of our proposed method on real event camera data, we construct a real-world dataset, EEH-R. Our dataset includes recordings of 8 subjects performing hand-object interactions under both well-lit and dark conditions. In total, it provides over two hours of recording across 85 sequences with 1,019,716 ground-truth annotations. Sample data from EEH-R are illustrated in Fig. 4 (bottom).

1) CAPTURE SYSTEM SETUP

We design a capture system to simultaneously record ego-centric event streams and accurate ground-truth hand/camera poses. An overview of our setup is illustrated in Fig. 5. For

TABLE 2. EEH-R dataset statistics by scene and lighting condition.

Scene	Lighting	#Sequences	#Annotations
Kitchen	Dark	23	287,566
Kitchen	Well-lit	19	238,340
Workspace	Dark	20	233,470
Workspace	Well-lit	23	260,340
Total		85	1,019,716

event data capture, we use a DAVIS346 [70], which provides both asynchronous events and synchronized grayscale frames. The grayscale frames are recorded at 30 fps.

To obtain precise local hand poses, we use MoCap Gloves [71] equipped with 16 IMU sensors, capable of capturing fine-grained finger articulations including finger curvature and palm arching. To preserve natural hand appearance in the event data, subjects wear plain fabric gloves over the MoCap gloves, concealing the sensors while maintaining texture characteristics similar to bare hands. For global positioning, we use an OptiTrack system consisting of 16 cameras (6 PrimeX22 and 10 PrimeX13) managed by Motive software [72]. Four optical markers are attached to both the top of the event camera, and another four optical markers are attached to the dorsal surface of each glove, enabling precise tracking of camera and hand positions. The ground-truth hand and camera poses are recorded at 120 fps. The average 3D calibration error of the motion-capture system was below 0.1972 mm, and the quality of the global hand and camera positions therefore depends on this calibration error.

2) SCENES

Our dataset covers two lighting conditions and two scene categories, recorded with eight subjects (6 male, 2 female). For lighting conditions, we capture sequences under both well-lit (average illuminance of 457 lux) and dark (average illuminance of 3.5 lux) conditions. For scene categories, we design kitchen and workspace scenarios to cover diverse hand-object interactions. The kitchen scene includes everyday objects such as bowls, spoons, cups, bottles, cutting boards, and knives, while the workspace scene contains notebooks, keyboards, mice, pens, and glasses. Each sequence lasts approximately 2 minutes, during which subjects freely interact with the objects. Multiple sequences were recorded for each combination of lighting and scene conditions. Detailed statistics of our dataset are summarized in Table 2.

3) ANNOTATIONS

The MoCap gloves provide 16 joint positions per hand (wrist and three joints per finger) in the local glove coordinate system. Following prior hand pose dataset construction pipelines [73], [74], we obtain MANO parameter annotations by fitting the MANO hand model to the joint positions provided by the MoCap gloves.

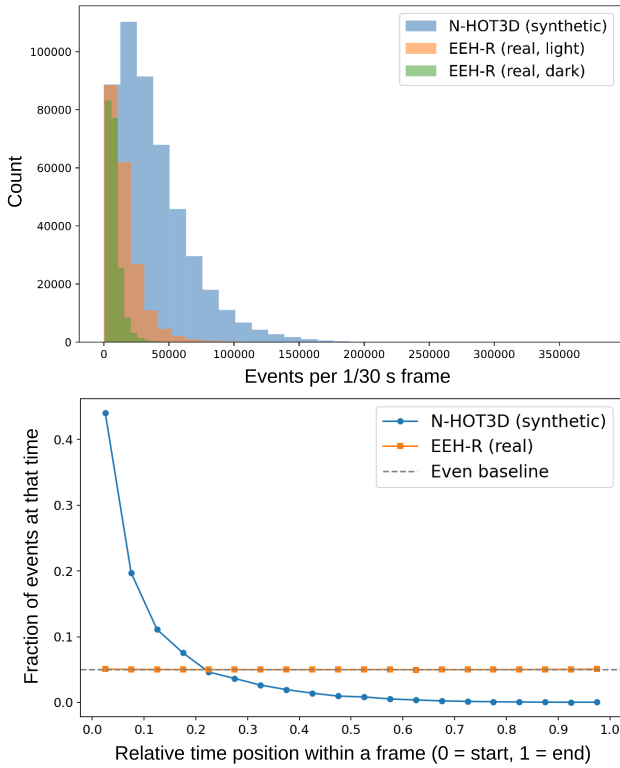


FIGURE 6. Analysis of event characteristics in the synthetic and real-world datasets. (Top) Distribution of the number of events per frame (1/30 s). (Bottom) Temporal distribution of events within a frame (1/30 s).

In addition to 3D annotations, we construct 2D hand segmentation masks for a subset of scenes. For well-lit scenes, we apply SAM3 [75] to grayscale images to automatically generate ground-truth hand masks, resulting in 198,410 annotated frames. For dark scenes, where grayscale images lack sufficient contrast, we manually annotate segmentation masks for 1,000 event frames. As in the original HOT3D dataset, all frames in EEH-R were manually and visually inspected to remove lower-quality poses.

EEH-R is split by subject identity across training, validation, and evaluation. Specifically, subjects P05, P07, P08, and P09 are used for training, P03 and P04 for validation, and P06 and P10 for evaluation. The EEH-R dataset contains a total of 1,019,716 frames, divided into 636,433 for training, 164,727 for validation, and 218,556 for evaluation. Approval for EEH-R dataset recording was obtained from the ethics committee of Keio University under 2025-131.

4) DOMAIN GAP BETWEEN SYNTHETIC AND REAL DATA

Compared with EEH-R, N-HOT3D involves more locomotion and head motion, which causes intense brightness changes across the entire scene. As shown in the top of Fig. 6, N-HOT3D consequently yields a larger number of events per frame. The bottom of Fig. 6 shows how events are distributed within the time window of each frame. While events in the real data occur almost uniformly over time, events in

N-HOT3D are concentrated at the beginning of each frame, resulting in an unnatural temporal distribution, since they are synthetically generated by an event simulator. In addition, EEH-R contains sensor noise inherent to real event cameras. The visual differences between the two datasets can be observed in the event streams shown in Fig. 4.

5) BIAS AND LIMITATIONS

EEH-R was collected in a controlled laboratory environment. Extending the dataset to in-the-wild settings is a promising direction for future work. Although the IMU-based sensor gloves provide accurate ground truth, they inevitably alter the visual appearance of the hands. Adopting a marker-less motion capture system would allow bare-hand recordings, further increasing the appearance diversity of the hands.

V. EXPERIMENTS

A. IMPLEMENTATION DETAILS

For training and evaluation, we follow the official train/validation/test splits of the N-HOT3D and EEH-R datasets. Input event frames are center-cropped and resized to a resolution of $H \times W = 224 \times 224$ pixels. All experiments were conducted using a single NVIDIA RTX 6000 Ada Generation GPU.

The Hand Detector is implemented using the YOLO26 framework [65] along with its training and inference ecosystem. The model is initialized from pretrained YOLO26 weights and fine-tuned for our task. Since LNES consists of two channels, we adapt the network by introducing additional empty channels to match the expected input format. The segmentation model is trained using Adam [76] with a learning rate of 1.0×10^{-3} for 100 epochs with a batch size of 64. The predicted segmentation masks are dilated using a kernel size of 7×7 with one iteration to extend the hand regions. When multiple detections of the same hand (left or right) are produced, we select the one with the highest confidence score. The detection confidence threshold is set to 0.5 for N-HOT3D and 0.8 for EEH-R.

For the hand reconstruction stage, the EfficientNetV2-S backbone produces feature maps with channel dimension $C = 1280$ and spatial resolution $H' \times W' = 7 \times 7$ pixels. The backbone is initialized with ImageNet-1K pretrained weights. The model is trained using the AdamW [77] optimizer with a learning rate of 2.0×10^{-5} for 30 epochs with a batch size of 32. The total hand loss weights are set as $\lambda_\gamma = 2.0$, $\lambda_\delta = 1.0$, $\lambda_\epsilon = 2.0$, and $\lambda_\zeta = 1.0$.

B. BASELINE METHODS

We compare our method with the following baselines to show its performance in event-based 3D hand mesh reconstruction. **EventHands** [8] is a frame-based approach that operates solely on event data. As it is regarded as one of the state-of-the-art methods for third-person event-based hand reconstruction, we adopt it as a baseline. Since this method

predicts only a single hand at a time, we trained two separate models for the left and right hands on N-HOT3D and EEH-R.

Ev2Hands [10] leverages only event data and extracts the features using a point cloud-based representation. It is the only prior method that predicts both hands simultaneously. Since it requires event data annotated as left hand, right hand, or background, whose labels are not available in N-HOT3D or EEH-R, we first train the model on the annotated Ev2Hands-S [10] dataset and then fine-tune it on N-HOT3D and EEH-R.

EventEgoHands [12] is the first egocentric hand mesh reconstruction method. This method estimates hand regions using mask-only estimation without distinguishing between left and right hands. The reconstruction stage is point cloud-based and employs only cross-attention between hand features to estimate hand meshes. Since this method does not differentiate between left and right hands, it always outputs two hands regardless of their presence.

C. EVALUATION METRICS

Following Ev2Hands [10], we adopt the Percentage of Correct Keypoints (PCK) and the area under the PCK curve (AUC) with thresholds ranging from 0 to 100 millimeters (mm). We also use the Relative AUC (R-AUC) and Right-root Relative AUC (RR-AUC) to assess the 3D hand pose estimation performance. The R-AUC is computed from joint positions expressed relative to the wrist of each hand, thereby measuring the accuracy of 3D joint positions for each hand independently. In contrast, RR-AUC is computed by expressing the joint positions of both hands relative to the right wrist, thereby evaluating the relative 3D configuration between the two hands. In addition, we report the Mean Per Joint Position Error (MPJPE) and Mean Per Vertex Position Error (MPVPE) in millimeters, which are standard metrics for 3D hand mesh reconstruction. Both MPJPE and MPVPE are computed after aligning each hand by its wrist joint position. For all the metrics above, only the hands successfully detected by the Hand Detector are included in the evaluation, and undetected hands are excluded from the aggregation of joint and mesh errors. The detection performance itself is evaluated separately using the metrics described below.

Furthermore, to evaluate the performance of the Hand Detector, we employ the mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds, namely mAP@50 and mAP@50–95. mAP@50 denotes the mAP at an IoU threshold of 0.5, while mAP@50–95 is averaged over IoU thresholds from 0.5 to 0.95 with a step size of 0.05. These metrics are widely used in instance segmentation and detection tasks, including the YOLO series [65].

VI. RESULTS

A. EVALUATION ON N-HOT3D DATASET

1) QUANTITATIVE EVALUATION

Table 3 presents the quantitative results on the N-HOT3D dataset. Our method significantly outperforms all existing methods across all metrics. Compared to the closest baseline

TABLE 3. Quantitative evaluation on N-HOT3D. We report the mean $\pm$ standard deviation over three runs with different random seeds. The best values are shown in bold.

	R-AUC ($\uparrow$)	RR-AUC ($\uparrow$)	MPJPE [mm] ($\downarrow$)	MPVPE [mm] ($\downarrow$)
EventHands [8]	0.253 $\pm$ 0.009	0.190 $\pm$ 0.008	105.62 $\pm$ 1.40	97.29 $\pm$ 1.29
Ev2Hands [10]	0.236 $\pm$ 0.001	0.209 $\pm$ 0.000	112.56 $\pm$ 0.46	107.00 $\pm$ 0.44
EventEgoHands [12]	0.417 $\pm$ 0.004	0.261 $\pm$ 0.021	64.83 $\pm$ 0.56	60.53 $\pm$ 0.52
Ours	0.661 $\pm$ 0.004	0.528 $\pm$ 0.003	43.01 $\pm$ 0.27	39.96 $\pm$ 0.26

EventEgoHands, our enhanced model demonstrates superior precision, reducing MPJPE and MPVPE by 33.7% and 34.0%, respectively.

2) QUALITATIVE EVALUATION

Fig. 7 shows qualitative comparisons on the N-HOT3D dataset. Our method reconstructs hand meshes that are closest to the ground truth compared with other methods. EventHands, which estimates each hand independently, often produces inconsistent hand positions and fails to capture the spatial relationship between the two hands. Ev2Hands, which is based on point cloud processing, limits the number of input event points due to computational constraints. In egocentric settings where a large number of events are generated, uniform subsampling reduces the number of events belonging to the hand region, which prevents the model from learning sufficiently detailed hand shape representations and often results in nearly identical hand poses. EventEgoHands improves the spatial consistency between the two hands by filtering hand regions. However, the reconstructed hand poses lack diversity and tend to produce similar poses across different scenes.

In contrast, the proposed method reconstructs more accurate hand shapes and better represents both inter-hand relationships and detailed finger motions such as fingertip positions and hand rotations. Notably, in scenes where only a single hand is visible, our approach predicts an accurate pose for the visible hand while correctly identifying the absence of the other. Whereas existing methods often produce incorrect poses or erroneously predict a second hand that is not present, our framework remains robust to such scenarios. This improvement is attributed to the Hand Detector and Adaptive Attention, which effectively learn both single-hand features and interactions between the two hands.

B. EVALUATION ON EEH-R DATASET

1) QUANTITATIVE EVALUATION

Table 4 shows the quantitative results on the real-world EEH-R dataset. Our method again achieves the best performance across all evaluation metrics. Compared with the best existing method, our method reduces MPJPE by 7.90 mm (18.8%) and MPVPE by 7.13 mm (18.1%). The improvement margin is smaller than that observed on the synthetic dataset. We attribute this difference to the characteristics of real-world data, where the camera is typically closer to the hands, resulting in larger hand regions and relatively easier detection. In contrast, the synthetic N-HOT3D dataset

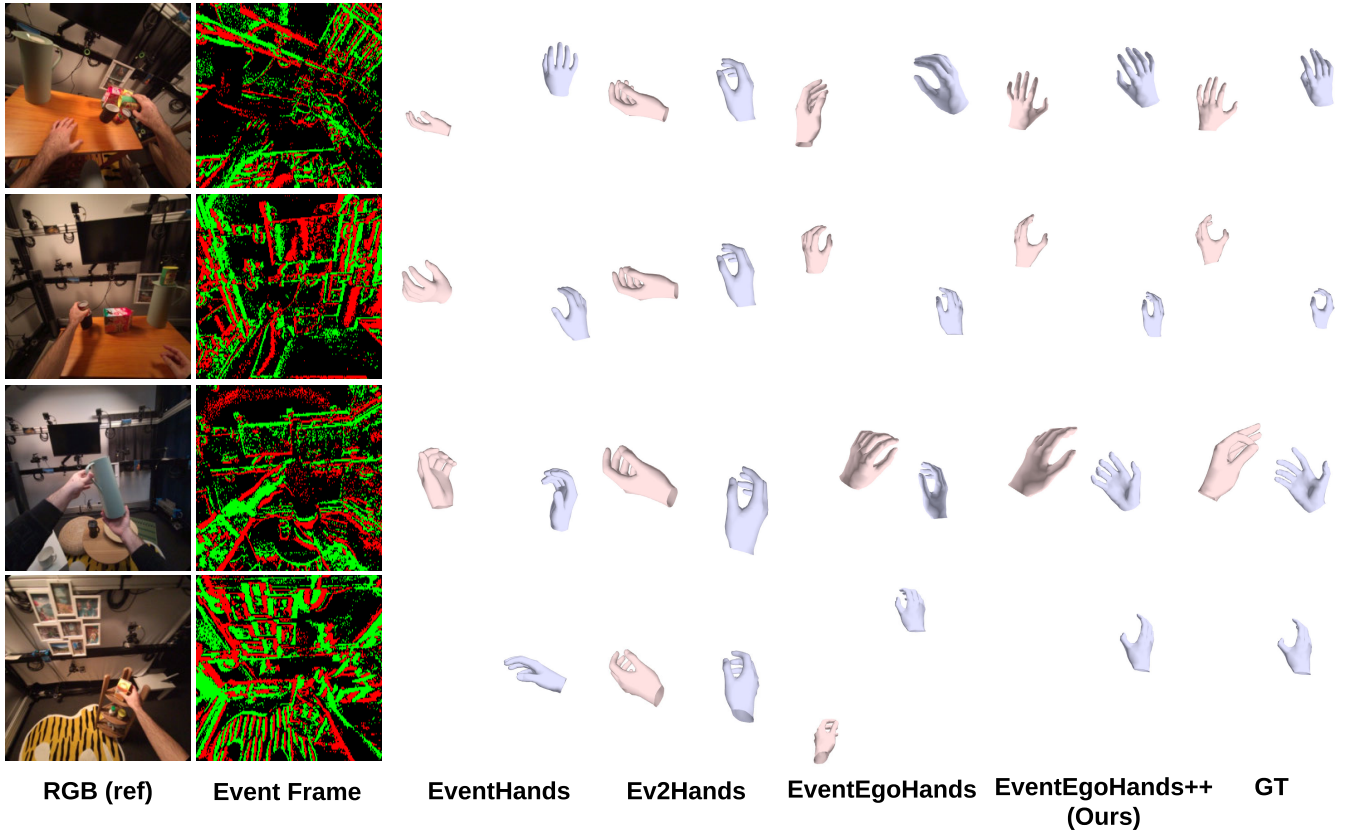


FIGURE 7. Qualitative evaluation on N-HOT3D. We compared our method with EventHands [8], Ev2Hands [10] and EventEgoHands [12]. RGB images were not used as input and are shown for reference only. In order from the top, the scenes show “pouring from a can”, “picking up a can”, “holding a pot”, and “picking up a carton”.

TABLE 4. Quantitative evaluation on EEH-R under different lighting conditions. All values are reported as mean $\pm$ standard deviation over three runs with different random seeds. The best values are shown in bold.

Method	Lighting	R-AUC ($\uparrow$)	RR-AUC ($\uparrow$)	MPJPE [mm] ($\downarrow$)	MPVPE [mm] ($\downarrow$)
EventHands [8]	All	0.583 ± 0.004	0.287 ± 0.004	42.08 ± 0.44	39.29 ± 0.41
	Well-lit	0.591 ± 0.005	0.298 ± 0.006	41.38 ± 0.59	38.56 ± 0.53
	Dark	0.576 ± 0.003	0.278 ± 0.004	42.71 ± 0.35	39.95 ± 0.34
Ev2Hands [10]	All	0.502 ± 0.002	0.250 ± 0.004	52.02 ± 0.54	48.59 ± 0.48
	Well-lit	0.492 ± 0.004	0.262 ± 0.002	53.57 ± 0.87	50.05 ± 0.80
	Dark	0.511 ± 0.005	0.239 ± 0.006	50.64 ± 0.69	47.29 ± 0.61
EventEgoHands [12]	All	0.478 ± 0.008	0.324 ± 0.012	56.59 ± 1.69	52.94 ± 1.64
	Well-lit	0.481 ± 0.003	0.340 ± 0.005	56.27 ± 1.13	52.65 ± 1.05
	Dark	0.475 ± 0.018	0.310 ± 0.022	56.95 ± 3.35	53.27 ± 3.12
Ours	All	0.691 ± 0.003	0.551 ± 0.005	34.18 ± 0.22	32.16 ± 0.23
	Well-lit	0.693 ± 0.003	0.606 ± 0.006	32.68 ± 0.30	30.68 ± 0.29
	Dark	0.689 ± 0.003	0.476 ± 0.005	35.57 ± 0.18	33.53 ± 0.22

contains scenarios where the hands appear smaller due to larger interaction distances and fisheye distortion correction, making hand detection and reconstruction more challenging.

Another notable observation is that the third-person baselines, EventHands and Ev2Hands, outperform the egocentric prior work EventEgoHands on several metrics. We attribute this to the characteristics of the real-world data. All sequences in EEH-R are captured during desk-based tasks, where the camera-to-hand distance is short and nearly constant, and

the hands occupy a large portion of the frame, making hand detection itself easier, as also indicated by the mAP in Table 6. Furthermore, as shown in Fig. 6, the number of events per frame in EEH-R is smaller than in N-HOT3D, so hand-related events are relatively dominant within each frame, creating a setting closer to the third-person scenario where only hand events are observed. In addition, R-AUC, MPJPE, and MPVPE evaluate hand pose errors in wrist-relative coordinates, and the local hand poses show little

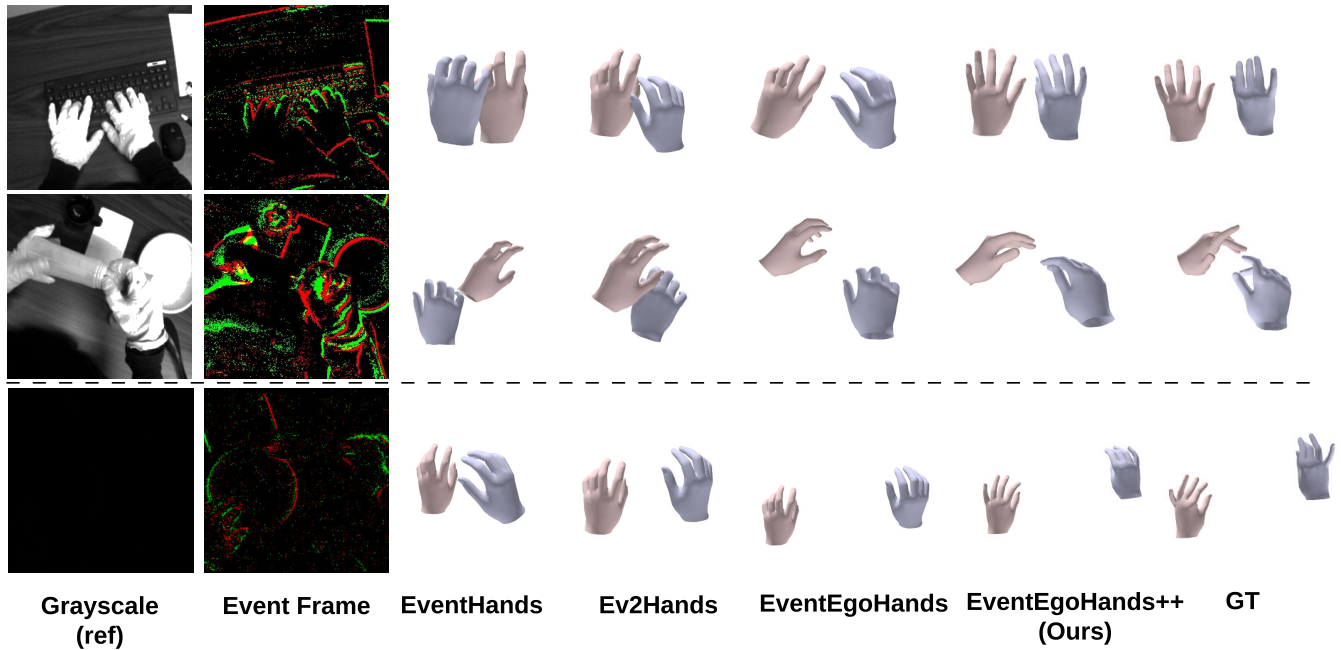


FIGURE 8. Qualitative evaluation on EEH-R. We compared our method with EventHands [8], Ev2Hands [10] and EventEgoHands [12]. Grayscale images were not used as input and are shown for reference only. The grayscale images at the bottom are actual captured images, which appear almost empty due to the low-light environment. In order from the top, the scenes show “typing on a keyboard”, “holding a bottle”, and “using a spoon”.

TABLE 5. Comparison of inference speed (FPS) across methods. Values are reported as the mean $\pm$ standard deviation.

Method	Segmentation	Hand	Overall
EventHands [8]	–	415.88 ± 143.65	415.88 ± 143.65
Ev2Hands [10]	–	11.16 ± 0.96	11.16 ± 0.96
EventEgoHands [12]	209.44 ± 39.24	12.53 ± 0.52	12.41 ± 0.35
Ours (Single Hand)	149.94 ± 36.18	62.46 ± 16.27	45.42 ± 9.10
Ours (Both Hands)	160.28 ± 35.91	64.09 ± 12.80	39.87 ± 6.63

difference across the baselines, as seen in Fig. 8. In contrast, EventEgoHands surpasses the third-person baselines on RR-AUC, which evaluates the relative position between the two hands, and this advantage is also evident in Fig. 8.

2) QUALITATIVE EVALUATION

Fig. 8 shows qualitative results on the EEH-R dataset. Consistent with the quantitative results, the differences between methods are smaller than those observed on the synthetic dataset. This is because EEH-R consists of static desk-based tasks with less subject motion than N-HOT3D.

Nevertheless, existing methods still exhibit limitations similar to those observed on the synthetic dataset. They often fail to accurately capture inter-hand relationships and fine details of intra-hand shape. In contrast, our method reconstructs hand meshes that are closest to the ground truth and produces more diverse hand poses across different scenes. These results demonstrate that the proposed method remains effective when applied to real event camera data, including challenging low-light environments.

3) RUNTIME ANALYSIS

Table 5 shows the inference speed of the proposed method and the baselines in frames per second (FPS), reported as the mean $\pm$ standard deviation over 100 samples. EventHands is extremely fast. However, as shown in Table 3 and Table 4, its accuracy is insufficient. Ev2Hands and EventEgoHands rely on point-cloud-based processing, which involves higher-dimensional inputs than image-based processing, resulting in low FPS, particularly for hand mesh reconstruction. In contrast, the proposed method is fast in both segmentation and hand mesh reconstruction, achieving practical inference speed in the overall pipeline. Moreover, when only one hand is detected, inference is slightly faster because certain operations, such as cross-attention in Adaptive Attention, are skipped.

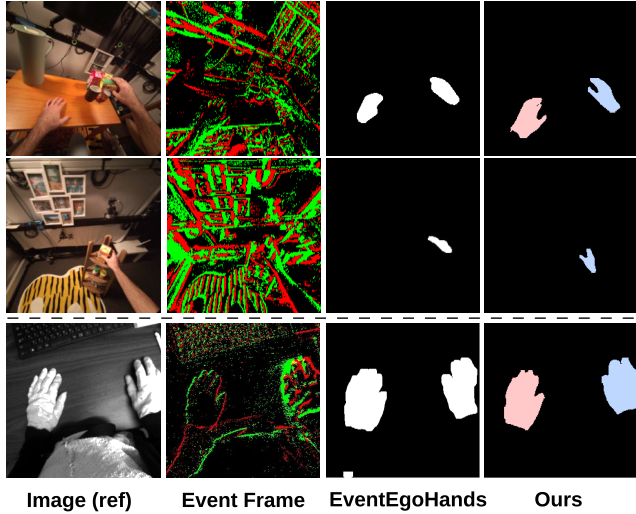
C. HAND SEGMENTATION PERFORMANCE

We conduct an experiment to validate the effectiveness of the proposed Hand Detector. Table 6 compares the performance of the mask segmentation methods on the N-HOT3D and EEH-R datasets. Qualitative segmentation results are shown in Fig. 9.

The Hand Detector significantly outperforms the U-Net on the N-HOT3D dataset, improving mAP@50 from 0.272 to 0.750 and mAP@50–95 from 0.042 to 0.407. The improved accuracy on N-HOT3D, where hands occupy only a small portion of the frame, suggests the effectiveness of instance segmentation that jointly learns bounding boxes and masks, compared to the simple mask estimation performed by U-Net.

TABLE 6. Performance comparison of segmentation methods on synthetic and real datasets. All values are reported as mean $\pm$ standard deviation over three runs with different random seeds. The best values are shown in bold.

Dataset	Method	All		Well-lit		Dark	
		mAP@50 ($\uparrow$)	mAP@50–95 ($\uparrow$)	mAP@50 ($\uparrow$)	mAP@50–95 ($\uparrow$)	mAP@50 ($\uparrow$)	mAP@50–95 ($\uparrow$)
N-HOT3D	U-Net in EventEgoHands	0.272 $\pm$ 0.025	0.042 $\pm$ 0.006	–	–	–	–
	Hand Detector	0.750 $\pm$ 0.013	0.407 $\pm$ 0.012	–	–	–	–
EEH-R	U-Net in EventEgoHands	0.928 $\pm$ 0.003	0.655 $\pm$ 0.002	0.929 $\pm$ 0.003	0.656 $\pm$ 0.002	0.899 $\pm$ 0.025	0.527 $\pm$ 0.021
	Hand Detector	0.901 $\pm$ 0.003	0.657 $\pm$ 0.005	0.901 $\pm$ 0.003	0.657 $\pm$ 0.005	0.887 $\pm$ 0.095	0.639 $\pm$ 0.091

**FIGURE 9.** Hand segmentation results. The top two rows correspond to N-HOT3D and the bottom row corresponds to EEH-R. Since the segmentation produced by EventEgoHands is a binary classification, the mask region is shown in white, whereas our method uses different colors for the left and right hands.

On EEH-R, we evaluate the performance under each lighting condition. Since EEH-R contains fewer events per frame and the hands are more visible within the scene, the mAP values are higher than those on N-HOT3D, and the Hand Detector performs slightly worse than the U-Net in terms of mAP@50. However, the Hand Detector outperforms the U-Net on mAP@50–95, which evaluates mask overlap more strictly. In particular, the improvement is substantial under the dark condition.

D. ABLATION STUDY

We conduct ablation studies to evaluate the contribution of each component to the proposed framework. All experiments are performed on the N-HOT3D dataset.

1) EFFECT OF THE PROPOSED COMPONENTS

Table 7 shows the results when removing the Hand Detector and the Adaptive Attention. Even when both components are removed, the proposed framework still outperforms existing methods. This suggests that the frame-based architecture, which jointly models two hands, is more effective than previous single-hand or point-cloud-based approaches. Note

TABLE 7. Ablation study on the proposed components on N-HOT3D. All values are reported as mean $\pm$ standard deviation over three runs with different random seeds. The best values are shown in bold.

Detector	Attention	R-AUC ($\uparrow$)	RR-AUC ($\uparrow$)	MPJPE [mm] ($\downarrow$)	MPVPE [mm] ($\downarrow$)
		0.548 $\pm$ 0.002	0.389 $\pm$ 0.002	46.79 $\pm$ 0.21	43.79 $\pm$ 0.19
	✓	0.553 $\pm$ 0.002	0.394 $\pm$ 0.002	47.15 $\pm$ 0.19	43.78 $\pm$ 0.18
✓		0.652 $\pm$ 0.001	0.469 $\pm$ 0.002	44.01 $\pm$ 0.03	41.04 $\pm$ 0.03
✓	✓	0.661 $\pm$ 0.004	0.528 $\pm$ 0.003	43.01 $\pm$ 0.27	39.96 $\pm$ 0.26

TABLE 8. Effect of introducing the Hand Detector to existing methods on N-HOT3D. All values are reported as mean $\pm$ standard deviation over three runs with different random seeds. underline indicates improvement within each method.

Method	R-AUC ($\uparrow$)	RR-AUC ($\uparrow$)	MPJPE [mm] ($\downarrow$)	MPVPE [mm] ($\downarrow$)
EventHands [8]	0.253 $\pm$ 0.009	0.190 $\pm$ 0.008	105.62 $\pm$ 1.40	97.29 $\pm$ 1.29
EventHands w/ Detector	<u>0.293 $\pm$ 0.001</u>	<u>0.234 $\pm$ 0.007</u>	108.09 $\pm$ 4.77	99.33 $\pm$ 4.31
Ev2Hands [10]	0.236 $\pm$ 0.001	0.209 $\pm$ 0.000	112.56 $\pm$ 0.46	107.00 $\pm$ 0.44
Ev2Hands w/ Detector	<u>0.242 $\pm$ 0.001</u>	<u>0.214 $\pm$ 0.001</u>	<u>110.05 $\pm$ 0.07</u>	<u>104.59 $\pm$ 0.06</u>
EventEgoHands [12]	0.417 $\pm$ 0.004	0.261 $\pm$ 0.021	64.83 $\pm$ 0.56	60.53 $\pm$ 0.52
EventEgoHands w/ Detector	<u>0.437 $\pm$ 0.013</u>	<u>0.277 $\pm$ 0.012</u>	68.64 $\pm$ 4.65	63.86 $\pm$ 4.33

that without the Hand Detector, the two hands cannot be localized or separated, so both hands are always treated as present ($m_l = m_r = 1$) and the Adaptive Attention always applies both self-attention and cross-attention.

When Adaptive Attention alone is introduced without the Hand Detector, the R-AUC slightly improves, while MPJPE slightly degrades. This indicates that attention alone is insufficient when hand regions are not properly localized; without the detector, the attention mechanism is always forced to process the entire frame, including background noise, rather than focusing on the actual hand instances.

In contrast, introducing the Hand Detector improves performance across all metrics, and combining it with Adaptive Attention yields further gains. In particular, the gain from Adaptive Attention is most pronounced in RR-AUC, which evaluates the relative position between the two hands, improving from 0.469 to 0.528 (a 12.6% relative improvement). These results indicate a complementary division of roles: the Hand Detector is the primary source of overall accuracy by localizing and identifying each hand, whereas Adaptive Attention specifically strengthens the inter-hand relative positioning.

2) GENERAL APPLICABILITY OF THE HAND DETECTOR

Table 8 evaluates the effect of integrating the Hand Detector into existing methods. Across all three methods, R-AUC and

TABLE 9. Ablation study of loss components on N-HOT3D. The best values are shown in bold.

$\mathcal{L}_{\text{MANO}}$	$\mathcal{L}_{\text{interhand}}$	$\mathcal{L}_{\text{joints}}$	$\mathcal{L}_{\text{vertices}}$	R-AUC ($\uparrow$)	RR-AUC ($\uparrow$)	MPJPE [mm] ($\downarrow$)	MPVPE [mm] ($\downarrow$)
✓				0.095	0.073	153.64	142.10
	✓			0.311	0.239	80.12	74.00
✓	✓	✓		0.657	0.521	43.50	40.43
✓	✓	✓	✓	0.661	0.528	43.01	39.96

TABLE 10. Comparison of Vision Encoders on N-HOT3D. The best values are shown in bold.

Vision Encoder	Params [M]	R-AUC ($\uparrow$)	RR-AUC ($\uparrow$)	MPJPE ($\downarrow$)	MPVPE ($\downarrow$)
ResNet50 [78]	23.5	0.653	0.519	43.98	40.94
ViT-B [35]	85.8	0.610	0.478	47.14	44.06
EfficientNetV2-S [67] (Ours)	20.2	0.661	0.528	43.01	39.96

TABLE 11. Comparison of Hand Detectors on N-HOT3D. The best values are shown in bold, and the second-best values are underlined.

Detector	mAP@50 ($\uparrow$)	mAP@50–95 ($\uparrow$)	FPS
YOLO11 [79]	0.699	0.375	221.03
RF-DETR [80]	0.769	<u>0.398</u>	68.49
YOLO26 [65] (Ours)	<u>0.750</u>	0.407	<u>209.54</u>

RR-AUC consistently improve after incorporating the proposed Hand Detector. These results indicate that extracting hand regions from event data is generally beneficial and can improve a wide range of existing hand reconstruction methods.

3) LOSS FUNCTIONS

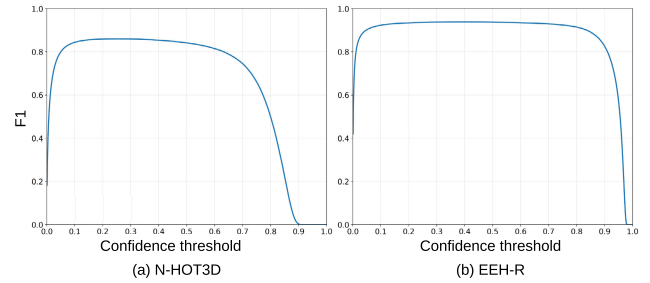
Table 9 shows the ablation study on the loss components. We adopt the loss functions used in Ev2Hands and EventEgoHands, which also output two hands. The results show that combining all loss components achieves the best performance.

4) VISION ENCODER

Table 10 shows the ablation study on the vision encoder. We compare EfficientNetV2-S [67] with ResNet50 [78], which has a comparable number of parameters and is widely used in vision tasks, and with ViT-B [35], which has recently been adopted in many vision tasks. The results confirm that EfficientNetV2-S, which is commonly employed in hand pose estimation studies [61], [63], [68], is also well suited for our task. As vision encoders continue to advance, our framework is expected to benefit from adopting newer encoders in the future.

5) HAND DETECTOR

Table 11 shows the ablation study on the hand detector. We compare the adopted YOLO26 with its previous official version, YOLO11, and with RF-DETR, a Transformer-based detector widely used alongside YOLO. Although RF-DETR achieves a higher mAP@50, YOLO26 is the most balanced model in terms of both accuracy and speed, as indicated by mAP@50–95 and FPS.

**FIGURE 10. F1 score of the Hand Detector at IoU = 0.5 for varying confidence thresholds. (a) N-HOT3D (synthetic). (b) EEH-R (real).****TABLE 12. Ablation study of attention components on N-HOT3D. The best values are shown in bold.**

SelfAttn	CrossAttn	R-AUC ($\uparrow$)	RR-AUC ($\uparrow$)	MPJPE [mm] ($\downarrow$)	MPVPE [mm] ($\downarrow$)
		0.652	0.469	44.01	41.04
✓		0.657	0.517	43.37	40.34
	✓	0.649	0.518	43.93	40.90
✓	✓	0.661	0.528	43.01	39.96

TABLE 13. Comparison of dilation kernel sizes on N-HOT3D. The best values are shown in bold, and the second-best values are underlined.

Method	R-AUC ($\uparrow$)	RR-AUC ($\uparrow$)	MPJPE [mm] ($\downarrow$)	MPVPE [mm] ($\downarrow$)
w/o dilation	0.636	0.467	44.84	41.71
kernel size 3	0.659	0.504	43.16	40.12
kernel size 5	0.667	<u>0.525</u>	42.60	39.58
kernel size 7	<u>0.661</u>	0.528	43.01	39.96
kernel size 9	0.644	0.520	44.30	41.16

6) DETECTION THRESHOLD ANALYSIS

Fig. 10 shows the F1 score at IoU = 0.5 for varying confidence thresholds on the synthetic and real datasets. On both datasets, the F1 score remains stable over a range of thresholds before dropping sharply beyond a certain point. Since false positives introduce non-existent hands into the subsequent reconstruction stage, we empirically select a relatively high threshold that does not substantially degrade the F1 score, namely 0.5 for the synthetic dataset and 0.8 for the real dataset.

7) ADAPTIVE ATTENTION

Table 12 shows the ablation study on the attention components in Adaptive Attention. Self-attention mainly improves wrist-relative pose accuracy, while cross-attention improves RR-AUC, which evaluates the relative position between the two hands, by exchanging information across hands. Combining both achieves the best performance on all metrics.

8) MASK DILATION KERNEL SIZE

Table 13 shows the ablation study on the kernel size used to dilate the masks estimated by the Hand Detector. Hand mesh reconstruction performance is highest with a kernel size of 5×5 or 7×7 . Dilating the mask compensates for incomplete or partially missing segmentation results, whereas an excessively large kernel includes more non-hand regions, indicating that there is a limit to the benefit of dilation.

TABLE 14. Cross-dataset evaluation. The best values are shown in bold.

Train	Finetune	Test	R-AUC ($\uparrow$)	RR-AUC ($\uparrow$)	MPJPE [mm] ($\downarrow$)	MPVPE [mm] ($\downarrow$)
EEH-R	–	EEH-R	0.691	0.551	34.184	32.163
N-HOT3D	–	EEH-R	0.079	0.038	137.229	126.446
N-HOT3D	EEH-R	EEH-R	0.695	0.557	34.172	32.207

TABLE 15. Cross-illumination evaluation of the Hand Detector on EEH-R. Each model is trained and tested under different illumination conditions. The best values for each test condition are shown in bold.

Train	Test	mAP@50 ($\uparrow$)	mAP@50–95 ($\uparrow$)
Well-lit	Well-lit	0.935	0.679
Dark	Well-lit	0.425	0.201
All	Well-lit	0.901	0.657
Well-lit	Dark	0.749	0.366
Dark	Dark	0.792	0.548
All	Dark	0.887	0.639
Well-lit	All	0.934	0.678
Dark	All	0.427	0.202
All	All	0.928	0.655

We adopt a kernel size of 7×7 , which achieves the highest RR-AUC, as we prioritize the correct relative positioning of the two hands.

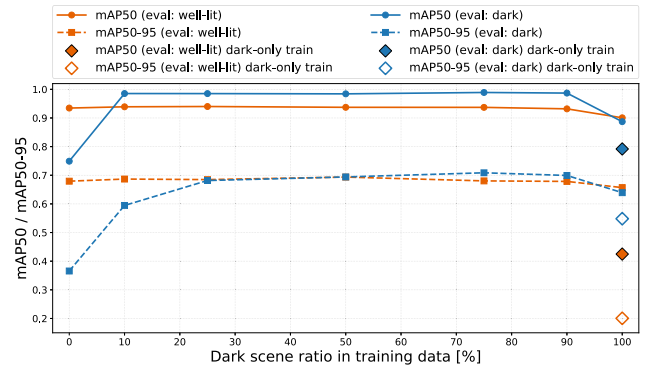
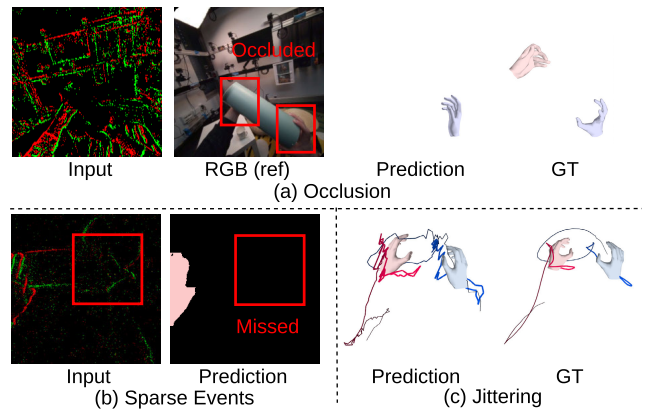
9) CROSS-DATASET EVALUATION

Table 14 shows the quantitative results of the cross-dataset evaluation. When the model is trained on N-HOT3D and evaluated on EEH-R, the performance drops substantially due to the domain gap between the two datasets. However, when the model pre-trained on N-HOT3D is fine-tuned on EEH-R, it achieves slightly higher accuracy than the model trained on EEH-R alone. This result indicates the potential of the synthetic dataset for pre-training.

10) IMPACT OF DARK-SCENE ANNOTATIONS ON THE HAND DETECTOR

Since the number of manually annotated dark-scene masks is much smaller than that of the automatically generated well-lit masks, we analyze how the amount of dark-scene supervision affects the detector performance on EEH-R. Table 15 compares models trained on each illumination condition separately. The model trained only on dark scenes performs worse under the dark condition than the model trained on both conditions, and its performance under the well-lit condition degrades substantially. This indicates that combining the automatically generated well-lit annotations with the manual dark-scene annotations is an effective supervision strategy.

We further analyze how many dark-scene annotations are required. We train the Hand Detector with all well-lit annotations in the training set while varying the ratio of dark-scene annotations from 0% to 100%, and evaluate the models separately under the well-lit and dark conditions. As shown in Fig. 11, the performance under the well-lit condition remains almost constant for both metrics regardless of the dark-scene data ratio. Under the dark condition,

**FIGURE 11. Effect of the dark-scene data ratio on the performance of the Hand Detector on EEH-R. All well-lit annotations are used, while the ratio of dark-scene annotations in the training set is varied from 0% to 100%.****FIGURE 12. Failure cases of the proposed method. (a) Missed detection and inaccurate pose estimation under severe occlusion by a held object. (b) Missed detection due to sparse events when the hand remains nearly static. (c) Temporal jittering in the predicted trajectory due to frame-by-frame estimation without temporal modeling. Red and blue lines indicate the wrist trajectories of the left and right hands, respectively.**

mAP@50–95 increases with the ratio and saturates once the ratio reaches approximately 25% of the dark training set. This demonstrates that adding only a small number of manually annotated dark frames to the well-lit data is sufficient for reliable detection in dark scenes.

E. FAILURE ANALYSIS

We identify three major failure patterns of the proposed method, as illustrated in Fig. 12.

1) OCCLUSION

When a hand is occluded by an object, the Hand Detector may fail to detect it. Even when the hand is detected, accurately

reconstructing the occluded parts, such as the fingertips, remains difficult when they are hidden by the object, as shown in Fig. 12 (a). Since hands frequently interact with surrounding objects in daily activities, explicitly modeling object shape and hand-object contact is a promising direction. While current event-based approaches primarily focus on hand regions, modeling hand-object interactions remains largely unexplored. Inspired by RGB-based methods [27], [81], incorporating distinct feature representations for both hand and object regions is a promising direction for improving reconstruction accuracy in complex interaction scenarios.

2) SPARSE EVENTS

Since an event camera only responds to brightness changes, events become sparse when the hand and head remain nearly static. In such cases, the hand may not be detected, as shown in Fig. 12 (b). This effect is more pronounced in low-light environments, where the detection performance is lower than in well-lit conditions, as indicated by the mAP in Table 6. Accumulating past events or introducing temporal processing would help mitigate this failure.

3) JITTERING

Since the proposed method operates on a frame-by-frame basis, the reconstructed hand meshes may exhibit temporal jittering across consecutive frames. As shown in the wrist trajectories in Fig. 12 (c), the predicted motion is less smooth than the ground truth. Note that such jittering is not specific to event-based methods, as frame-based RGB methods such as HaMeR [5] and WiLoR [6] also suffer from it. Leveraging the temporal nature of event streams and extending the framework to 4D hand mesh reconstruction is an important direction for future work. Recent approaches such as HaPTIC [38], which predicts coherent 4D hand trajectories from monocular videos by fusing temporal information across frames with attention and by directly estimating inter-frame depth changes rather than per-frame absolute depth, suggest promising strategies to resolve this issue.

VII. CONCLUSION

In this paper, we propose EventEgoHands++, a framework for event-based egocentric 3D hand mesh reconstruction. To address the challenges of background noise and the inability of prior work to distinguish between hands, we introduce a Hand Detector that performs instance-level hand detection by jointly estimating bounding boxes and masks. Furthermore, to overcome the performance degradation caused by fixed attention that ignores hand visibility, we also propose Adaptive Attention that dynamically adjusts the attention process depending on the presence of detected hands. In addition, we extend the synthetic N-HOT3D dataset with refined masks and newly added bounding-box annotations, and newly construct the real-world EEH-R dataset. Experiments on both synthetic and

real datasets demonstrate that the proposed method significantly improves reconstruction performance compared with existing approaches. We hope that this work contributes to advancing event-based hand reconstruction research and facilitates future studies on egocentric event-based vision systems.

REFERENCES

- [1] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3D hand pose estimation in single depth images: From single-view CNN to multi-view CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3593–3601.
- [2] C. Xu and L. Cheng, "Efficient hand pose estimation from a single depth image," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3456–3462.
- [3] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Real-time 3D hand pose estimation with 3D convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 956–970, Apr. 2019.
- [4] A. Boukhayma, R. de Bem, and P. H. S. Torr, "3D hand shape and pose from images in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10835–10844.
- [5] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3D with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Mali, Jun. 2024, pp. 9826–9836.
- [6] R. Alexandros Potamias, J. Zhang, J. Deng, and S. Zafeiriou, "WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 12242–12254.
- [7] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.
- [8] V. Rudnev, V. Golyanik, J. Wang, H.-P. Seidel, F. Mueller, M. Elgharib, and C. Theobalt, "EventHands: Real-time neural 3D hand pose estimation from an event stream," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12365–12375.
- [9] J. Jiang, J. Li, B. Zhang, X. Deng, and B. Shi, "EvHandPose: Event-based 3D hand pose estimation with sparse supervision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 9, pp. 6416–6430, Sep. 2024.
- [10] C. Millerdurai, D. Luvizon, V. Rudnev, A. Jonas, J. Wang, C. Theobalt, and V. Golyanik, "3D pose estimation of two interacting hands from a monocular event camera," in *Proc. Int. Conf. 3D Vis. (3DV)*, Mar. 2024, pp. 291–301.
- [11] R. Liu, T. Ohkawa, T. Ho Elden Tse, M. Zhang, A. Yao, and Y. Sato, "Leveraging RGB images for pre-training of event-based hand pose estimation," 2025, *arXiv:2509.16949*.
- [12] R. Hara, W. Ikeda, M. Hatano, and M. Isogawa, "EventEgoHands: Event-based egocentric 3D hand mesh reconstruction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2025, pp. 1199–1204.
- [13] A. Bandini and J. Zariffa, "Analysis of the hands in egocentric vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6846–6866, Jun. 2023.
- [14] C. Plizzari, G. Goletto, A. Furnari, S. Bansal, F. Ragusa, G. Maria Farinella, D. Damen, and T. Tommasi, "An outlook to the future of egocentric vision," *Int. J. Comput. Vis.*, vol. 132, no. 11, pp. 4880–4936, Nov. 2024.
- [15] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Proc. CVPR*, Jun. 2011, pp. 3281–3288.
- [16] S. Lee, S. Bambach, D. J. Crandall, J. M. Franchak, and C. Yu, "This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 557–564.
- [17] C. Li and K. M. Kitani, "Model recommendation with virtual probes for egocentric hand detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2624–2631.
- [18] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3570–3577.
- [19] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1949–1957.

- [20] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 409–419.
- [21] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2O: Two hands manipulating objects for first person interaction recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10118–10128.
- [22] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12943–12954.
- [23] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, R. Newcombe, R. Wang, J. J. Engel, and T. Hodan, "HOT3D: Hand and object tracking in 3D from egocentric multi-view videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 7061–7071.
- [24] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–17, Dec. 2017.
- [25] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng, "End-to-end hand mesh recovery from a monocular RGB image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2354–2364.
- [26] S. Baek, K. I. Kim, and T.-K. Kim, "Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1067–1076.
- [27] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee, "HandOccNet: Occlusion-robust 3D hand mesh estimation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1486–1495.
- [28] X. Chen, Y. Liu, Y. Dong, X. Zhang, C. Ma, Y. Xiong, Y. Zhang, and X. Guo, "MobRecon: Mobile-friendly hand mesh reconstruction from monocular image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20512–20522.
- [29] D. Kulon, H. Wang, R. A. Güler, M. M. Bronstein, and S. Zafeiriou, "Single image 3D hand reconstruction with mesh convolutions," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019, p. 45.
- [30] D. Kulon, R. A. Güler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou, "Weakly-supervised mesh-convolutional hand reconstruction in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4989–4999.
- [31] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.
- [32] A. T. Aboukhadra, J. Malik, A. Elhayek, N. Robertini, and D. Stricker, "THOR-Net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Mali, Jan. 2023, pp. 1001–1010.
- [33] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1954–1963.
- [34] K. Lin, L. Wang, and Z. Liu, "Mesh graphormer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12919–12928.
- [35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–22.
- [36] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang, C. C. Loy, L. Yang, and Z. Liu, "SMPLer-X: Scaling up expressive human pose and shape estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 11454–11468.
- [37] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4D: Reconstructing and tracking humans with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 14737–14748.
- [38] Y. Ye, Y. Feng, O. Taheri, H. Feng, S. Tulsiani, and M. J. Black, "Predicting 4D hand trajectory from monocular videos," in *Proc. Int. Conf. 3D Vis. (3DV)*, Mar. 2026, pp. 860–870.
- [39] J. Zhang, J. Deng, C. Ma, and R. Alexandros Potamias, "HaWoR: World-space hand motion reconstruction from egocentric videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Cleveland, OH, USA, Jun. 2025, pp. 1805–1815.
- [40] Z. Yu, S. Zafeiriou, and T. Birdal, "Dyn-HaMR: Recovering 4D interacting hand motion from a dynamic camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 27716–27726.
- [41] Z. Xiao and X. Wang, "Event-based video super-resolution via state space models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 12564–12574.
- [42] T. Yura, A. Mirzaei, and I. Giltchenski, "EventSplat: 3D Gaussian splatting from moving event cameras for real-time rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 26876–26886.
- [43] R. Maeda, Y. Moon, and S.-H. Baek, "Event ellipsometer: Event-based mueller-matrix video imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 21804–21813.
- [44] C. Geckeler, N. Neugebauer, M. Muglikar, D. Scaramuzza, and S. Mintchev, "Event spectroscopy: Event-based multispectral and depth sensing using structured light," *IEEE Robot. Autom. Lett.*, vol. 11, no. 3, pp. 2658–2665, Mar. 2026.
- [45] A. Kar, V. Raj, and G.-M. Su, "Event camera guided visual media restoration & 3D reconstruction: A survey," 2025, *arXiv:2509.09971*.
- [46] V. Rudnev, M. Elgharib, C. Theobalt, and V. Golyanik, "EventNeRF: Neural radiance fields from a single colour event camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4992–5002.
- [47] T. Nakabayashi, N. Kairanda, H. Saito, and V. Golyanik, "Ev4DGS: Novel-view rendering of non-rigid objects from monocular event streams," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2025, pp. 1–15.
- [48] H. Deguchi, M. Masuda, T. Nakabayashi, and H. Saito, "E2GS: Event enhanced Gaussian splatting," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2024, pp. 1676–1682.
- [49] W. Yu, C. Feng, J. Li, J. Tang, J. Yang, Z. Tang, M. Cao, X. Jia, Y. Yang, L. Yuan, and Y. Tian, "EvaGaussians: Event stream assisted Gaussian splatting from blurry images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 24780–24790.
- [50] Y. Xue, H. Li, S. Leutenegger, and J. Stieckler, "Event-based non-rigid reconstruction from contours," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2022, pp. 1–22.
- [51] S. Ghosh and G. Gallego, "Event-based stereo depth estimation: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 10, pp. 9130–9149, Oct. 2025.
- [52] K. Horikawa, M. Isogawa, H. Saito, and S. Mori, "Dense depth from event focal stack," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 4545–4553.
- [53] W. Ikeda, M. Hatano, R. Hara, and M. Isogawa, "Event-based egocentric human pose estimation in dynamic environment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2025, pp. 2336–2341.
- [54] C. Millerdurai, H. Akada, J. Wang, D. Luvizon, C. Theobalt, and V. Golyanik, "EventEgo3D: 3D human motion capture from egocentric event streams," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1186–1195.
- [55] C. Millerdurai, H. Akada, J. Wang, D. Luvizon, A. Pagani, D. Stricker, C. Theobalt, and V. Golyanik, "EventEgo3D++: 3D human motion capture from a head-mounted event camera," *Int. J. Comput. Vis.*, vol. 133, no. 9, pp. 6432–6455, Sep. 2025.
- [56] R. Hori, M. Isogawa, D. Mikami, and H. Saito, "EventPointMesh: Human mesh recovery solely from event point clouds," *IEEE Trans. Vis. Comput. Graphics*, vol. 31, no. 9, pp. 5593–5610, Sep. 2025.
- [57] Z. Xu, G. Lu, C. Gao, and Q. Chen, "EvHand-FPV: Efficient event-based 3D hand tracking from first-person view," 2025, *arXiv:2509.13883*.
- [58] J. Park, G. Moon, W. Xu, E. Kaseman, T. Shiratori, and K. M. Lee, "3D hand sequence recovery from real blurry images and event stream," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 343–359.
- [59] J. Jiang, X. Zhou, B. Wang, X. Deng, C. Xu, and B. Shi, "Complementing event streams and RGB frames for hand mesh reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 24944–24954.
- [60] A. Prakash, R. Tu, M. Chang, and S. Gupta, "3D hand pose estimation in everyday egocentric images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 183–202.
- [61] T. Ohkawa, K. He, F. Sener, T. Hodan, L. Tran, and C. Keskin, "AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12999–13008.

- [62] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric RGB-D sensor," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1163–1172.
- [63] W. Mucha, M. Wray, and M. Kampel, "Towards egocentric 3D hand pose estimation in unseen domains," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2026, pp. 5776–5786.
- [64] F. Lin and T. Martinez, "Ego2HandsPose: A dataset for egocentric two-hand 3D global pose estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 4363–4371.
- [65] G. Jocher, J. Qiu, M. Liu, S. Lyu, F. C. Akyon, and M. E. Kalfaoglu, "Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models," Ultralytics Inc., Frederick, MD, USA, Tech. Rep., 2026, doi: [10.48550/arXiv.2606.03748](https://arxiv.org/abs/2606.03748). [Online]. Available: <https://arxiv.org/abs/2606.03748>
- [66] R. Papo, S. Gershov, T. Friedman, I. Or, G. Bolotin, and S. Laufer, "RoHan: Robust hand detection in operation room," 2025, *arXiv:2501.08115*.
- [67] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
- [68] W. Mucha and M. Kampel, "In my perspective, in my hands: Accurate egocentric 2D hand pose and action recognition," in *Proc. IEEE 18th Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2024, pp. 1–9.
- [69] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic DVS events," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1312–1321.
- [70] iniVation. (2019). *DAVIS346*. [Online]. Available: <https://inivation.com/wp-content/uploads/2019/08/DAVIS346.pdf>
- [71] AiQ Synterial Ltd. *16 IMUs MoCap Gloves*. [Online]. Available: <https://eu1.hubs.ly/H0bkfKm0>
- [72] OptiTrack. *Motive*, Sep. 15, 2026. [Online]. Available: <https://optitrack.com/software/motive>
- [73] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "HOnnotate: A method for 3D annotation of hand and object poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3193–3203.
- [74] S. Brahmabhatt, C. Tang, C. D. Twigg, Charles C. Kemp, and J. Hays, "ContactPose: A dataset of grasps with object contact and hand pose," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 361–378.
- [75] N. Carion et al., "SAM 3: Segment anything with concepts," 2025, *arXiv:2511.16719*.
- [76] Diederik P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–15.
- [77] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–19.
- [78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [79] G. Jocher and J. Qiu, "Ultralytics YOLO11," 2024. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [80] I. Robinson, P. Robicheaux, M. Popov, D. Ramanan, and N. Peri, "RF-DETR: Neural architecture search for real-time detection transformers," 2025, *arXiv:2511.09554*.
- [81] Z. Lin, C. Ding, H. Yao, Z. Kuang, and S. Huang, "Harmonious feature learning for interactive hand-object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12989–12998.



RYOSEI HARA (Graduate Student Member, IEEE) received the B.S.Eng. degree in information and computer science from Keio University, Japan, in 2025, where he is currently pursuing the M.S.Eng. degree in science and technology. His research interests include event-based vision, egocentric vision, and 3D hand mesh reconstruction.



WATARU IKEDA (Graduate Student Member, IEEE) received the M.S. degree in information and computer science from Keio University, Japan, in 2025, where he is currently pursuing the Ph.D. degree in science and technology. His research interests include human motion capture and egocentric vision.



MASASHI HATANO (Graduate Student Member, IEEE) received the M.S. degree in computer science from Keio University, Japan, and the Diplôme d'ingénieur from CentraleSupélec, France, in 2023. He is currently pursuing the Ph.D. degree in computer science with Keio University. His research interests include egocentric vision and human motion understanding and generation.



MARIKO ISOGAWA (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Osaka University, Japan, in 2011, 2013, and 2019, respectively. She was a Visiting Scholar at Carnegie Mellon University, USA, from 2019 to 2020. She is currently an Associate Professor with the Department of Information and Computer Science, Faculty of Science and Technology, Keio University, Japan. Her research interests include computer vision and pattern recognition.

...