EVENTEGOHANDS: EVENT-BASED EGOCENTRIC 3D HAND MESH RECONSTRUCTION

Supplementary Material

Contents

A Overview of the Supplementary Material 1

B Hand Segmentation Module 1
- B.1 Locally-Normalised Event Surfaces (LNES) 1
- B.2 Event Cloud 1
- B.3 Qualitative Results 2

C Analysis of Computational Cost 2

D Video Qualitative Evaluation 2

E Ablation Analysis 2

F Failure Case 3

G References 3

A. OVERVIEW OF THE SUPPLEMENTARY MATERIAL

This supplementary document contains additional details and discussions of our EventEgoHands. Please also refer to the supplementary video for video qualitative evaluation. We highlight reference numbers assosiated with the main paper in blue, and those associated with this supplementary document in red.

B. HAND SEGMENTATION MODULE

B.1. Locally-Normalised Event Surfaces (LNES)

Locally-Normalised Event Surfaces (LNES) [1] is one of the frame-based event representations. The LNES preserves temporal information within a time window by applying temporal weights. The LNES representation $\boldsymbol{I}$ is expressed by the following equation:

$$\boldsymbol{I}(x_i, y_i, p_i) = \frac{t_i - t_s}{t_l - t_s}, \tag{1}$$

where x_i, y_i are the pixel coordinates, p_i represents the polarity, and t_i is the timestamp of the i-th event point. Here, t_s is the timestamp of the first event, and t_l is the timestamp of the last event within a time window.

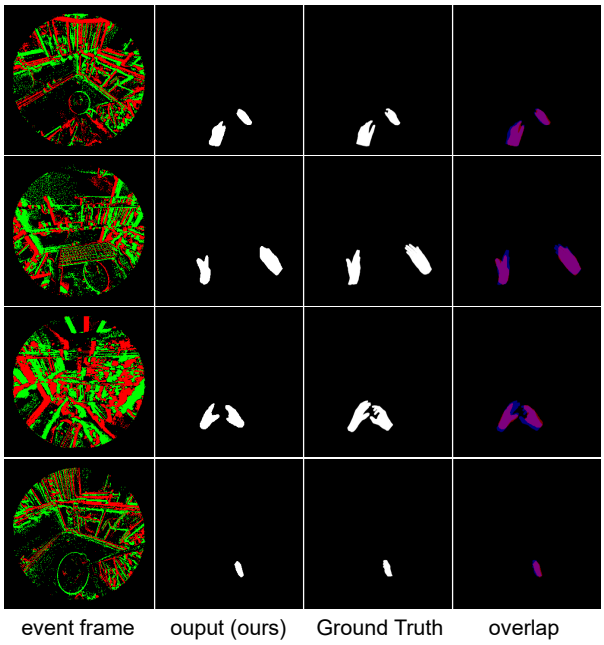
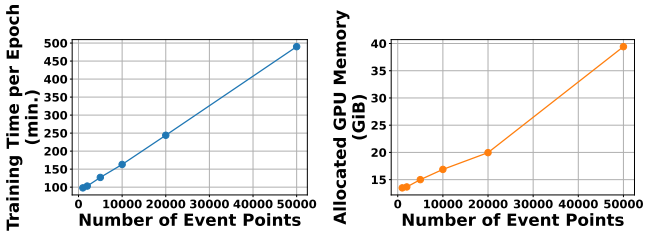


Fig. 1: **Hand segmentation results.** We show qualitative results of our hand segmentation. In the overlap column, the mask in red and blue represents our predicted hand mask and ground truth, respectively. The overlap region between two is shown in purple.

We adopt LNES as the event-frame representation for the Hand Segmentation Module. Our method takes $\boldsymbol{I}^{1:T}$ as input, which is a continuous LNES $\boldsymbol{I}$ within a fixed-width time window T. When the camera wearer's movement is slight, the number of triggered events decreases, which can result in a decline in hand estimation accuracy. We can address situations where fewer events occur by incorporating multiple time steps as input. In our proposed method, $T = 3$ is used, which corresponds to a very short interval of 3 frames at 30 fps. Since the mask does not change significantly within a short time interval, the mask at the latest timestamp is used as the filtering mask.

B.2. Event Cloud

The Event Cloud [2] is one of the point-cloud-based event representations. A single event point is represented as fol-



(a) Training time cost (b) GPU memory consumption

Fig. 2: **Analysis of computational cost w.r.t the number of event points**.

lows:

$$\boldsymbol{E_k} = (x_k, y_k, t_k, p_k, n_k), \tag{2}$$

where x_k, y_k are the pixel coordinates, t_k is the timestamp, and p_k, n_k are the positive and negative polarity of the k-th event point. The Event Cloud $\boldsymbol{E} \in \mathbb{R}^{N\times 5}$ is a point cloud consisting of N event points $\boldsymbol{E_k}$. The total number of event points is $N = 2048$.

B.3. Qualitative Results

The results of the mask region estimated by the Hand Segmentation Module are shown in Fig. 1. The results indicate that the Hand Segmentation Module can estimate the hand region from LNES even when only one hand is visible.

C. ANALYSIS OF COMPUTATIONAL COST

To investigate how the Hand Segmentation Module contributes to reducing computational cost by decreasing the number of events, we conducted an analysis of computational cost. We analyze the training time per epoch (in minutes) and GPU memory usage (in gibibytes) while varying the number of event points. Most of the N-HOT3D datasets we created contain between a few thousand (K) and 100,000 (100K) event points within a single time window at 30 fps. The number of events in our proposed method is $N = 2048$. Therefore, we can reduce the number of events by approximately 1/50 at most. As shown in Fig. 2, training time and GPU memory usage increase as the number of events increases. Therefore, reducing the number of events helps to lower the computational cost.

D. VIDEO QUALITATIVE EVALUATION

We include the supplementary video showcasing the 3D hand mesh reconstruction performance of EventEgoHands. The video is also available on the page where this document can be found. The video contains the qualitative evaluation results, including comparisons with the baseline methods and the ablation study. The video demonstrates that our method produces the most stable and closest output to the ground truth. The EventEgoHands is designed for single time window output without considerations for temporal coherence, which may result in jittery outputs when applied to video evaluation.

Table 1: **Balancing Hyperparameters for Hand Segmentation Module**. The yellow row indicates the value we adopted.

$\lambda_\alpha : \lambda_\beta$	IoU (↑)
0.8 : 0.2	0.529
0.7 : 0.3	**0.567**
0.6 : 0.4	0.519

Table 2: **Balancing Hyperparameters for Hand Reconstruction Module**. The yellow row indicates the value we adopted.

λ	R-AUC (↑)	MPJPE [mm] (↓)	MPVPE [mm] (↓)
$\lambda_\gamma = 1$	0.440	60.89	43.78
$\lambda_\gamma = 0.1$	**0.450**	**59.51**	**42.92**
$\lambda_\gamma = 0.01$	0.419	63.26	44.71
$\lambda_\delta = 10$	0.440	60.97	43.83
$\lambda_\delta = 1$	**0.450**	**59.51**	**42.92**
$\lambda_\delta = 0.1$	0.446	59.94	43.16
$\lambda_\epsilon = 10$	0.444	60.24	43.53
$\lambda_\epsilon = 1$	**0.450**	**59.51**	**42.92**
$\lambda_\epsilon = 0.1$	0.441	61.12	43.98
$\lambda_\zeta = 30$	0.383	67.92	47.68
$\lambda_\zeta = 20$	**0.450**	**59.51**	**42.92**
$\lambda_\zeta = 10$	0.441	60.93	43.60

E. ABLATION ANALYSIS

We conduct an ablation study to analyze the impact of balancing hyperparameters for losses in Hand Segmentation Module and Hand Reconstruction Module. We use predefined sets of loss weights for Hand Segmentation Module: $(\lambda_\alpha, \lambda_\beta) \in \{(0.8, 0.2), (0.7, 0.3), (0.6, 0.4)\}$. Regarding Hand Reconstruction Module, we systematically vary the values of each loss weight using the following ranges: λ_γ from 0.01 to 1, λ_δ from 0.1 to 10, λ_ϵ from 0.1 to 10, and λ_ζ from 10 to 30. We vary one loss weight at a time while keeping the others fixed at their default value: 0.1, 1, 1, and 20 for λ_γ, λ_δ, λ_ϵ, and λ_ζ, respectively. We use the Intersection over Union (IoU) as a metric for hand segmentation, which is commonly employed in segmentation tasks, to assess the overlap between the predicted mask and the ground truth mask. R-AUC, MPJPE, and MPVPE are adopted as metrics of hand reconstruction. Tab. 1 and Tab. 2 show comparisons of hyperparameters for the Hand Segmentation and Hand Reconstruction Module, respectively.

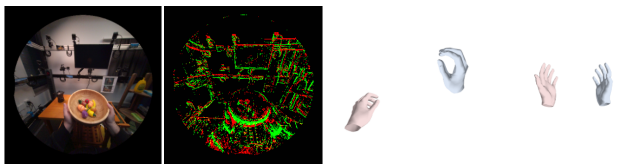


Fig. 3: **Example of failure case.**

F. FAILURE CASE

Although the proposed method demonstrates improvements over existing methods, there is a challenge when interacting with objects. Fig. 3 illustrates a failure case where, like many other 3D hand mesh reconstruction methods, our method struggles when an object occludes the hand. Since our method does not consider hand-object interactions, it will be necessary to incorporate this aspect in future work.

G. REFERENCES

[1] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, and Christian Theobalt, "EventHands: real-time neural 3D hand pose estimation from an event stream," in *ICCV*, 2021, pp. 12385–12395.

[2] Christen Millerdurai, Diogo Luvizon, Viktor Rudnev, André Jonas, Jiayi Wang, Christian Theobalt, and Vladislav Golyanik, "3D pose estimation of two interacting hands from a monocular event camera," in *3DV*, 2024, pp. 291–301.