Leveraging RGB Images for Pre-Training of Event-Based Hand Pose Estimation

Ruicong Liu [1, 2] Takehiko Ohkawa[1] Tze Ho Elden Tse[2] Mingfang Zhang[1]
Angela Yao[2] Yoichi Sato[1]
[1] The University of Tokyo, Japan [2] National University of Singapore, Singapore

Abstract

This paper presents RPEP, the first pre-training method for event-based 3D hand pose estimation using labeled RGB images and unpaired, unlabeled event data. Event data offer significant benefits such as high temporal resolution and low latency, but their application to hand pose estimation is still limited by the scarcity of labeled training data. To address this, we repurpose real RGB datasets to train event-based estimators. This is done by constructing pseudo-event-RGB pairs, where event data is generated and aligned with the ground-truth poses of RGB images. Unfortunately, existing pseudo-event generation techniques assume stationary objects, thus struggling to handle non-stationary, dynamically moving hands. To overcome this, RPEP introduces a novel generation strategy that decomposes hand movements into smaller, step-by-step motions. This decomposition allows our method to capture temporal changes in articulation, constructing more realistic event data for a moving hand. Additionally, RPEP imposes a motion reversal constraint, regularizing event generation using reversed motion. Extensive experiments show that our pre-trained model significantly outperforms state-of-the-art methods on real event data, achieving up to 24% improvement on EvRealHands. Moreover, it delivers strong performance with minimal labeled samples for fine-tuning, making it well-suited for practical deployment.

1. Introduction

Capturing 3D hands from RGB images has been studied extensively [2, 5, 13, 15–17, 23, 24], but it remains vulnerable under challenging conditions such as wide brightness variation, complex lighting, and fast hand motion [7, 14]. Event cameras [8] are an alternative to RGB cameras that asynchronously capture per-pixel intensity changes. They can record high dynamic range images with an ultra-high frame rate (up to 1 µs latency) [6, 9, 18–20, 25].

Despite their advantages, datasets with real event data for 3D hand pose estimation are scarce due to the difficulty of annotation. Unlike RGB images, event data lacks texture information, making it difficult to annotate accurate 3D hand poses. The EvRealHand dataset [6] is the first and only large-scale dataset to provide annotated real event data of hands. Although it offers event streams with 3D hand annotations, it heavily relies on a complex rig equipped with synchronized multi-camera RGB and event sensors. Such reliance on a studio-style capture rig limits its diversity and authenticity.

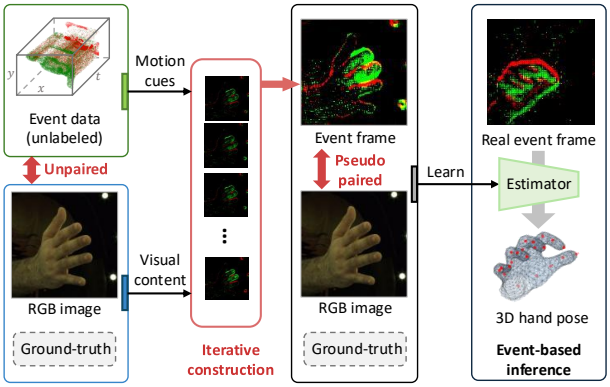

Figure 1. We propose a pre-training method for event-based hand pose estimation using labeled RGB images and unpaired, unlabeled event data. At the core of our approach is an iterative construction module that generates a pseudo-event frame for each input RGB image, forming pseudo RGB-event pairs that reflect dynamic hand movements.

The lack of evnt datasets motivates our approach called **RPEP** (Fig. 1): leveraging **R**GB images for **P**re-training of **E**vent-based hand **P**ose estimation, helping to reduce the dependence on event annotations. To learn event-based models from RGB images, previous studies [11, 21] construct pseudo-paired RGB–event data. In this process, the generated pseudo-event frames are aligned with the ground-truth pose of the RGB image. This alignment allows using the ground-truth hand pose annotations of the RGB images

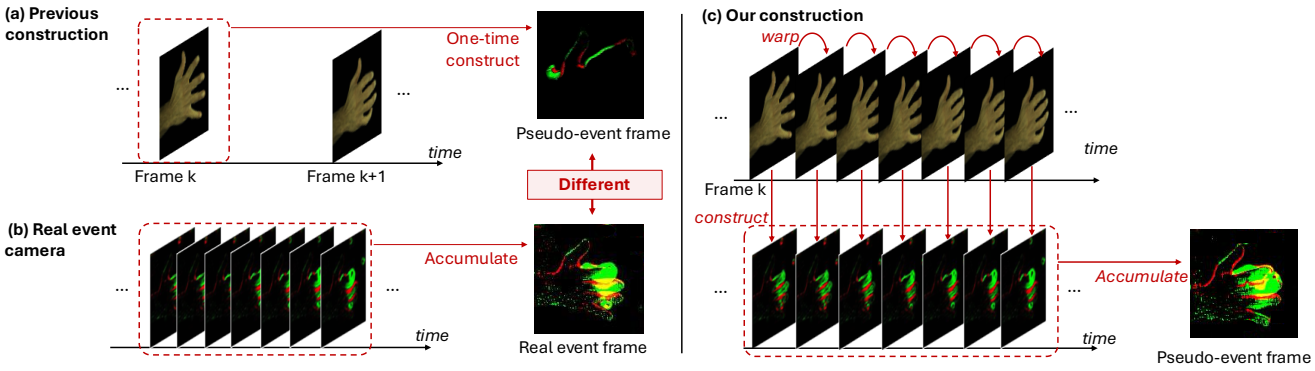

Figure 2. Comparison of (a) RGB-event pair construction of previous methods [4, 11], (b) real event camera's capture process, and (c) our construction method for hands.

to train event-based estimators. However, existing algorithms for constructing RGB–event pairs [4, 11] are not well suited for hand data, due to their poor reconstruction quality. As shown in Fig. 2 (a), previous methods generate event frames that are sparsely distributed along image edges. In contrast, the real event frame in Fig. 2 (b) exhibits a denser distribution in the finger regions, where finger articulations naturally trigger events. Such discrepancy in event distribution brings a huge domain gap between real and constructed event data, which hinders the learning process and degrades the model's generalization to real environment.

Such poor construction quality stems from the stationary assumption underlying existing algorithms. As shown in Fig. 2 (a), prior methods [4, 11] generate pseudo-events only once from a single RGB frame. This one-time construction implicitly assumes that the hand undergoes only rigid pose changes (*i.e.*, translation and rotation) between frame k and frame $k+1$. It thereby ignores non-rigid articulations, which will induce additional events within this interval. Consequently, the constructed events appear only along static image edges. In contrast, due to the high temporal resolution of event cameras, real sensors capture much denser articulations that occur from frame k to $k+1$, as illustrated in Fig. 2 (b). These events are distributed across articulation regions such as moving fingers and the palm, covering all areas of motion.

To construct more realistic pseudo-event data, we reformulate the original one-time construction into a process that simulates the real event accumulation process, as shown in Fig. 2 (c). From a single frame k, our method generates multiple intermediate RGB frames by warping, simulating the image changes caused by articulations. For each RGB frame, we then perform a one-time construction [4, 11] to generate corresponding pseudo-event frame. Over time, the final pseudo-event frame is generated by accumulating all previous pseudo-event frames. Unlike prior methods [4, 11], our construction process allows for articulations of the hand, thereby facilitating a more authentic simulation of the event accumulation.

To achieve the above process, we implement an iterative construction module, which separates the event accumulation into T iterations. At each iteration, this module generates an optical flow map to both 1) warp the RGB image and 2) construct pseudo-event frame. As the process progresses, the image evolves, leading to the construction of event data that captures dynamic articulations. To estimate the optical flow map, we use the RGB image and an unpaired, unlabeled event frame as input. In detail, we extract 1) the *visual appearance feature* of the RGB input and 2) the *motion priors*, such as direction and trajectory, from the event input. The motion priors provide necessary movement information, thereby enabling the static RGB hand to "articulate". The appearance feature and motion priors are then fed into a decoder to estimate the flow map.

Additionally, RPEP imposes a novel motion reversal constraint, ensuring the the semantic correctness of the motion priors. In other words, it ensures that the motion prior truly represent information such as physical moving direction and trajectory. Starting from a constructed pseudo-event frame, we create a reversal frame with a completely opposite motion direction and trajectory. Our method then maximizes the difference between the two motion priors from pseudo-event and reversal frames. This constraint ensures the consistency between the extracted motion priors and the physical motion dynamics.

We evaluate our method across multiple challenging scenarios (*e.g.*, flash and strong light) using the EvRealHands dataset [6] as the evaluation target. Extensive experimental results demonstrate the superior performance of our method compared to existing transfer learning [3, 11, 22, 26] and pre-training [1] methods. In comparison to these methods, RPEP exhibits a relative improvement rate of 24% in normal scenes, 20% in strong light scenes, and 14% in flash light scenes.

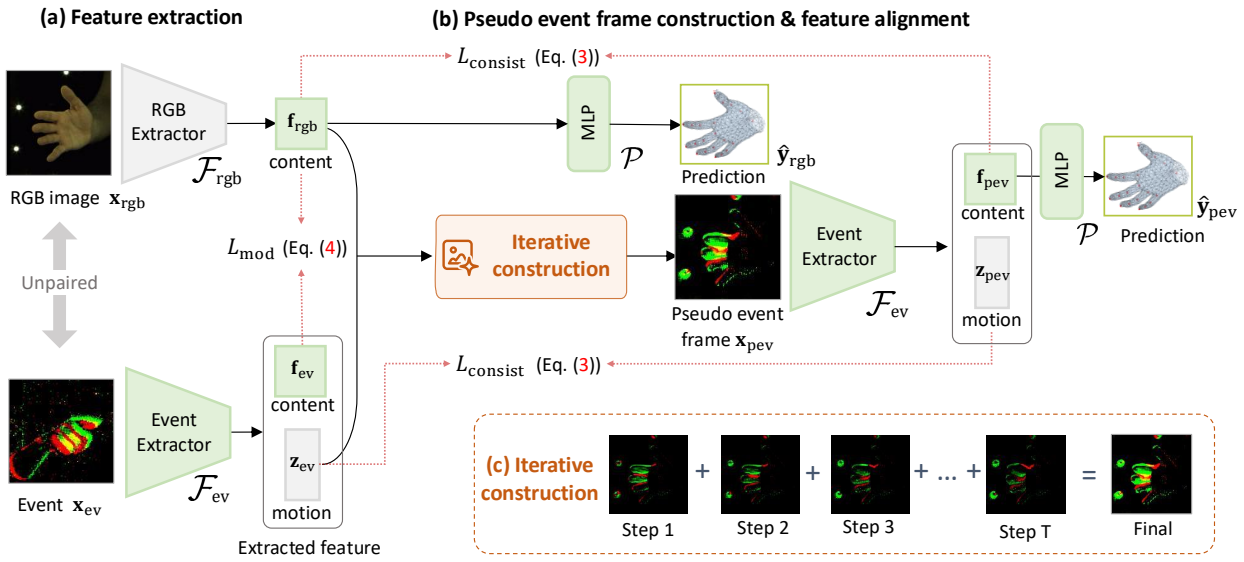

Figure 3. Overview of the proposed method, where labeled RGB images and unpaired, unlabeled event frames serve as inputs for training. (a) Feature extractors, $\mathcal{F}_{rgb}$ and $\mathcal{F}_{ev}$, extract features from their respective inputs. (b) Our iterative construction module constructs pseudo-event frames from these features, which are then fed into $\mathcal{F}_{ev}$ for feature alignment. The features $\mathbf{f}_{rgb}$, $\mathbf{f}_{ev}$, and $\mathbf{f}_{pev}$ are all aligned to the same feature space. (c) Illustration of the iterative construction process. Event frames from all iterations are accumulated together to get the final pseudo-event frame.

2. Proposed Method

We introduce RPEP, a novel pre-training method for event-based hand pose estimation leveraging RGB images. Our goal is to learn an event-based hand pose estimator, $\mathcal{H} = \mathcal{F}_{ev} \circ \mathcal{P}$, which comprises a feature extractor $\mathcal{F}_{ev}$ and a multi-layer perceptron (MLP) $\mathcal{P}$.

As illustrated by Fig. 3 (a), our method employs two feature extractors: an RGB extractor $\mathcal{F}_{rgb}$ and an event extractor $\mathcal{F}_{ev}$. They process unpaired inputs, an RGB image $\mathbf{x}_{rgb} \in \mathbb{R}^{H\times W\times 3}$ and event frame $\mathbf{x}_{ev} \in \mathbb{R}^{H\times W\times 2}$ (event histogram [10]), respectively. We design $\mathcal{F}_{ev}$ to capture the following two features from $\mathbf{x}_{ev}$. **1) The appearance feature $\mathbf{f}_{ev}$ represents the hand pose, 2) and the motion priors $\mathbf{z}_{ev}$ represents the moving direction and trace.** The RGB extractor $\mathcal{F}_{rgb}$ solely extracts the appearance feature $\mathbf{f}_{rgb}$ from $\mathbf{x}_{rgb}$, since the hand image is stationary and contains no motion information. The $\mathbf{x}_{rgb}$ and $\mathbf{z}_{ev}$ are combined and fed into our iterative construction module to construct the pseudo-event frame $\mathbf{x}_{pev}$.

Feature alignment. To enable effective knowledge transfer and representation sharing across modalities, the MLP $\mathcal{P}$ is shared between both RGB and event. This encourages the extracted features $\mathbf{f}_{ev}$ and $\mathbf{f}_{rgb}$ to be aligned in the same latent space. We further employ adversarial learning to explicitly align $\mathbf{f}_{ev}$ and $\mathbf{f}_{rgb}$. In addition, we align features between the original input data and the constructed pseudo-event frames, *i.e.*, $\mathbf{f}_{rgb}$ with $\mathbf{f}_{pev}$, and $\mathbf{z}_{ev}$ with $\mathbf{z}_{pev}$.

Iterative construction. To simulate to process of Fig. 2 (c), we separate the time window $\Delta\tau$ into multiple iterations and develop an iterative construction module. In each iteration t, we use a decoder $\mathcal{G}$ to generate an optical flow map $\hat{\mathbf{v}}^{(t)}$. Its input contains the appearance feature of the input RGB image and the motion prior of the input event frame, *i.e.*, $\hat{\mathbf{v}} = \mathcal{G}(\mathbf{f}_{rgb}, \mathbf{z}_{ev})$. The estimated flow map $\hat{\mathbf{v}}^{(t)}$ is used for two purposes: **1) generating a sub-pseudo event frame $\mathbf{x}_{pev}^{(t)}$**, and **2) warping the RGB image** to reflect the dynamic image changes caused by articulation. After T iterations, all event frames are accumulated to form the final pseudo-event frame $\mathbf{x}_{pev}$:

$$\mathbf{x}_{pev} = \sum_{t=1}^{T} \mathbf{x}_{pev}^{(t)} = \lfloor \sum_{t=1}^{T} \nabla \mathbf{x}_{rgb}^{(t-1)} \cdot \hat{\mathbf{v}}^{(t)} \rfloor. \quad (1)$$

This way, our construction process allows for variations in the RGB image, thereby facilitating a more authentic simulation of the event accumulation process. Empirically, we set $T = 6$. Fig. 4 displays the constructed event frame, flow map, and warped RGB image of each iteration.

Motion reversal constraint To ensure that the motion priors, $\mathbf{z}_{ev}$ and $\mathbf{z}_{pev}$, truly reflect actual physical dynamics, we introduce a motion reversal constraint. Our method reverses the flow map $\hat{\mathbf{v}}^{(t)}$ across iterations from 1 to T, generating reverse events $\mathbf{x'}_{pev}^{(t)}$ as:

$$\mathbf{x'}_{pev} = \sum_{t=1}^{T} \mathbf{x'}_{pev}^{(t)} = \lfloor \sum_{t=1}^{T} \nabla \mathbf{x}_{rgb}^{(t+1)} \cdot -\hat{\mathbf{v}}^{(t)} \rfloor. \quad (2)$$

Here, we reverse the direction of motion using $-\hat{\mathbf{v}}^{(t)}$ and

Table 1. Comparison with state-of-the-art methods. The evaluation set EvRealHands [6] is separated according to collection scenarios: "Normal", "Strong light", and "Flash". Within each scenario, samples are further divided into "Scripted" and "Unscripted" hand poses.

	Method	Metrics	Normal		Strong light		Flash	
			Scripted	Unscripted	Scripted	Uncripted	Scripted	Uncripted
Baseline	w/o pre-train	3D-MPJPE	27.98	49.12	31.95	52.94	36.20	51.53
		PA-MPJPE	13.38	17.95	16.96	19.36	15.04	22.29
	Synthetic [18]	3D-MPJPE	37.04	56.77	37.76	56.73	37.33	57.70
		PA-MPJPE	18.78	20.09	20.47	20.60	17.42	25.76
Main results	SimCLR [1]	3D-MPJPE	32.72	52.49	30.94	53.90	41.28	58.04
		PA-MPJPE	13.92	17.14	16.41	17.88	15.40	22.95
	Vid2E [3]	3D-MPJPE	25.92	44.59	26.83	47.47	33.81	51.39
		PA-MPJPE	13.32	16.89	15.17	18.94	14.72	20.80
	CycleGAN [26]	3D-MPJPE	24.32	46.78	28.57	50.65	31.27	51.03
		PA-MPJPE	13.13	16.34	15.95	18.08	15.08	21.98
	ADDA [22]	3D-MPJPE	25.79	45.80	31.06	51.02	34.25	49.17
		PA-MPJPE	13.32	16.35	16.41	18.28	15.27	22.09
	RPG-EV [11]	3D-MPJPE	29.51	47.81	34.34	53.19	34.15	51.59
		PA-MPJPE	15.51	16.89	18.05	18.58	16.56	22.61
	RPEP (Ours)	3D-MPJPE	**21.26**	**39.91**	**26.08**	**44.28**	**30.97**	**48.55**
		PA-MPJPE	**12.11**	**15.45**	**14.47**	**17.85**	**14.06**	**20.11**

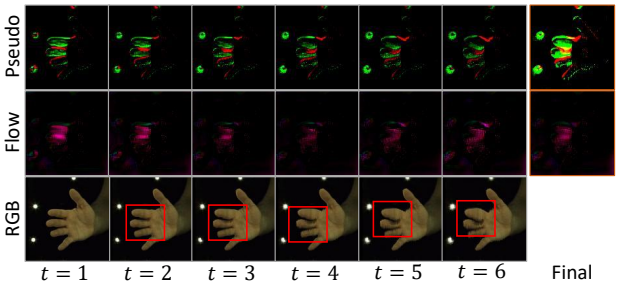

Figure 4. Constructed event frame, flow map, and RGB image from each iteration, with image changes caused by articulation highlighted in red rectangles.

invert the trace by employing $\mathbf{x}'^{(t+1)}_{\text{pev}}$ from the subsequent iteration $t+1$.

Given that the movement direction and trace of $\mathbf{x}'_{\text{pev}}$ are opposite to those of $\mathbf{x}_{\text{pev}}$, their motion features, $\mathbf{z}'_{\text{pev}}$ and $\mathbf{z}_{\text{pev}}$, should exhibit high divergence. Therefore, we propose a divergence loss $\mathcal{L}_{\text{div}}$ to minimize their cosine similarity.

3. Experiment

In Tab. 1, we compare our method with state-of-the-art approaches. In all experiments, we first pre-trains an event-based pose estimator from labeled RGB data from InterHand2.6M [12] and unlabeled event data from EvRealHands[6]. Then, we fine-tune the pre-trained estimator using a few labeled samples from the EvRealHands.

The compared methods include SimCLR [1], an unsupervised pre-training approach; Vid2E [3] and CycleGAN [26], which convert RGB videos into event representations; ADDA [22], a domain adaptation technique; and RPG-EV [11], a transfer learning method that utilizes labeled RGB data along with unlabeled event data to train event-based networks. For completeness, we also include two baselines: 1) fine-tuning a randomly initialized estimator without any pre-training, and 2) pre-training on a synthetic event dataset, EventHands [18].

Tab. 1 presents the results after fine-tuning, showing that our method consistently outperforms all state-of-the-art approaches under various lighting conditions, achieving the lowest MPJPE error in every case. Interestingly, the "Synthetic" setup performs worse than even the "w/o pre-train" baseline, suggesting the huge domain gap between synthetic and real event data. Most other RGB-pre-training methods outperform the baselines, highlighting the benefit of using real RGB data for pre-training. Notably, our method surpasses RPG-EV by a significant margin—achieving an 8mm lower error—demonstrating the effectiveness of our iterative pseudo-event construction in improving estimation performance.

4. Conclusion

In this paper, we introduce RPEP, a novel pre-training method for event-based 3D hand pose estimation leveraging RGB images. The core innovation of RPEP is its iterative construction module, which generates pseudo-event frames that effectively accommodate dynamic hand motions. Additionally, our method incorporates a motion reversal constraint to refine the extracted motion priors, leading to enhanced construction results. Evaluation results demonstrate that RPEP outperforms state-of-the-art techniques, achieving significant performance gains across a range of challenging scenarios.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 2, 4

[2] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, Fei Li, Zheng Liu, Feng Lu, Karim Abou Zeid, Bastian Leibe, Jeongwan On, Seungryul Baek, Aditya Prakash, Saurabh Gupta, Kun He, Yoichi Sato, Otmar Hilliges, Hyung Jin Chang, and Angela Yao. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 428–448, 2024. 1

[3] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 2, 4

[4] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Eklt: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, 128(3):601–618, 2020. 2

[5] Markus Höll, Markus Oberweger, Clemens Arth, and Vincent Lepetit. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. In *2018 IEEE conference on virtual reality and 3D user interfaces (VR)*, pages 175–182. IEEE, 2018. 1

[6] Jianping Jiang, Jiahe Li, Baowen Zhang, Xiaoming Deng, and Boxin Shi. Evhandpose: Event-based 3d hand pose estimation with sparse supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 4

[7] Jianping Jiang, Xinyu Zhou, Bingxuan Wang, Xiaoming Deng, Chao Xu, and Boxin Shi. Complementing event streams and rgb frames for hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24944–24954, 2024. 1

[8] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μ s latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43 (2):566–576, 2008. 1

[9] Xianlei Long, Xiaxin Zhu, Fangming Guo, Chao Chen, Xiangwei Zhu, Fuqiang Gu, Songyu Yuan, and Chunlong Zhang. Spike-brgnet: Efficient and accurate event-based semantic segmentation with boundary region-guided spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

[10] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5419–5427, 2018. 3

[11] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 1, 2, 4

[12] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 4

[13] Takehiko Ohkawa, Ryosuke Furuta, and Yoichi Sato. Efficient annotation and learning for 3D hand pose estimation: A survey. *International Journal on Computer Vision (IJCV)*, 131:3193—3206, 2023. 1

[14] Joonkyu Park, Gyeongsik Moon, Weipeng Xu, Evan Kaseman, Takaaki Shiratori, and Kyoung Mu Lee. 3d hand sequence recovery from real blurry images and event stream. In *European Conference on Computer Vision*, pages 343–359. Springer, 2024. 1

[15] Vladimir I Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):677–695, 1997. 1

[16] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. User-defined gestures for augmented reality. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 955–960. 2013.

[17] Siddharth S Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review*, 43:1–54, 2015. 1

[18] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12385–12395, 2021. 1, 4

[19] Chenyang Shi, Yuzhen Li, Ningfang Song, Boyi Wei, Yibo Zhang, Wenzhuo Li, and Jing Jin. Identifying light interference in event-based vision. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4800–4816, 2023.

[20] Chenyang Shi, Boyi Wei, Xiucheng Wang, Hanxiao Liu, Yibo Zhang, Wenzhuo Li, Ningfang Song, and Jing Jin. Polarity-focused denoising for event cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

[21] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 1

[22] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2, 4

[23] Christian Von Hardenberg and François Bérard. Bare-hand human-computer interaction. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–8, 2001. 1

[24] Huilong Xie, Wenwei Song, and Wenxiong Kang. Learning an augmented rgb representation for dynamic hand gesture authentication. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

[25] Xu Zheng and Lin Wang. Eventdance: Unsupervised source-free cross-modal adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17448–17458, 2024. 1

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 4