Complementing Event Streams and RGB Frames for Hand Mesh Reconstruction

Jianping Jiang[†1,2,3], Xinyu Zhou[†4], Bingxuan Wang[1,2,3], Xiaoming Deng[‡5,6], Chao Xu[4], Boxin Shi[‡1,2,3]
[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3]AI Innovation Center, School of Computer Science, Peking University
[4]National Key Lab of General AI, School of Intelligence Science and Technology, Peking University
[5]Institute of Software, Chinese Academy of Sciences [6]University of Chinese Academy of Sciences

Abstract

Reliable hand mesh reconstruction (HMR) from commonly-used color and depth sensors is challenging especially under scenarios with varied illuminations and fast motions. Event camera is a highly promising alternative for its high dynamic range and dense temporal resolution properties, but it lacks salient texture appearance for hand mesh reconstruction. In this paper, we propose EvRGBHand – the first approach for 3D hand mesh reconstruction with an event camera and an RGB camera compensating for each other. By fusing two modalities of data across time, space, and information dimensions, EvRGBHand can tackle overexposure and motion blur issues in RGB-based HMR and foreground scarcity as well as background overflow issues in event-based HMR. We further propose EvRGBDegrader, which allows our model to generalize effectively in challenging scenes, even when trained solely on standard scenes, thus reducing data acquisition costs. Experiments on real-world data demonstrate that EvRGBHand can effectively solve the challenging issues when using either type of camera alone via retaining the merits of both, and shows the potential of generalization to outdoor scenes and another type of event camera. For code, models, and dataset, please refer to https://alanjiang98.github.io/evrgbhand.github.io/.

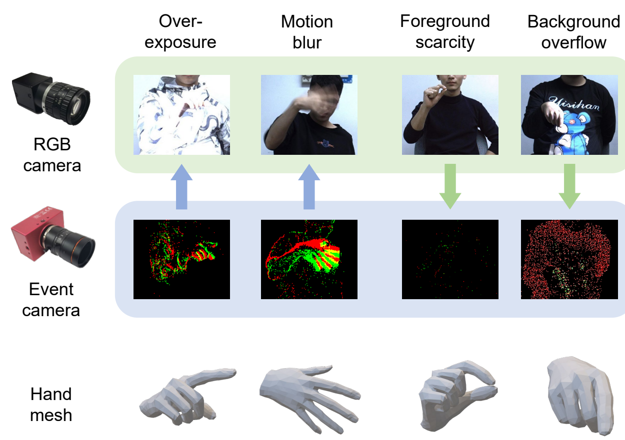


Figure 1. Due to the differences in RGB camera and event camera imaging mechanisms, it is promising to make complementary use of both modalities of data to achieve robust hand mesh reconstruction and tackle their respective challenging issues listed at the top. The arrows between the first and second rows point to the compensated data domain using the data from their tails.

1. Introduction

Reliable 3D hand mesh reconstruction (HMR) is essential for various applications in virtual reality and robotics. Although great progress on HMR has been made for color [4, 6, 25], depth [9, 10, 19, 32], and event cameras [37, 39], HMR based on a single sensor can not achieve satisfactory performance for different scenarios. The frame-based RGB or depth imaging mechanism inevitably faces degenerated issues, such as **overexposure** under strong light conditions and **motion blur** when hands move fast, which poses challenges to conducting robust HMR.

Recently, event cameras have shown great potential in HMR for high dynamic range (HDR) and fast motion scenes [39] thanks to their superior properties from neuromorphic imaging mechanism in dynamic range and temporal resolution. Being generated asynchronously by measuring per-pixel intensity changes, event streams [26] are incapable in preserving effective texture and edge information in two typical scenarios. First, events triggered from hands are rare when hands keep static (we call it "**foreground scarcity**" issue). Second, events triggered from the background are excessive when illumination significantly changes, which can heavily confuse the events from hand motion (we call it "**background overflow**" is-

[†] equal contribution, [‡] corresponding author

sue). We show these issues in Fig. 1, which motivate us to combine RGB frames and event streams to compensate for each other and improve the performance on their respective issues. Advantages from such fusion have been demonstrated on several vision tasks, such as feature tracking [34], super-resolution [14, 30, 50], and data association [58], but there has been no work specially designed for HMR yet.

Combining RGB frames and asynchronous event streams for HMR faces two challenges. First, event streams and RGB images differ in data format, space, temporal distribution, and visual information carried. It is still an open problem to conduct a fully adaptive multi-modal fusion strategy for HMR with images and events. Second, it is difficult to obtain high-quality 3D hand annotations, especially in challenging scenes (*e.g*. strong light, fast motion, flash at a large scale). Hence, *how to enable models to generalize well from limited training data in normal scenes to real-world challenging scenes* remains an open problem.

To tackle these challenges, we propose EvRGBHand – an transformer-based [47] framework for 3D HMR to make complementary benefits of event streams and RGB frames as shown in Fig. 2. We design EvImHandNet to bridge the gap in data distribution across two modalities by spatial alignment, complementary fusion, and temporal attention on event streams and RGB images. To effectively enhance the model's generalization capability, we further propose EvRGBDegrader, a data augmentation module for event and image pairs, enabling our model to be trained solely on normal scenes and yet significantly improve performance in challenging settings. To evaluate our method, we collect a real-world event-based hand dataset EVREALHANDS with 3D annotations and build a large-scale synthetic dataset to enlarge training data. Experiments on real-world data show that EvRGBHand can effectively tackle the challenging issues by compensating for each other and well balance between computational cost and accuracy, even with the most vanilla transformer-based fusion strategy [1, 20]. Preliminary qualitative analysis shows that EvRGBHand, once trained solely on indoor scenes captured by the DAVIS346 event camera [26], demonstrates cross-environment generalization to outdoor scenes and cross-camera adaptability to another type of event camera. The main contributions of this paper can be summarized as follows:

1. We investigate the feasibility of using events and images for HMR, and propose the first solution to 3D HMR by complementing event streams and RGB frames.
2. We introduce EvImHandNet, a novel approach for effectively fusing event streams and RGB images across spatial, temporal, and informational dimensions.
3. We propose EvRGBDegrader, a data augmentation method specifically designed for enhancing the generalization capability of models in challenging scenes for HMR with events and images.

2. Related work

2.1. RGB-based HMR

Prior works on 3D HMR can be divided into two categories: parametric and non-parametric methods [6]. Parametric methods [2, 3, 5, 60] estimate the parameters of a hand model such as MANO [38] while non-parametric methods [4, 23, 29, 61] directly regress the positions of the hand mesh vertices. Although parametric methods involve the hand shape prior into the approaches, they ignore spatial correlations [27] and regressing 3D rotations is a challenging task [31]. Recent network architectures such as graph convolutional neural network (GCN) [22] and transformer [47] significantly improve the performance of non-parametric methods. GCN-based methods [4, 23] can model the vertex-to-vertex correlations, and transformer-based methods [6, 28, 29] can learn the relationships among joints and mesh vertices, thus tackling the partial occlusion issue effectively. Considerable progress has been made in HMR based on a single RGB frame, but sequence-based studies are still inadequate. Prior sequence-based methods involve the temporal information by recurrent networks [21, 54] or a tracking framework [15, 48]. However, these sequence-based methods cannot simultaneously achieve multi-modal fusion.

2.2. Event-based HMR

Event cameras [26] generate asynchronous events by measuring per-pixel brightness changes and have several merits over RGB cameras, such as high dynamic range (120 dB), high time resolution (up to 1 μs), low redundancy, and low power consumption. Recent researches have shown their potential in several vision tasks, such as detection [35], tracking [13], optical flow estimation [62], super-resolution [14], human pose estimation [64], *etc*. EventHands [39] is the first learning-based approach to conduct event-based 3D HMR solution and qualitatively demonstrates the benefits of event cameras for 3D HMR in strong light and fast motion scenes. Jalees *et al*. [37] propose an event-based hand tracking system in an energy-based optimization paradigm. Since both methods are solely based on event streams, they inevitably face low spatial resolution, foreground scarcity, and background overflow issues. As far as we know, there is no existing hand mesh reconstruction approach using both event streams and RGB frames. The closest work is EventCap [53], which applies to human pose estimation from event streams and gray-scale images for the first time. It first obtains an initial pose from gray-scale images and reconstructs human motion with high frame rate by event trajectories. Nevertheless, the initialization from gray-scale images is not robust to strong light scenes and the fitting approach using event trajectories cannot involve the appearance information from gray-scale images. In contrast

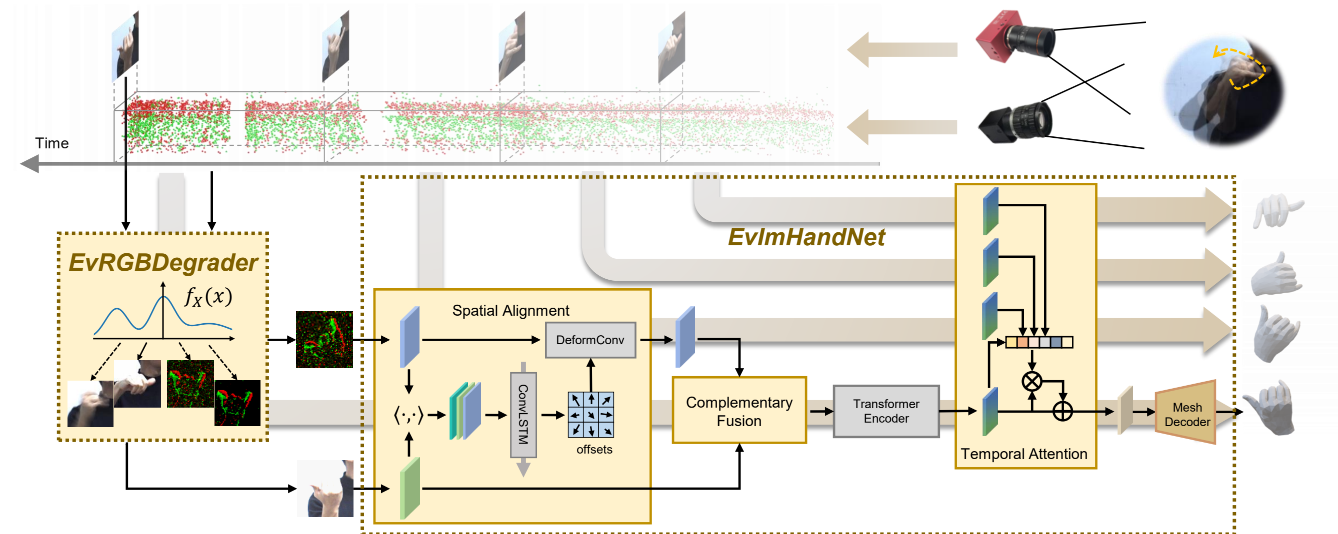


Figure 2. Overview of our pipeline. During training, we first generate various challenging scene data from normal scene sequences via EvRGBDegrader. Then we achieve spatial alignment of the event and image features using the Deformable module with temporal motion clues. Once aligned, we feed these features subsequently to complementary fusion module (detailed architecture in Fig. 3) for scene-aware fusion, the transformer encoder to learn non-local correlations and mapping them to the latent hand space. We then apply temporal attention on context hand features to leverage the spatial-temporal consistency of hand motions. Finally, the mesh decoder maps the hand features into the 3D coordinates of hand vertices and joints. In evaluation, we deactivate EvRGBDegrader.

to loose data association in EventCap [53], our approach utilizes tight feature-level fusion of the two modalities, enabling the two cameras to complement each other in HMR.

2.3. Event-image Fusion

The fusion of event streams and RGB images faces diverse challenges across data format, space, time, and information dimensions. Current fusion approaches can be broadly categorized into two main types: pixel-level and feature-level approaches. Pixel-level approaches [33, 44, 46, 50, 59] align events and images at the pixel level, leveraging the imaging constraints of event cameras for fusion. They are commonly used in low-level vision tasks. Feature-level methods [30, 34, 45] align events and images in the feature space, utilizing spatial-temporal relationships for fusion, and are frequently used in middle-level and high-level vision tasks. Since HMR aims to estimate the motion of a 3D non-rigid mesh, it is necessary to consider the complementary usage of two modal information in imaging, the spatial alignment of two free-viewpoint data, and the spatial-temporal consistency of the hand motion. This presents greater challenges than previous tasks.

3. Method

The pipeline of EvRGBHand is illustrated in Fig. 2. EvRGBHand consists of EvImHandNet to complement events and images for robust HMR in Sec. 3.2 and EvRGBDegrader to enable the model to generalize well in challenging scenes in Sec. 3.3. In EvImHandNet, we adopt spatial alignment, complementary fusion, and temporal attention to estimate hand shapes and 3D joints from the events and image pair. To address the difficulty in obtaining challenging scene data with 3D annotations, we apply data augmentation on normal scene training data through EvRGBDegrader. This effectively enhances the generalization performance of our model under challenging scenarios to outdoor scenes and another type of event camera.

3.1. Preliminaries

Hand model representation. We adopt a differentiable hand parametric MANO model [38] as hand model representation. Mesh vertices of MANO can be obtained by function $\mathbf{V} = M(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{778\times 3}$ and 3D joints $\mathbf{J}_{\text{3D}} \in \mathbb{R}^{21\times 3}$ can be recovered by regression function $\mathbf{J}_{\text{3D}} = J_{\text{reg}}(M(\boldsymbol{\theta}, \boldsymbol{\beta}))$ with pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$.

Event camera. Event cameras generate asynchronous event streams by recording the changes of per-pixel intensity $I(x, y, t)$. An event $e_i = (x_i, y_i, t_i, p_i)$ is triggered at pixel (x_i, y_i) at time t_i when the logarithmic brightness change meets the condition:

$$\log I(x_i, y_i, t_i) - \log I(x_i, y_i, t_p) = p_i C, \quad (1)$$

where t_p is the last event triggering timestamp at the same pixel, $p_i \in \{-1, 1\}$ is the polarity, C is the threshold.

3.2. EvImHandNet

To make the asynchronous event streams compatible with modern deep learning architectures [47, 49], we use the time surface representation from EventHands [39]. Considering events triggered from the hand at timestamp t are sparse, we use N events (denoted as E_t^N) before timestamp t to form a two-channel stacked frame $I_{\text{Ev},t}$ by iterating each event e_i in E_t^N as:

$$I_{\text{Ev},t}(x_i, y_i, p_i) = \frac{t_i - t_s}{t - t_s}, \tag{2}$$

where t_s is the timestamp of the first event in E_t^N. The stacked frame $I_{\text{Ev},t}$ with two channels can effectively record hand motions by assigning higher weights to events closer to the target time.

Spatial alignment. Since HMR is a task that estimates the 3D coordinates of hand vertices and joints from camera observations, aligning spatial information in both events and images is crucial. In practical applications, events and images can be captured from the same viewpoint, such as in DAVIS [26], or from different viewpoints, as seen in hybrid cameras [46, 62]. Consequently, the approach based on epipolar geometry [17, 56] lacks generality. Meanwhile, methods based on cost volumes [55] or vanilla transformer architectures [28] have a high computational cost, which is not suitable with the low-power nature of event camera. To address these challenges, we directly achieve HMR through the correlation between the data, being unaware of the relative camera positions.

To achieve spatial alignment between two modalities, we first use a shallow CNN module f^{C} (ResNet34 [16]) to extract 24×24 feature maps $F^{\text{C}}_{\text{Im},t}$, $F^{\text{C}}_{\text{Em},t}$ from the images $I_{\text{Im},t}$ and event stacked frames $I_{\text{Ev},t}$. Further, drawing inspiration from the Deformable Convolution [7, 51], we use the feature maps to estimate the offsets between events and images. To alleviate the temporal jitter in spatial alignment caused by texture mismatching between events and images, we exploit the temporal motion clues via a ConvLSTM[41] layer:

$$\Delta P = \text{ConvLSTM}(F^{\text{C}}_{\text{Im},t}, F^{\text{C}}_{\text{Ev},t}), \tag{3}$$

where ΔP are the offsets. Since the offsets are learned from the feature correlation between events and images, we can achieve alignment without estimating the relative camera pose or disparity. Leveraging these offsets, we can obtain the aligned features F_t^{A} of events and images using Deformable Convolution f^{DC} [7]:

$$F^{\text{A}}_{\text{Ev},t} = f^{\text{DC}}(F^{\text{C}}_{\text{Ev},t}, \Delta P). \tag{4}$$

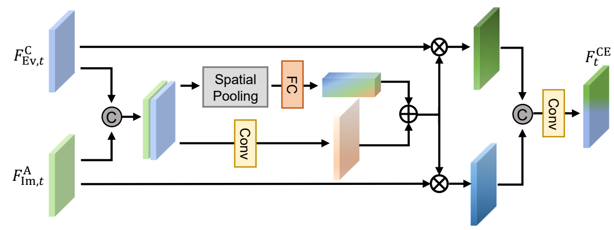


Figure 3. Detailed architecture of complementary fusion module.

Complementary fusion. Given the complementary nature of events and images, we expect our model to learn the relationship between scene and feature selection for robust HMR. To this end, we design the complementary fusion module f^{CF} [52, 57] as illustrated in Fig. 3, which can automatically compute weights based on the two modality features to obtain the complementary features:

$$F_t^{\text{CF}} = f^{\text{CF}}(F^{\text{A}}_{\text{Ev},t}, F^{\text{C}}_{\text{Im},t}), \tag{5}$$

where F_t^{CF} are down-sampled to 8×8 for latter processing.

Inspired by FastMETRO [6], we use the transformer encoder framework to build non-local relationships among the complementary features. The features F_t^{CF} are flattened as transformer tokens, and fed into the transformer encoder f^{TE} which consists of L sequential transformer blocks. The outputs of transformer blocks are latent hand features $F_t^{\text{H}} = \{F_t^l, l = 1, 2, ..., L\}$:

$$F_t^{\text{H}} = f^{\text{TE}}(F_t^{\text{CF}}). \tag{6}$$

The transformer encoder can effectively exploits non-local associations of hand observations within the feature map, addressing the self-occlusion issue in HMR.

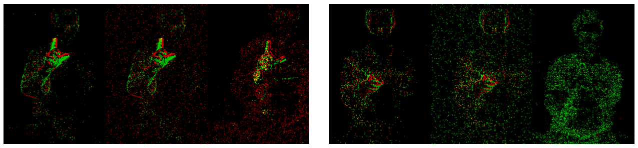
Temporal attention. Hand motion exhibits spatial-temporal continuity, and the event streams contain rich temporal and motion information. Therefore, we propose a temporal attention mechanism to effectively leverage the hand motion context information. We employ relative position encoding [40] to apply temporal attention f^{TA} for each token within the hand feature for sequential S steps:

$$F_t^{\text{TAH}}(x, y) = f^{\text{TA}}(\{F^{\text{H}}_{t+s}(x, y), \quad s = -S, ..., 0\}), \tag{7}$$

where F_t^{TAH} are the final latent hand features. On one hand, the temporal attention mechanism ensures smooth hand mocap. On the other hand, it can utilize motion information from other moments to compensate for the current instance, leading to more stable HMR.

We use a transformer decoder architecture with L transformer blocks to regress the mesh vertices and joints, which has also been adopted in FastMETRO [6]. The transformer decoder takes the learnable joint tokens $\{\mathbf{q}_1^J, \mathbf{q}_2^J, ..., \mathbf{q}_{21}^J\}$

(a) Color jitter augmentation for overexposure



(b) Salt and pepper noise for background overflow

Figure 4. Visualization of train-evaluation gap and EvRGBDegrader. For each triplet from left to right, we show original data, degraded data, real data with challenging issues.

and vertex tokens $\{\mathbf{q}_1^V, \mathbf{q}_2^V, ..., \mathbf{q}_{195}^V\}$ as input, where $\mathbf{q}_i^J, \mathbf{q}_i^V \in \mathbb{R}^D$. Given latent hand features F_t^{TAH}, the transformer decoder learns non-local correlations among vertices and joints by passing joint and vertex features through cross-attention and self-attention layers. An MLP-based 3D coordinate regressor estimates the hand mesh vertices of the coarse mesh and 3D joints using the outputs of the transformer decoder. For mesh vertices, we use an MLP layer to upsample the coarse mesh (195 vertices) to a fine mesh (778 vertices) as the hand MANO model.

3.3. EvRGBDegrader

The acquisition and annotation of high-quality 3D hand datasets are of high cost, especially under challenging scenes such as strong light, fast motion, and flash. This prompts us to leverage data under normal scenes to endow models with the capability to generalize to challenging scenes. As shown in Fig. 4, we observe that for the data pair $(I_{\mathrm{Im}}, I_{\mathrm{Ev}})$, the degradation process under challenging conditions is traceable. For instance, the brightness of image I_{Im} is high under strong light, while the distribution of I_{Ev} remains relatively stable. In flashing scenes, the mean value of I_{Ev} increases significantly along a dimension, while the texture and sharpness of I_{Im} are little affected. Therefore, EvRGBDegrader consists of three core augmentations:

- **Overexposure (OE):** For RGB frames, we use color jitter augmentation to change the image brightness and augment the strong light scenes.
- **Motion blur (MB):** For simulating motion blur, we warp the original image with optical flow following [11] in OpenCV to interpolate frames and average them.
- **Background overflow (BO):** We add salt and pepper noise on training event stacked streams to simulate the leak noise. Each pixel of the stacked frames will emit salt and pepper noise randomly.

During the training process, we apply degradation to a data pair $(I_{\mathrm{Im}}, I_{\mathrm{Ev}})$ at a certain probability to yield a degraded data pair $(I_{\mathrm{Im}}^{\mathrm{DG}}, I_{\mathrm{Ev}}^{\mathrm{DG}})$:

$$(I_{\mathrm{Im}}^{\mathrm{DG}}, I_{\mathrm{Ev}}^{\mathrm{DG}}) = f_X((I_{\mathrm{Im}}, I_{\mathrm{Ev}})), \tag{8}$$

where f_X is the degradation probability distribution of OE, MB, and BO. The t-SNE visualization in Fig. 5 implies that challenging scenes (test) and normal scenes (w/o Deg) exhibit distribution gap in the imaging descriptor space, which can be bridged by EvRGBDegrader (w/ Deg)

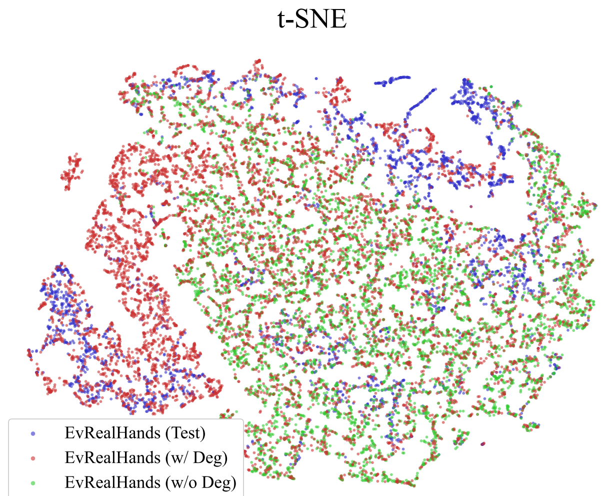


Figure 5. Visualization for events and image descriptor vectors by t-SNE. The descriptor vector has four dimensions: image sharpness, image brightness, and the means of positive and negative polarity events.

3.4. Training

Following the common practice in transformer-based mesh reconstruction method [6, 28, 29], we use vertex loss and joint loss as supervisions on each predicted result:

$$\mathcal{L}_{\mathbf{V}} = \frac{1}{M}\|\mathbf{V} - \hat{\mathbf{V}}\|_1, \quad \mathcal{L}_{\mathbf{J}} = \frac{1}{K}\|\mathbf{J} - \hat{\mathbf{J}}\|_2, \tag{9}$$

where $\mathbf{V}, \mathbf{J}$ are predicted mesh vertices and 3D joints, $\hat{\mathbf{V}}, \hat{\mathbf{J}}$ are respective ground truths, $M = 778$, and $K = 21$.

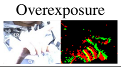
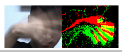
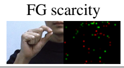
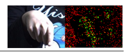
For supervision on sequential data, the total loss is the sum of vertex losses and joint losses of one hand mesh from RGB-based HMR and S sequential hand meshes from event-based HMR:

$$\mathcal{L}_{\mathrm{all}} = \lambda_{\mathbf{V}}\mathcal{L}_{\mathbf{V}}^{\mathrm{Im}} + \lambda_{\mathbf{J}}\mathcal{L}_{\mathbf{J}}^{\mathrm{Im}} + \sum_{s=1}^{S}(\lambda_{\mathbf{V}}\mathcal{L}_{\mathbf{V},s}^{\mathrm{Ev}} + \lambda_{\mathbf{J}}\mathcal{L}_{\mathbf{J},s}^{\mathrm{Ev}}). \tag{10}$$

4. Datasets and metrics

To demonstrate our method under various challenging scenarios, we collect the real-world event-based hand dataset EvRealHands with 3D annotations, which covers the

Table 1. Scenes and their corresponding issues that challenge RGB or event-based HMR in our EvRealHands datasets. (FG and BG are short for foreground and background.)

Scenes	RGB		Event	
	Overexposure	Motion blur	FG scarcity	BG overflow
				
Normal	—	—	✓	—
Strong light	✓	—	✓	—
Flash	—	—	✓	✓
Fast motion	—	✓	—	—

typical challenging issues for RGB images and events (examples in Fig. 1). To supplement training data for better performance, we develop a synthetic dataset from the RGB-based hand dataset InterHand2.6M [36].

4.1. Real-world data

The indoor sequences of EvRealHands are captured using a multi-camera system following [15, 42] with 7 RGB cameras (FLIR, 2660×2300 pixels, 15 FPS) and an event camera (DAVIS346, 346×260 pixels) capturing data from different views simultaneously. We collect 4,452 seconds of event streams and RGB images from 10 subjects. Each subject performs 15 fixed poses [8] and random hand poses. To include challenging issues caused by RGB and event imaging mechanisms, we set up strong light, flash, and fast motion scenes in addition to normal scenes. The scenes and their corresponding issues are listed in Tab. 1. Additionally, we capture data in outdoor scenes through a hybrid camera system for qualitative evaluation. The system is composed of an RGB camera (FLIR BFS-U3-51S5) and an event camera (DAVIS346 Mono [26] or PROPHESEE GEN 4.0 [12]) via a beamsplitter (Thorlabs CCM1-BS013). We collect 12 outdoor sequences (6 for DAVIS346, 6 for PROPHESEE) from 3 subjects, including sequences with fast motion, variant illuminations.

4.2. Synthetic data

To better model the distribution of real hand poses, we synthesize event streams from existing real RGB datasets. We apply v2e [18] event simulator on InterHand2.6M [36] to get synthetic event streams from RGB sequences. We select right hand sequences of 9 camera views from 4 subjects for simulation.

5. Experiments

In this section, we first introduce the experimental settings in Sec. 5.1. We then show the experimental results of EvRGBHand to demonstrate the complementary effects, generalization, and efficiency in Sec. 5.2. We also show the ablation studies in Sec. 5.3. More information about the dataset, experimental results can be found in our video and supplementary material.

5.1. Settings

Baselines. In order to demonstrate the complementary benefits of events, we compare our method with Mesh Graphormer [28], and FastMETRO [6] (denoted as FastMETRO-RGB), which are RGB-based methods on the top of FreiHand [63] leaderboard. For event-based HMR, we use EventHands [39], the only event-based HMR method with learning framework, as one of the baselines. Considering that EventHands [39] is a parametric approach, a comparison between EventHands [39] and our non-parametric approach is not sufficient in demonstrating the complementary benefits of RGB images. Therefore, we introduce FastMETRO-Event, which uses the same architecture as FastMETRO [6] and takes the same stacked event frames as input. While there are no existing methods for HMR using both events and images, we believe comparing EvRGBHand with HMR based on a single sensor would be unfair. Drawing inspiration from recent advancements in the multi-modal domain [1, 20, 24, 43], we introduce a vanilla version of event and RGB fusion for HMR (denoted as "EvRGBHand-vanilla"). Built upon the FastMETRO [6] architecture, it directly inputs the event features $F^{\mathrm{C}}_{\mathrm{Im},t}$ and the image features $F^{\mathrm{C}}_{\mathrm{Em},t}$ as tokens into the transformer encoder for fusion. The detailed architecture can be found in the supplementary materials.

Training and evaluation data. We collect 24 sequences of normal scenes 8 subjects in EvRealHands and all the InterHand2.6M [36] synthetic data as training data. And we set indoor sequences from the rest 2 subjects and all the outdoor sequences in EvRealHands as evaluation data. Our evaluation data include 4 sequences of normal scenes, 5 sequences under strong light, 2 sequences under flash light, and 3 sequences of fast motion. Following [6, 28], we only use the right hand data. We conduct both quantitative and qualitative evaluations on indoor data with 3D annotations. For data without 3D annotations (fast motion or outdoor sequences), we conduct qualitative assessments.

5.2. Results

Complementary effects on imaging issues. As quantitative results shown in Tab. 2 and qualitative results shown in Fig. 6, EvRGBHand outperforms HMR methods based on a single RGB camera or event camera and the vanilla fusion method. EvRGBHand outperforms Mesh Graphormer [28] and FastMETRO-RGB [6] on MPJPE 12 $\sim$ 18 mm lower in strong light scenes. As shown in Fig. 6, HMR methods based on RGB cameras face overexposure and motion blur issues under strong light and fast motion scenes. EvRGBHand can leverage the stable event sequences to compensate for these issues. For event-based HMR, EvRGBHand outperforms EventHands [39] and FastMETRO-Event on MPJPE 7 $\sim$ 33 mm lower in normal and flash scenes.

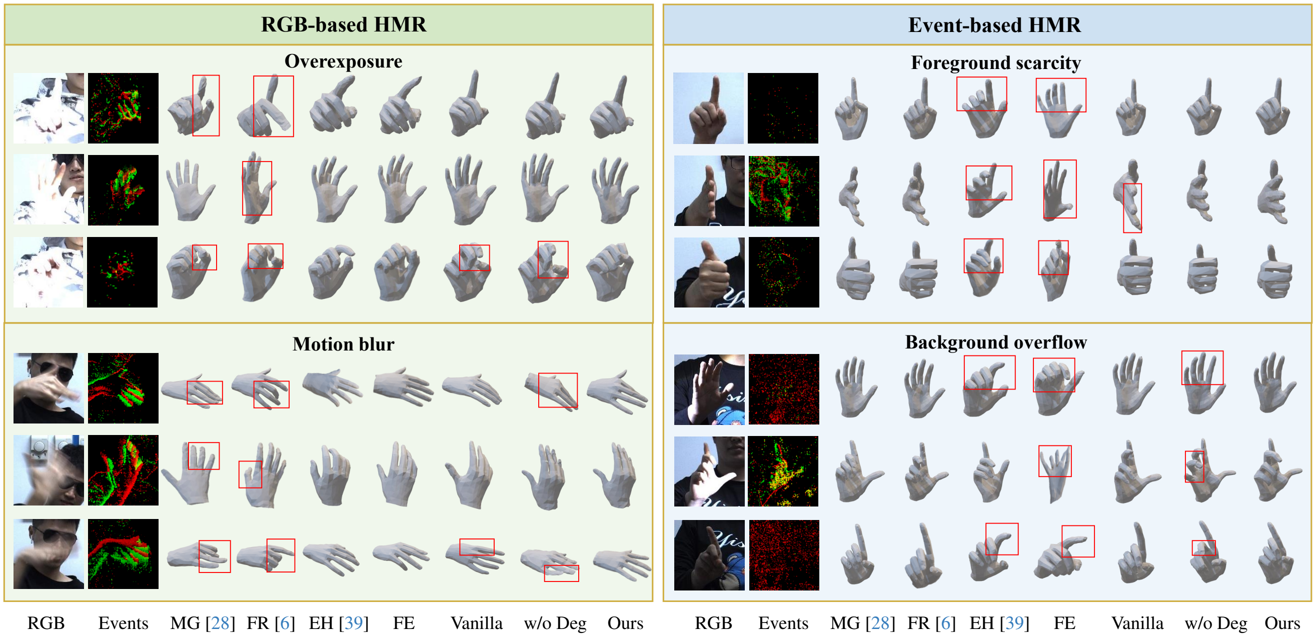


Figure 6. Qualitative analysis of HMR methods under challenging issues. For each issue, columns from left to right are RGB images, events, results from Mesh Graphormer (MG) [28], FastMETRO-RGB (FR) [6], EventHands (EH) [39], FastMETRO-Event (FE), EvRGBHand-vanilla (Vanilla), EvRGBHand without EvRGBDegrader (w/o Deg) and EvRGBHand (Ours). For easy reference, results of issues from RGB images (left side) are aligned to the event camera view and results of issues from events (right side) are aligned to the RGB camera view. EvRGBHand successfully tackles challenging issues of RGB images and event streams by compensating for each other.

Table 2. Quantitative comparison among HMR based on a single sensor or complementary usage in several scenes.

Scenes	Methods	MPJPE ↓	MPVPE ↓	PA-MPJPE ↓
Normal	Mesh Graphormer [28]	11.57	11.68	5.49
	FastMETRO-RGB [6]	11.71	12.03	5.56
	EventHands [39]	21.13	20.12	9.05
	FastMETRO-Event	18.36	17.81	7.85
	EvRGBHand-vanilla	11.84	11.98	5.07
	Ours	**11.47**	**11.63**	**5.02**
Strong Light	Mesh Graphormer [28]	40.59	38.19	13.96
	FastMETRO-RGB [6]	35.02	33.52	13.53
	EventHands [39]	27.17	25.88	9.99
	FastMETRO-Event	23.75	22.81	9.67
	EvRGBHand-vanilla	25.26	24.12	10.01
	Ours	**22.34**	**21.36**	**9.47**
Flash	Mesh Graphormer [28]	23.41	22.85	10.09
	FastMETRO-RGB [6]	24.43	23.99	9.69
	EventHands [39]	53.69	51.29	14.37
	FastMETRO-Event	36.30	35.29	13.38
	EvRGBHand-vanilla	23.13	22.88	10.02
	Ours	**20.44**	**20.47**	**8.98**

Results from Fig. 6 show that the failure of event-based methods in these scenes derives from the dynamic imaging mechanism, low texture information and noises. However, EvRGBHand can utilize the rich texture information and high pixel resolution of RGB images to improve the performance via complementary fusion. Results of EvRGBHand-vanilla and EvRGBHand in Tab. 2 and Fig. 6 indicate that, compared to the method that employs transformers for direct fusion, meticulously considering the relationships between the two modalities in spatial, temporal, and information dimensions can yield superior performance enhancements with limited training data. Furthermore, the results of EvRGBHand-vanilla also suggest that even with the most rudimentary fusion strategy, using events and images for HMR can achieve better performance than those methods based on a single sensor by MPJPE 2 $\sim$ 16 mm lower, underscoring the potential of HMR with events and images.

Generalization. As qualitative results shown in Fig. 7, although EvRGBHand was trained under normal indoor scenes using the DAVIS346 camera [26], it still generalize well in challenging outdoor environments (natural various lighting, fast motion) and data captured by the PROPHESEE GEN 4.0 [12], significantly outperforming other methods. This can be attributed, on one hand, to our fusion strategy across temporal, spatial, and informational dimensions. On the other hand, it derives from the efforts of EvRGB-Degrader in bridging the distribution gap between indoor-outdoor data and normal-challenging scenes.

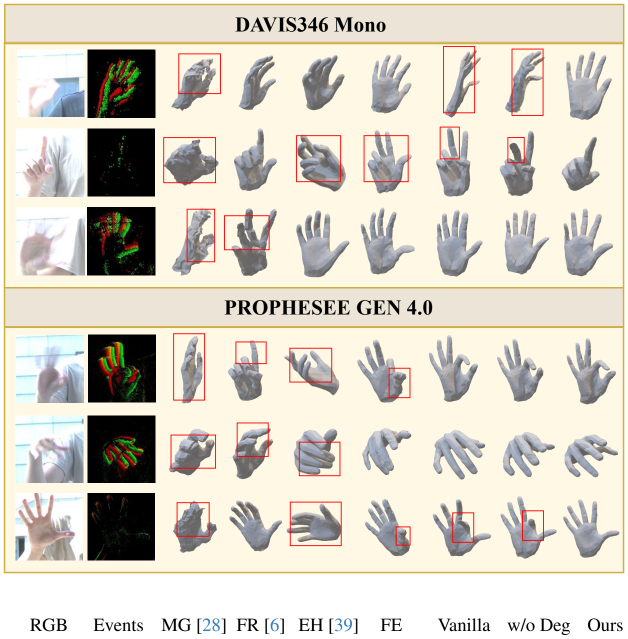


Figure 7. Qualitative analysis of HMR methods on outdoor DAVIS346 sequences and PROPHESEE GEN 4.0 sequences. EvRGBHand generalizes better than other methods.

Table 3. Computational cost and average accuracy.

Methods	Params↓	FLOPs↓	MPJPE↓	MPVPE↓
EventHands [39]	22.68 M	2.81 G	30.44	29.24
FastMETRO-Event	141.68 M	10.79 G	23.59	23.12
EvRGBHand-vanilla	277.02 M	17.90 G	17.45	17.30
Ours	55.92 M	8.15 G	16.66	16.43

Efficiency. As shown in Tab. 3, EvRGBHand has 60.5% fewer Params and requires 24.5% fewer FLOPs than FastMETRO-Event, while achieves better performance with 6.9 mm average MPJPE lower. Compared with EvRGBHand-vanilla, EvRGBHand with carefully designed architecture can achieves 79.8% fewer Params and 54.5% fewer FLOPs with better average accuracy.

5.3. Ablation Studies

EvImHandNet. As quantitative results shown in Sec. 5.3, spatial alignment (SA), complementary fusion (CF), and temporal attention (TA) all contribute to the stable HMR performance. Compared to the vanilla fusion strategy, these modules collectively lead to an improvement of $2.5 \sim 3$ mm MPJPE in challenging scenes.

EvRGBDegrader Quantitative results in Sec. 5.3 shows that the simulations of overexposure (OE) and background overflow (BO) significantly improve the performance on indoor challenging scenes ($8 \sim 20$ mm MPJPE lower). Qualitative results in Fig. 6 and Fig. 6 between “w/o Deg” and “Ours” show that EvRGBDegrader effectively promote the performance in strong light and fast motion scenes. This indicates that EvRGBDegrader can effectively bridge the data distribution gap between normal collection settings and outdoor evaluation scenarios.

Table 4. Ablation studies.

EvImHandNet			EvRGBDegrader			MPJPE (mm)↓		
SA	CF	TA	OE	MB	BO	Normal	Strong light	Flash
✗	✗	✗				11.81	25.35	22.99
	✗	✗				11.60	24.23	22.86
		✗				11.57	23.87	22.52
			✗			11.53	45.11	28.52
				✗		11.48	23.50	21.13
					✗	11.63	27.83	23.33
			✗	✗	✗	11.73	47.34	29.02
						11.47	**22.34**	**20.43**

Figure 8. RGB images and failure cases of EvRGBHand.

6. Conclusion

In this paper, we explore the potential of complementary usage of event cameras and color cameras for hand mesh reconstruction tasks. To this end, we introduce a framework, EvRGBHand, which leverages the strengths of both event camera and color camera imaging to achieve robust and efficient HMR. Through multi-modal information fusion and degradation augmentation, our approach demonstrates potential generalization capabilities with low data cost in outdoor scenes and another type of event camera.

Limitations. As shown in Fig. 8, when overexposure and motion blur issues are observed together with challenging hand poses, it is challenging for EvRGBHand to output proper predictions. Besides, the respective performance from complementary use of event streams and RGB frames in our experiment is affected by the different pixel resolutions. As event cameras evolve, we expect future work to collect data from event cameras with higher image resolution and lower noise to rigorously validate the effects of complementing event streams and RGB images.

Acknowledgements

This work is supported by Beijing Natural Science Foundation (Grant No. L233024), and National Natural Science Foundation of China (Grand No. 62088102, 62136001).

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, 2021. 2, 6

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *CVPR*, 2019. 2

[3] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3D hand pose estimation from monocular RGB images. In *ECCV*, 2018. 2

[4] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. MobRecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, 2022. 1, 2

[5] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3D hand reconstruction via self-supervised learning. In *CVPR*, 2021. 2

[6] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In *ECCV*, 2022. 1, 2, 4, 5, 6, 7, 8

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 4

[8] Quentin De Smedt, Hazem Wannous, J-P Vandeborre, Joris Guerry, B Le Saux, and David Filliat. 3D hand gesture recognition using a depth and skeletal dataset: Shrec'17 track. In *Proceedings of the Workshop on 3D Object Retrieval*, 2017. 6

[9] Xiaoming Deng, Yuying Zhu, Yinda Zhang, Zhaopeng Cui, Ping Tan, Wentian Qu, Cuixia Ma, and Hongan Wang. Weakly supervised learning for single depth-based hand shape recovery. *IEEE Transactions on Image Processing*, 30:532–545, 2020. 1

[10] Linpu Fang, Xingyan Liu, Li Liu, Hang Xu, and Wenxiong Kang. JGR-P2O: Joint graph reasoning based pixel-to-offset prediction network for 3D hand pose estimation from a single depth image. In *ECCV*, 2020. 1

[11] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, 2003. 5

[12] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick T. Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch. A 1280×720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86μm pixels, 1.066GEPS readout, programmable event-rate controller and compressive data-formatting pipeline. In *IEEE International Solid- State Circuits Conference*, pages 112–114, 2020. 6, 7

[13] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *ECCV*, 2018. 2

[14] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. EvIntSR-Net: Event guided multiple latent frames reconstruction and super-resolution. In *ICCV*, 2021. 2

[15] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. MEgATrack: Monochrome egocentric articulated hand-tracking for virtual reality. *ACM TOG*, 2020. 2, 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[17] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *CVPR*, 2020. 4

[18] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. v2e: From video frames to realistic DVS events. In *CVPRW*, 2021. 6

[19] Lin Huang, Jianchao Tan, Ji Liu, and Junsong Yuan. Hand-transformer: Non-autoregressive structured modeling for 3D hand pose estimation. In *ECCV*, 2020. 1

[20] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. 2021. 2, 6

[21] Leyla Khaleghi, Alireza Sepas-Moghaddam, Joshua Marshall, and Ali Etemad. Multi-view video-based 3D hand pose estimation. *TAI*, 2022. 2

[22] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2

[23] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 2

[24] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas M. Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *ICLR*, 2021. 6

[25] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, 2022. 1

[26] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbrück. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *JSSC*, 2008. 1, 2, 4, 6, 7

[27] Guan Ming Lim, Prayook Jatesiktat, and Wei Tech Ang. Mobilehand: Real-time 3D hand shape and pose estimation from color image. In *ICONIP*, 2020. 2

[28] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 2, 4, 5, 6, 7, 8

[29] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2, 5

[30] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In *CVPR*, 2023. 2, 3

[31] Siddharth Mahendran, Haider Ali, and René Vidal. A mixed classification-regression framework for 3D pose estimation from 2D images. In *BMVC*, 2018. 2

[32] Jameel Malik, Soshi Shimada, Ahmed Elhayek, Sk Aziz Ali, Vladislav Golyanik, Christian Theobalt, and Didier Stricker. Handvoxnet++: 3D hand shape and pose estimation using voxel-based neural networks. *IEEE TPAMI*, 2021. 1

[33] Nico Messikommer, Stamatios Georgoulis, Daniel Gehrig, Stepan Tulyakov, Julius Erbach, Alfredo Bochicchio, Yuanyou Li, and Davide Scaramuzza. Multi-bracket high dynamic range imaging with event cameras. In *CVPR*, 2022. 3

[34] Nico Messikommer, Carter Fang, Mathias Gehrig, and Davide Scaramuzza. Data-driven feature tracking for event cameras. In *CVPR*, pages 5642–5651, 2023. 2, 3

[35] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. 2018. 2

[36] 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 *ECCV*, 2020. 6

[37] Jalees Nehvi, Vladislav Golyanik, Franziska Mueller, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Differentiable event stream simulator for non-rigid 3D tracking. In *CVPRW*, 2021. 1, 2

[38] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied Hands: Modeling and capturing hands and bodies together. *ACM TOG*, 2017. 2, 3

[39] 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. 1, 2, 4, 6, 7, 8

[40] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. 2018. 4

[41] Xingjian Shi, Zhourong Chen, Hao Wang, D. Y. Yeung, Wai-Kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 4

[42] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 6

[43] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022. 6

[44] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *ECCV*, 2022. 3

[45] Abhishek Tomy, Anshul Paigwar, Khushdeep Singh Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and RGB camera for robust object detection in adverse conditions. 2022. 3

[46] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time Lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *CVPR*, 2022. 3, 4

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 4

[48] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. RGB2Hands: real-time tracking of 3D hand interactions from monocular RGB video. *ACM TOG*, 2020. 2

[49] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 2021. 4

[50] Zihao W. Wang, Peiqi Duan, Oliver Cossairt, Aggelos K. Katsaggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *CVPR*, 2020. 2, 3

[51] Junwu Weng, Mengyuan Liu, Xudong Jiang, and Junsong Yuan. Deformable pose traversal convolution for 3d action and gesture recognition. In *ECCV*, pages 136–152, 2018. 4

[52] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 4

[53] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. EventCap: Monocular 3D capture of high-speed human motions using an event camera. In *CVPR*, 2020. 2, 3

[54] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. SeqHAND: RGB-sequence-based 3D hand pose and shape estimation. In *ECCV*, 2020. 2

[55] Jiayu Yang, Wei Mao, José M. Álvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *CVPR*, 2020. 4

[56] Lixin Yang, Jian Xu, Licheng Zhong, Xinyu Zhan, Zhicheng Wang, Kejian Wu, and Cewu Lu. POEM: reconstructing hand in a point embedded multi-view stereo. In *CVPR*, 2023. 4

[57] Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, and Boxin Shi. Learning event guided high dynamic range video reconstruction. In *CVPR*, pages 13924–13934, 2023. 4

[58] Dehao Zhang, Qiankun Zhou, Duan Peiqi, Zhou Chu, and Boxin Shi. Data association between event streams and intensity frames under diverse baselines. In *ECCV*, 2022. 2

[59] Jiqing Zhang, Yuanchen Wang, Wenxi Liu, Meng Li, Jinpeng Bai, Baocai Yin, and Xin Yang. Frame-event alignment and fusion network for high frame rate tracking. In *CVPR*. IEEE, 2023. 3

[60] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *ICCV*, 2019. 2

[61] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, 2020. 2

[62] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, 2019. 2, 4

[63] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan C. Russell, Max J. Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, 2019. 6

[64] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. EventHPE: Event-based 3D human pose and shape estimation. In *ICCV*, 2021. 2