Continuous-Time Human Motion Field from Event Cameras

Ziyun Wang[1,2], Ruijun Zhang[1], Zi-Yan Liu[1], Yufu Wang[1], Kostas Daniilidis[1,3]
[1]University of Pennsylvania, USA
[2]Johns Hopkins University, USA
[3]Archimedes, Athena RC

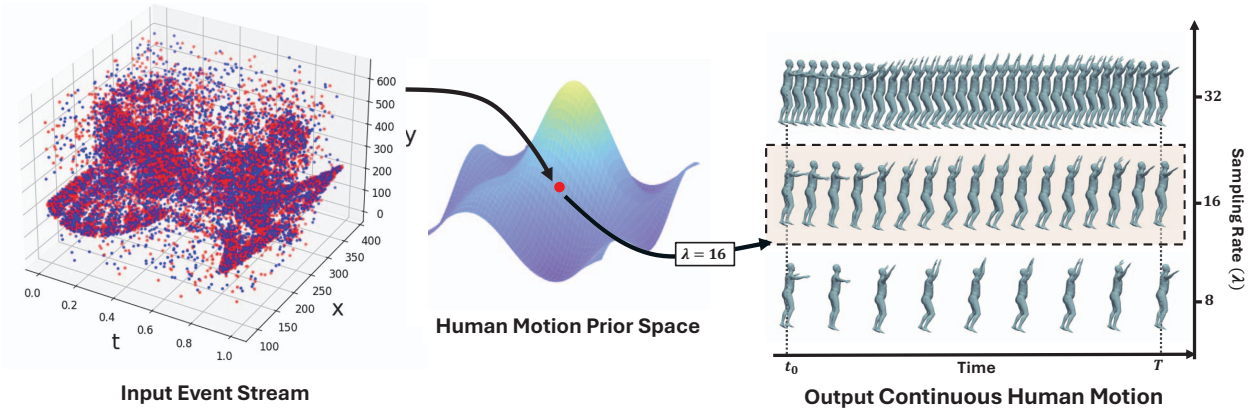

Figure 1. **EvHuman** predicts a set of global and local latent codes from an event stream to represent continuous-time human motions. The latent codes are decoded by a neural human motion prior in a time-continuous MLP network that can be queried at any time resolution in parallel efficiently. The human sequences on the rigth are decoded with a test event stream in MMHPSD [74].

Abstract

This paper addresses the challenges of estimating a continuous-time human motion field from a stream of events. Existing human motion estimation methods rely predominantly on frame-based approaches, which are prone to aliasing and inaccuracies due to limited temporal resolution and motion blur. In this work, we predict a continuous-time human motion field directly from events, by leveraging a recurrent feed-forward neural network to predict human motion in the latent space of possible human motions. Prior state-of-the-art event-based methods rely on computationally intensive optimization across a fixed number of poses at high frame rates, which becomes prohibitively expensive as we increase the temporal resolution. In comparison, we present the first work that replaces traditional discrete-time predictions with a continuous human motion field represented as a time-implicit function, enabling parallel pose queries at arbitrary temporal resolutions. Despite the promises of event cameras, few benchmarks have tested the limit of high speed human motion estimation. We introduce Beam-splitter Event Agile Human Motion Dataset—a hardware-synchronized high-speed human dataset to fill this gap. On this new data, our method improves joint errors by 23.8 % compared to previous event human methods, while reducing the computational time by 69%. More details of the work can be found on the project page: ziyunclaudewang.github.io/evhuman.html.

1. Introduction

Human Mesh Recovery (HMR) methods recover the full 3D mesh of a moving human from a video, which has been a core research problem in computer vision. However, with highly dynamic human motions, performing such tasks with traditional cameras is challenging because frame-based cameras can only provide sampling of human motions at a limited frame rate. First, it is challenging to predict the correct motion when the subject is moving fast and the time resolution of a video is low. Additionally, fast motions are often accompanied by motion blur that squashes motion information over time, preventing the network from obtaining the correct pose information.

To address these issues, researchers have begun exploring event cameras as an alternative sensor modality [8, 66, 73–75]. Event cameras are known for their high temporal resolution, high dynamic range, and low data throughput. Due to their asynchronous design, there is no fixed global

Work completed while Ziyun Wang was at University of Pennsylvania.

shutter time, which helps mitigate motion blur. Event data captured with a static camera are inherently background and aliasing free, providing a continuous motion signal rather than discretized frames. Additionally, event cameras can estimate the motion of 2D pixels more robustly because only changes in the scene are recorded [16, 18, 63–65, 71, 72]. These advantages make event cameras ideal sensors for capturing high-speed human motion under various lighting conditions. Despite the higher temporal resolution and better motion signals, existing approaches assume that the predicted poses are represented as a sequence of discrete poses, which is computationally expensive to optimize and requires a fixed number of predicted poses known a priori. For high-speed prediction, the optimization performance can be up to 6750 times slower than real time [66]. Although learning-based methods are faster [8, 74], the inference time scales linearly with the number of query poses, and the pose error increases due to chaining short predictions.

In this work, we introduce *EvHuman*, the first learning-based human motion estimation approach that directly outputs a continuous-time motion field from events, enabling the prediction of human pose at any arbitrary timestamp within the event stream. Our approach significantly outperforms the prior methods across a variety of HMR metrics while significantly reducing the computational time of prediction sequences of human poses at high temporal resolutions.

Unlike existing methods that predict human poses frame-by-frame, EvHuman learns latent codes of human motion, which are then decoded with a human motion prior network pretrained on a wide range of diverse human motions. The decoder itself is a time-continuous function, predicting both root and local poses for any specified query time. A global motion predictor takes in the predicted poses, joint positions, and velocities, and maps them to global velocities. Our training process incorporates traditional supervised losses and introduces a novel event-based contrast maximization loss using vertex optical flow derived from the predicted human meshes. During inference, unlike optimization methods like EventCap [66], EvHuman does not need a fixed set of initial guessed poses. Instead, it encodes an entire event stream once and can predict pose prediction at any time resolution by evaluating at arbitrary timestamps.

We evaluated our method against state-of-the-art event and image human methods on MMHPSD [74] and our novel **Beam-splitter Event Agile Human Motion Dataset (BEAHM)**. BEAHM was collected with a custom-built event/image beam splitter and multiple high frame-rate cameras to capture high-speed human mesh labels. Precise hardware synchronization aligns events and images temporally for accurate benchmarking and ground-truth labeling. We make publicly available all raw events, images, the beam splitter design, and data collection software. Our main contributions are as follows.

- We introduce the first feed-forward event-based continuous-time human motion field leveraging neural human motion priors, advancing the state of the art performance for event-based human mesh by 23.8 % while reducing the computational time by 69 %.
- We design a novel event-based human mesh motion loss that explicitly maximizes event contrast based on flows rendered from our continuous-time human motion field.
- We collected a new high-resolution event-based human pose dataset, Beam-splitter Event Agile Human Motion Dataset, that provides ground truth meshes at 120 FPS.

2. Related Work

Event-based Human Pose Estimation 3D human pose estimation is grouped into two categories. The first category estimates 3D skeletal joint positions [40, 42, 47, 48, 57, 59]. The second category, which is more related to our problem, recovers a parametric 3D human mesh, such as the SMPL model [35]. To recover SMPL parameters, methods employ an optimization-based approach by fitting to the image evidence [3, 6, 11], or learn from the data to directly regress the pose and shape parameters [17, 23, 26, 28, 46, 58]. Our method is a regression approach. However, instead of directly regressing the parameters, we predict a latent representation [19] that is decoded to the SMPL parameters. Recovery of global human motion from a dynamic camera is more challenging and often requires additional sensors [20, 25, 60] or integration with SLAM techniques [27, 34, 52, 61, 68]. In this study, we assume a static camera setup; however, event-based approaches face difficulties with static humans because no events are generated.

Event-based human pose estimation has advanced through datasets like DHP19 [8], which support 2D joint detection and triangulation. Recent developments include TORE's volume-based representation for joint lifting [4], Scarpellini's end-to-end single-camera framework [53], and Chen's point aggregation approach [9]. For 3D mesh recovery, EventCap [66] optimizes human mesh through tracking joints through events, while EventHPE [74] learns 3D human pose through poses and optical flow supervision. A spiking-based extension is used to improve energy efficiency for event-based HPE [75].

Unsupervised Event Optical Flow Estimation Learning-based flow estimation from events has been extensively studied in recent years [5, 16, 37, 44, 67, 69, 71, 72]. Contrast Maximization (CM) methods have demonstrated competitive performance in optical flow estimation using only event data [13, 18, 56, 67, 70]. Ye et al. [67] introduced a pipeline for learning Egomotion, which is guided by aligning adjacent event slices using predicted rigid flow. Zhu et al. [72] introduced a novel timestamp-based motion

loss to enhance the robustness of contrast calculations. Gallego et al. [13, 14] introduced a comprehensive framework, which extends the application of contrast maximization to both flow and depth estimation. A key advantage of the contrast maximization approach is that it requires only event data as input, enabling fine motion supervision even when ground truth is unavailable.

Learned Human Motion Priors Various techniques have been proposed to provide priors for motion estimation [7, 22, 29, 31, 32, 32, 33, 51]. Unlike physics-based methods, these approaches learn probabilistic transitions between states from motion capture data [39]. Motion variational auto-encoders can be trained to animate single characters by sampling the distribution of possible motions [32]. HuMoR [50] learns a 3D human dynamical model based on the conditional variational autoencoder, which describes the transition probability between two consecutive human states. He et al. [19] employ an implicit function to represent continuous human motion. PACE [27] extends this method and shows superior performance in world-grounded human motion estimation. Unlike NeMF-based optimization, which fits to a fixed number of initialized poses, our approach takes full advantage of the continuous poses in training, by computing motion induced optical flow to self-supervise event networks.

3. Preliminaries

Event Modeling and Representation. We denote the brightness at spatial coordinate (x, y) at time t as $I(x, y, t)$. Each event is triggered if the logarithmic brightness $|\log I(x, y, t) - \log I(x, y, t - \Delta t)| > C$, where C is a contrast threshold and Δt is the time since the last event at this pixel. Each event is a tuple of $e_i = (x_i, y_i, t_i, p_i)$, where (x_i, y_i) is the event spatial coordinate, t_i is the event timestamp, and p_i (polarity) is the sign of $\log I(x_i, y_i, t_i) - \log I(x_i, y_i, t_i - \Delta t)$. Given a stream of events $\mathcal{E} = \{e_k\}_{k=0}^{N}$, an event volume computes a tensor representation of the events as

$$E(x, y, t) = \sum_i p_i k_b (x - x_i) k_b (y - y_i) k_b (t - t_i),$$

where k_b is a bilinear sampling kernel function. Following [62, 71, 72], we discretize t into T bins and represent the event volume as a tensor $E \in \mathbb{R}^{H \times W \times T}$.

Representing 3D Humans. We adopt SMPL [36] to represent 3D humans, whose parameters include body pose $\theta \in \mathbb{R}^{24\times3}$, shape $\beta \in \mathbb{R}^{10}$, and root translation $\tau \in \mathbb{R}^3$ in the camera space. The parameters map to set of 3D vertices $\mathcal{M} = \mathcal{S}(\theta, \beta, \tau) \in \mathbb{R}^{V\times3}$ with a differentiable SMPL layer. The human motion is commonly treated as a discrete set of prediction $\{\mathcal{M}_t\}$ at a fixed frame rate. In this study, we leverage the high temporal resolution of events to recover a continuous-time function $\mathcal{M}(t)$, as described in Sec. 4.1.

4. Method

Given a stream of events triggered by the human motion, our goal is to recover the continuous 3D human motion of this event duration. We follow previous event-based HMR methods [66, 74] to focus on motion tracking, assuming the shape parameter β and first poses θ_0 are known or could be initialized by RGB-based methods.

4.1. Human Motion Field

A human motion field represents a motion sequence as a continuous function $\mathcal{M} : t \rightarrow \theta_t$, which maps the continuous temporal coordinates t to the human pose at t. Similar to other neural fields [41, 45], $\mathcal{M}$ can be approximated by a multilayer perceptron (MLP). Describing motion as a field takes advantage of temporally dense event signals, allowing us to recover smooth motions at a flexible sampling rate. Given an event volume E, our method recovers a human motion field as:

$$\mathcal{M}(t) = \mathcal{F}(E) = F_m(t; F_e(E)), \tag{1}$$

where $\mathcal{M}(t)$ is composed of two functions F_e and F_m. F_e is an event-based encoder that maps events E to a latent point z, which resides in the latent space of a Variational Autoencoder (VAE) pre-trained on the AMASS [39] human motion dataset. The motion decoder F_m, a fixed VAE decoder, reconstructs the motion parameters θ_t from the input time t and the latent representation z. Following NeMF [19], we model human motion as a generative latent variable model:

$$\theta_t = \mathcal{F}_m(t; z), \tag{2}$$

where z is a latent code. The decoder $\mathcal{F}_m$ can be trained with only motion-captured human data. We freeze the pre-trained decoder $\mathcal{F}_m$ from NeMF and focus on learning F_e, which maps events into the learned latent space, which provides a prior for plausible human movements.

To predict a human motion from events, we infer z as

$$z = \mathcal{F}_e(E). \tag{3}$$

In NeMF, the latent code is decomposed into local codes z_l and global code z_g to facilitate training. To ensure consistency with the pre-trained decoder, we also predict z_l and z_g. We describe the implementation of $\mathcal{F}_e$ in Sec. 4.2 next.

4.2. Event Human Motion Predictor

The Event Human Motion Predictor $\mathcal{F}_e$ predicts the latent codes z_l and z_g from the input events E. We first encode each event slice E_t with a convolutional image encoder

$$f_t = \text{Encoder}(E_t). \tag{4}$$

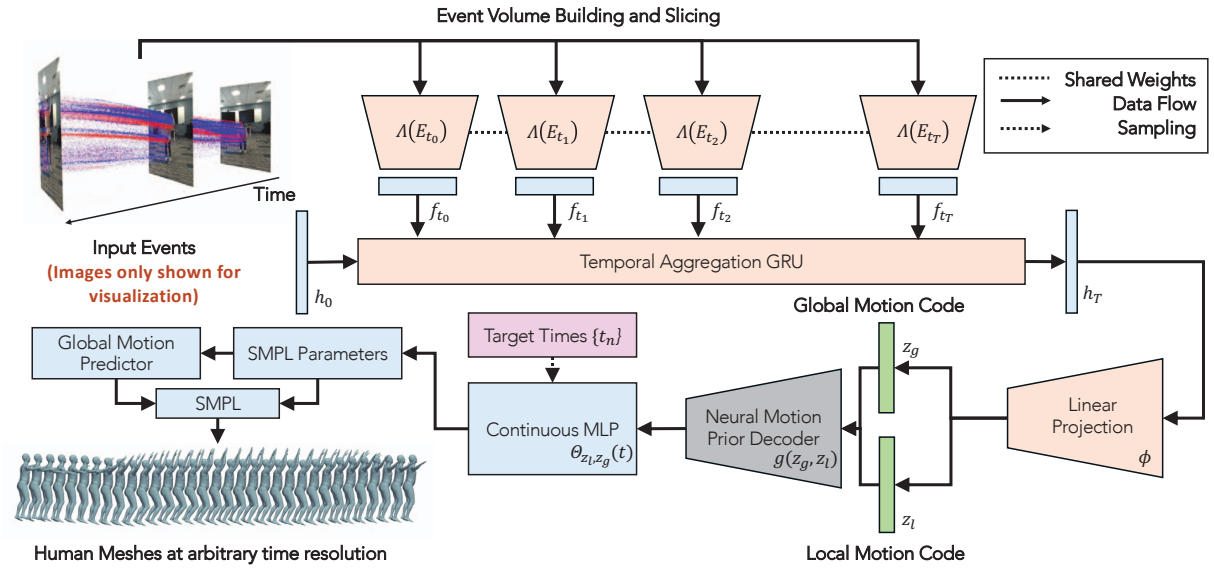

Figure 2. **Pipeline:** EvHuman first aggregate features computed from event volumes with a recurrent network. We project the aggregated terminal hidden state into a pair of latent codes, which are decoded with a pre-trained neural motion prior decoder to produce a continuous function as a MLP network. Finally, we estimate the global translations from the joint velocities and rotations.

The features from different temporal slices are aggregated with a GRU-based recurrent network [10]. The recurrent network predicts a final feature h_T, which is projected to z_l and z_g with two MLPs.

$$h_t = \text{Encoder}_{\text{GRU}}(h_{t-1}, f_t) \tag{5}$$

$$z_g, z_l = \text{Head}(h_T) \tag{6}$$

An important difference between events and images is that events do not exist for non-moving parts of the scene, assuming constant lighting. Thus, it is crucial to consider the temporal relationship among the stream of events, rather than predicting the pose independently for each event frame as in [8].

4.3. Global Motion Estimation

The recovered human motion field $\mathcal{M}(t)$ is inherently a local representation, encoding body poses $\{\theta_t\}_1^T$ but not the root locations $\{\tau_i\}_1^T$ in the world or camera coordinate. Therefore, we follow NeMF and train a global motion predictor (GMP) to predict the root velocity from the predicted body poses.

Specifically, at time t, we use the SMPL layer to obtain the 3D joints $\mathcal{J} \in \mathcal{R}^{23\times 3}$ from the prediction θ_t. In addition, we use neighboring times to compute the velocity $\dot{\mathcal{J}}$ and angular velocity $\dot{\theta}$. We then use an MLP network to predict the root velocity based on these inputs

$$\dot{\tau}_t = \text{GMP}(\theta, \dot{\theta}, \mathcal{J}, \dot{\mathcal{J}}) \tag{7}$$

Finally, the velocity is integrated using Euler's method iteratively forward $\tau_{t+\delta t} = \tau_t + \dot{\tau}_t \delta t$. Although NeMF also provides a pretrained GMP, it makes predictions in AMASS's world coordinate, while we are interested in predicting translation in the camera coordinate. Therefore, we train this component from scratch in our training procedure. Please see details of GMP in the Supplementary Material.

4.4. Human Mesh Event Contrast Maximization

Given a continuous motion model $\Theta_{\mathbf{z}_l,\mathbf{z}_g}(t)$ and the translation τ_t computed in Sec. 4.3, the vertices can obtained by:

$$V_t^v = \mathcal{W}(\mathcal{S}(\Theta_{\mathbf{z}_l,\mathbf{z}_g}(t), \beta, \tau_t), v) \tag{8}$$

where $\mathcal{S}$ is the parametric SMPL [36] that returns a global human mesh parameterized by joint poses and shape parameter β, $\mathcal{W}$ is a skinning function on the mesh, and v the vertex index. This continuous-time motion model gives us the full trajectory of vertices in 3D. The motion field defined between two times t_i and t_j can be computed by subtracting the 2D location of the same vertex:

$$\mathbf{F}_{i,j,v}^{shape} = \mathbf{1}_{vis}(v)(\pi(V_{t_i}^v) - \pi(V_{t_j}^v)), \tag{9}$$

where π is the perspective projection function given known intrinsics. While it is easy to use all vertices in contrast maximization, the back of the human motion (with respect to the camera) can produce erroneous flow. We use differentiable renderers to render optical flow on mesh triangles using Barycentric coordinates [30] so that flow can be differentiable with respect to the SMPL parameters. We denote the visibility of the vertices as $\mathbf{1}_{vis}(v)$ based on mesh rasterization. Inspired by unsupervised event-based optical flow methods [14, 18, 67, 72], we maximize the variance

of the Image of Warped Events (IWE). The IWE of events looks sharp if we properly predict the 2D flow of humans. Flow from low-parameter models such as SMPL is uniquely compatible with Contrast Maximization because it avoids problems such as event collapse. Qualitative results can be found in Fig. 5. Given the events $E_{ij} = \{\mathbf{x_k}, p_k, t_k\}$ between t_i and t_j, the motion-compensated events become

$$\mathbf{x'_k} = \mathbf{x_k} - \mathbf{F}^{shape}_{i,j,v}(\mathbf{x_k})(t_r - t_k). \tag{10}$$

The image of warped events (IWE) is defined as

$$I(\mathbf{x}; \mathbf{F}^{shape}_{i,j,v}) = \sum_{k}^{|E_{ij}|} b_k \delta(\mathbf{x} - \mathbf{x'_k}), \tag{11}$$

where b_k is the polarity indicator that separates events into a positive and a negative image, and δ is the bilinear kernel function. Various objective functions are described in detail [14]. We maximize the image variance:

$$\text{Var}(I) = \frac{1}{MN} \sum_{u=1}^{M} \sum_{v=1}^{N} (I(u,v) - \mu)^2 \tag{12}$$

$$\mu = \frac{1}{MN} \sum_{u=1}^{M} \sum_{v=1}^{N} I(i,j). \tag{13}$$

Since we maximize the variance, the loss function is the negative variance $\mathcal{L}_c = -\text{Var}(I)$.

4.5. Training

Loss Functions. We follow best practices from human mesh regression [23, 28] and train our model with a combination of 2D and 3D losses. We compute the losses between the SMPL predictions and the ground truth labels

$$\mathcal{L}_\theta = \sum_{t=1}^{T} ||\hat{\theta}_t - \theta_t||_F^2$$
$$\mathcal{L}_t = \sum_{t=1}^{T} ||\hat{\tau}_t - \tau_t||_F^2$$

where the hat operator denotes the ground truth labels. We also compute losses on the joints obtained from SMPL meshes and their 2D projection using ground truth camera parameters

$$\mathcal{L}_{3D} = \sum_{t=1}^{T} ||\hat{\mathcal{J}}_{3D} - \mathcal{J}_{3D}||_F^2$$
$$\mathcal{L}_{2D} = \sum_{t=1}^{T} ||\hat{\mathcal{J}}_{2D} - \Pi(\mathcal{J}_{3D})||_F^2$$

where $\mathcal{J}_{3D}$ are the 3D joints from the SMPL model and Π is the camera reprojection operator.

Additionally we follow EventHPE [74] to compute the cosine difference between the vertex flow computed from the predicted human motion field $F^{shape}_{t_i,t_j}$ and the event-based flow $F^{e}_{t_i,t_j}$ when a pre-trained flow network is available:

$$\mathcal{L}_{flow} = \sum_{t=1}^{T} \sum_{v} \frac{< \mathbf{F}^{shape}_{t,v} \mathbf{F}^{e}_{t,v} >}{||\mathbf{F}^{shape}_{t,v}||_2 \cdot ||\mathbf{F}^{e}_{t,v}||_2} \tag{14}$$

The flow network is trained in an unsupervised fashion using images and events, following Ev-FlowNet [71], to constrain the pairwise motion [55]. Finally, we compute the unsupervised flow loss $\mathcal{L}_c$ as described in Sec. 4.4. The final loss function is the weighted sum of the terms above:

$$\mathcal{L} = \lambda_\theta \mathcal{L}_\theta + \lambda_t \mathcal{L}_t + \lambda_{3D} \mathcal{L}_{3D} + \lambda_{2D} \mathcal{L}_{2D} + \lambda_{flow} \mathcal{L}_{flow} + \lambda_c \mathcal{L}_c. \tag{15}$$

Training Strategy. We first train a global motion predictor (GMP) (Sec. 4.3) using ground-truth local motions as input. Then, we freeze the global motion predictor and train the event human motion predictor (Sec. 4.2). In the end, we freeze the local motion prediction and fine-tune the GMP for 1 epoch. Details of training and full hyperparameters can be found in the Supplementary Material.

Dataset	Label	Label FPS	Sync	Motions
CDEHP [54]	2D joints	60	-	25
DHP19 [8]	2D joints	100	Hard	33
MMHPSD [74]	Mesh	15	Soft	21
EventCap [66]	Mesh	100	-	12
BEAHM	**Mesh**	**120**	**Hard**	**40**

Table 1. Comparison of event human pose estimation datasets.

Figure 3. Two example sequences of BEAHM. **(A)** Multi-cam images with ground truth annotations, including bounding boxes and skeletons. **(B)** Events plotted on the RGB image. **(C)** Ground truth human meshes superimposed on the RGB images.

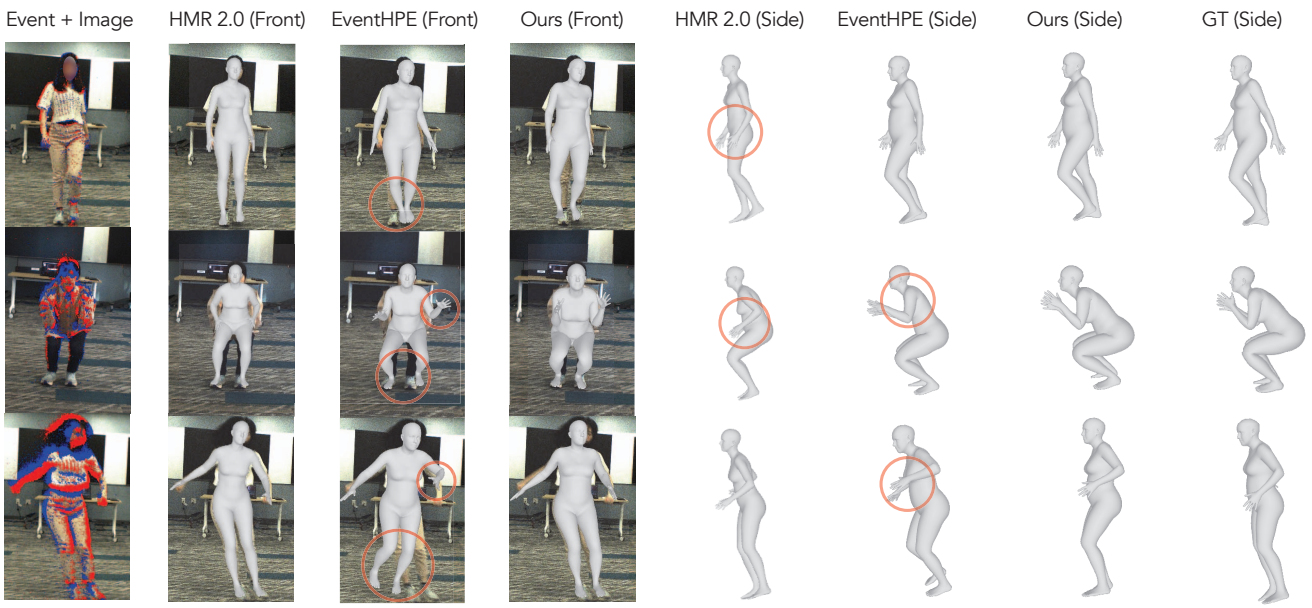

Figure 4. **Qualitative results of comparison with baseline methods**. We mark the erroneous predictions with red circles. For each method, we include the front view and the side view. In the first row, HMR 2.0 predicts erroneous unnatural motion seen from the side. The bottom two rows show the robustness of our method in fast motions compared to baseline methods.

Figure 5. Sharpness of IWE improves when predicted motion is correct. **Left**: Image of raw events. **Middle**: Image of motion-compensated events. **Right:** Optical flow from our continuous-time human motion field. Color indicates direction of optical flow.

5. Beam-splitter Event Agile Human Motion Dataset (BEAHM)

A key challenge in studying event-based HMR is the lack of high-speed labeled datasets. We need high-speed labels of the human mesh, a diverse set of motions, and precise synchronization between events, images, and ground truth. To address this, we collect Beam-splitter Event Agile Human Motion Dataset (BEAHM), which has the desired properties compared to prior datasets (Tab. 1). We used a single-objective beam splitter with a shared lens to obtain aligned events and images [21]. The ground truth is obtained through four calibrated RGB cameras using a state-of-the-art multi-view human reconstruction technique [1]. To ensure precise timing, we built a custom trigger board to synchronize event camera and 120 FPS RGB cameras. We designed 40 diverse motions from slow walking to fast Karate kicking, covering a wide spectrum of difficulties.

High speed labels. Accurate pose ground truth at a high frame rate is crucial for capturing fast human movements. While simple interpolation of joint rotations and translations can be sufficient for upsampling slow motions, it fails to upsample rapid movements. Our analysis reveals an average error of up to 25 mm for Slerp-interpolated joints, underscoring the need for high-speed labeling.

6. Experiments

In this section, we present a comprehensive evaluation of EvHuman through qualitative and quantitative analyses. We begin by introducing the datasets and metrics used for evaluation (Sec. 6.1), and the baseline methods (Sec. 6.2). We then provide a detailed comparison of pose accuracy on the selected datasets (Sec. 6.3). In addition, we conducted an ablation study to investigate the impact of key design choices in our model (Sec. 6.4) and show variable frame rate decoding unique to our method (Sec. 6.5). Finally, we analyze the computational efficiency of our method in comparison to existing approaches (Sec. 6.6).

6.1. Datasets and Metrics

MMHPSD [74] is chosen for the evaluation of pose precision. It is a recent event-based human dataset with a multi-camera setup with 3D mesh annotations, making it well-suited for our target scenario. It includes a software-synchronized events captured by a CeleX-V event sensor along with grayscale images, with ground-truth annotations provided at 15 FPS.

BEAHM. A detailed description of our data is provided

Table 2. Quantitative comparison on MMHPSD [74] and our BEAHM. DHP19† uses the groundtruth depth for each joint. Entries labeled with * are re-implementations of the original papers.

Models	MPJPE ↓	PA-MPJPE ↓	PEL-MPJPE ↓	PCKh@0.5 ↑	Input Modality
MMHPSD					
HMR [24]	-	64.78	95.32	0.61	Images
HMR2.0 [17]	-	60.63	79.53	0.73	Images
EventCap* [66]	-	62.62	89.95	0.64	Images + Events
EventHPE [74]	71.79	43.90	54.96	0.85	Images + Events
EventHPE [74] (*w/o HMR Feats*)	81.06	45.86	58.90	0.84	Events
DHP19† [8]	72.42†	65.87†	74.04†	0.81†	Events
Ours	**67.66**	**39.16**	**52.23**	**0.86**	Events
BEAHM					
HMR [24]	-	68.73	112.55	0.42	Images
HMR 2.0 [17]	-	52.12	80.19	0.60	Images
EventCap* [66] (*w/ HMR Init*)	-	65.81	93.94	0.70	Images + Events
DHP19†*[8]	46.75†	42.39†	48.15†	0.86†	Events
EventHPE [74] (*w/o HMR Feats*)	65.28	38.91	52.31	0.86	Events
Ours *w/o Fine-tune*	67.01	39.16	52.23	0.86	Events
Ours *w/o* $\mathcal{L}_c$	50.77	30.22	41.26	0.91	Events
Ours	**49.74**	**30.05**	**41.06**	**0.92**	Events

in Sec. 5. For fairness, we follow EventHPE [74] to evaluate pose accuracy for a roughly 1-second window, which corresponds to skipping 16 labeled frames at 120 FPS, with 8 evaluation timestamps in each window. In addition, we provide evaluation at 120 FPS in Tab. 3.

Metrics. We follow the evaluation protocols of previous work to report 3D human joint metrics, including MPJPE, Pelvis-Adjusted MPJPE, Procrustes-Aligned MPJPE for 3D joint positions, and PCKh@0.5 for the percentage of predicted 2D joints that fall within half the head length of the ground-truth 2D joints. Pelvis-Adjusted MPJPE eliminates root transformation, which is commonly used with centered human crops. We also report unadjusted MPJPE to include global translation in our evaluation.

6.2. Baseline Methods

We compare EvHuman with image and event-based baselines. Following EventHPE [74], we include image-based methods **HMR** and **HMR 2.0**, computing per-image predictions and interpolating only when the frame rate is lower than the ground truth (Tab. 3). For event-based methods, we compare with **EventHPE** [74], **DHP19** [8], and **EventCap** [66]. Since EventHPE incorporates HMR features, making it not purely event-based, we re-trained it without HMR features for fair event-based comparison. Additionally, we re-trained **DHP19** [8] on the 24 SMPL joints, lifting its 2D keypoints into 3D using ground-truth depth. **EventCap** [66] has no open-source code, so we reimplemented it and reported the results.

6.3. Comparisons

Comparison on MMHPSD We present quantitative and qualitative comparisons on MMHPSD in Tab. 2 and Fig. 4, respectively. EvHuman surpasses all baseline methods across all metrics, improving PA-MPJPE by 4.8 mm and PEL-MPJPE by 2.7 mm over the second-best method, *EventHPE*, which uses both image and event data. EvHuman outperforms the best image-based method by 21.47 mm in PA-MPJPE. Unlike *EventHPE*, which allows arbitrary poses at each frame time, our model leverages motion priors from the entire sequence, enabling it to predict only plausible natural human motions constrained by the neural human motion prior.

Comparison on BEAHM Since [74] evaluates MMHPSD at 15 Hz, we perform quantitative comparisons at the same frame rate in Tab. 2. EvHuman improves MPJPE by 15.54 mm, PA-MPJPE by 8.86 mm and 0.07 in PCKh@0.5, over EventHPE. Additionally, EvHuman outperforms the best image-based method by 21.07 mm in PA-MPJPE and 0.32 in PCKh@0.5. HMR 2.0, limited by its per-frame processing, produces an erroneous mesh, as illustrated in Fig. 4. Leveraging the high-speed labels of BEAHM, we provide a comparison at 120 Hz in Tab. 3. Our method improves the absolute joint errors over the best event-based baseline in MPJPE by 17 mm, PA-MPJPE by 5.9 mm , and PCKh@0.5 by 0.03. During training, EventHPE showed instability in translation estimation, which caused significant drifts in joint errors.

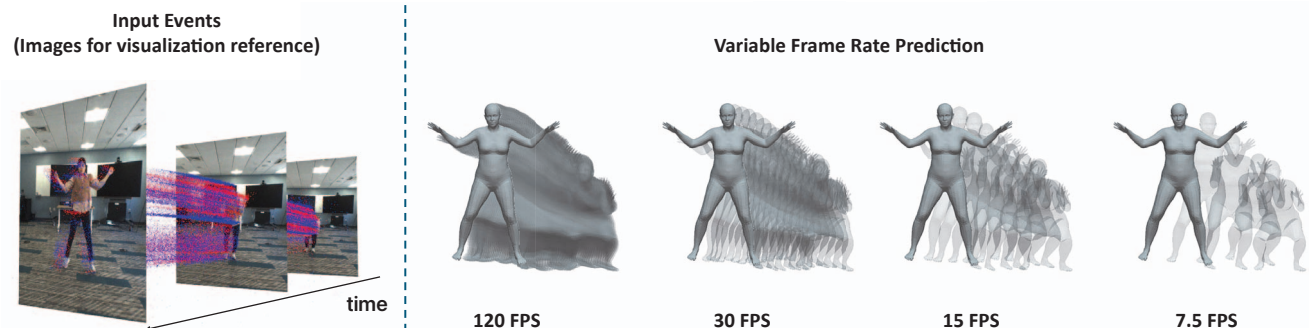

Figure 6. **Left:** Events triggered by human motion. Aligned RGB images shown **only for visualization**. **Right:** Human motion prediction at variable frame rates without re-training the network.

6.4. Ablation

We performed two ablation studies on the proposed component on our BEAHM dataset in Tab. 2. *w/o* $\mathcal{L}_c$ refers to model trained without the human mesh event contrast maximization in Sec. 4.4. *w/o Fine-tune* means we do not fine-tune the global motion predictor function (GMP), as described in Sec. 4.5. The global motion predictor fine-tuning at the end significantly boosts the global translation performance of the model, by improving the MPJPE by 17.27 mm, PA-MPJPE by 8.11 mm, PEL-MPJPE by 11.17 mm and 0.06 in PCKh@0.5. Vertex flow contrast maximization improves the global and local MPJPE performance.

6.5. Variable Frame Rate Prediction

Due to the continuous-time nature of our motion parameterization, human motion can be predicted at any time resolution **without retraining**. A series of meshes can be predicted simply by inputting the desired timestamps within the events. Prediction at a high frame rate is computationally efficient as we do run full inference for every frame. Qualitative results are shown in Fig. 6.

Table 3. Evaluation at 120 FPS on BEAHM. DHP19† uses the groundtruth depth for each joint. We include their upper-bound performance but exclude their performance from ranking.

Models	MPJPE	PA-MPJPE	PEL-MPJPE	PCKh@0.5
HMR 2.0 [17]	-	56.23	93.08	0.51
DHP19† [8]	46.15†	41.34†	47.23†	0.86†
EventHPE [74]	66.76	35.97	48.24	0.89
Ours	**49.76**	**30.07**	**41.08**	**0.92**

6.6. Computational Speed

Due to the time continuity of our model, we can sample an arbitrary number of poses in parallel. An analysis of the computation time is provided in Fig. 7. For HMR 2.0, we present two settings: with video frame interpolation (*HMR 2.0 + Upsampled)* and with raw high-speed input *(HMR 2.0 + High FPS*). For *HMR 2.0 + Upsample*, we upsample frames to the desired frame rate using FILM [49]. Our event baseline, EventHPE, predicts 8 frames at a time. As shown in Fig. 7, inference speeds are similar across methods at the training sampling rate (8 frames). However, the computational time for all three baseline methods increases significantly as the prediction length grows because the full network runs at each frame. All evaluations are conducted with an RTX 4080 deskptop GPU.

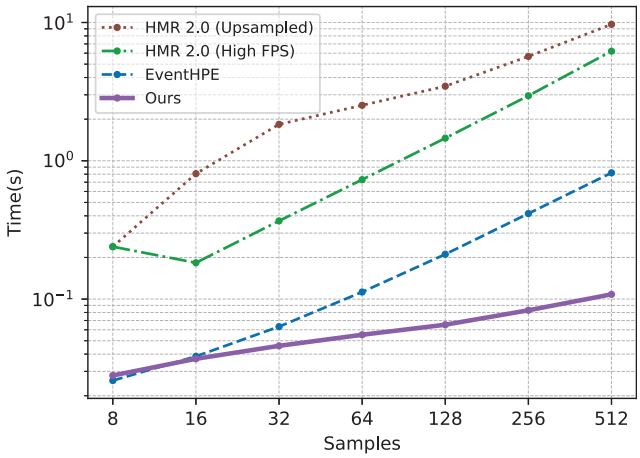

Figure 7. **Computational Speed Comparison.** The computational time is plotted against the number of prediction frames.

7. Conclusion and Discussion

This work introduces the first method that estimates a continuous human motion field from events. Leveraging a learned motion-prior latent space and an implicit motion decoder, our method allows for fast arallel inference at arbitrary temporal resolutions. The proposed method outperforms state-of-the-art event-based human pose methods while achieving 69 % faster inference. We also contribute a hardware synchronized event-based human mesh dataset with high temporal resolution labeling, opening up valuable research opportunities in studying event-based human motion estimation. Despite its strengths, the method has limitations, including reliance on voxelized events and initial pose estimation. Future efforts will focus on reducing latency and eliminating the need for initialization.

Acknowledgment

We gratefully acknowledge support by the following grants: NSF FRR 2220868, NSF IIS-RI 2212433, ARO MURI W911NF-20-1-0080, and ONR N00014-22-1-2677.

References

[1] Easymocap - make human motion capture easier. Github, 2021. 6, 2, 4

[2] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62–1, 2020. 1

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 2

[4] R Wes Baldwin, Ruixu Liu, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Time-ordered recent event (tore) volumes for event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2519–2532, 2022. 2

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

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[7] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192, 2000. 3

[8] Enrico Calabrese, Gemma Taverni, Christopher Awai Easthope, Sophie Skriabine, Federico Corradi, Luca Longinotti, Kynan Eng, and Tobi Delbruck. Dhp19: Dynamic vision sensor 3d human pose dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1, 2, 4, 5, 7, 8

[9] Jiaan Chen, Hao Shi, Yaozu Ye, Kailun Yang, Lei Sun, and Kaiwei Wang. Efficient human pose estimation via 3d event point cloud. In *2022 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2022. 2

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 4

[11] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7792–7801, 2019. 2

[12] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286. IEEE, 2013. 2

[13] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3867–3876, 2018. 2, 3

[14] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12280–12289, 2019. 3, 4, 5

[15] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. *Asynchronous, Photometric Feature Tracking Using Events and Frames*, page 766–781. Springer International Publishing, 2018. 5

[16] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 197–206. IEEE, 2021. 2

[17] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 2, 7, 8

[18] Friedhelm Hamann, Ziyun Wang, Ioannis Asmanis, Kenneth Chaney, Guillermo Gallego, and Kostas Daniilidis. Motion-prior contrast maximization for dense continuous-time motion estimation. In *European Conference on Computer Vision*, pages 18–37. Springer, 2025. 2, 4

[19] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. *Advances in Neural Information Processing Systems*, 35:4244–4256, 2022. 2, 3

[20] Dorian F. Henning, Christopher Choi, Simon Schaefer, and Stefan Leutenegger. Bodyslam++: Fast and tightly-coupled visual-inertial camera and human motion tracking. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3781–3788, 2023. 2

[21] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2022. 6

[22] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 3

[23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2, 5

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and

pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[25] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14632–14643, 2023. 2

[26] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2

[27] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. In *2024 International Conference on 3D Vision (3DV)*, pages 397–408. IEEE, 2024. 2, 3

[28] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 2, 5

[29] Taku Komura, Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, and Joe Yearsley. A recurrent variational autoencoder for human motion synthesis. In *The 28th British Machine Vision Conference*, 2017. 3

[30] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 4

[31] Yan Li, Tianshu Wang, and Heung-Yeung Shum. Motion texture: a two-level statistical model for character motion synthesis. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 465–472, 2002. 3

[32] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020. 3

[33] C Karen Liu, Aaron Hertzmann, and Zoran Popović. Learning physics-based motion style with nonlinear inverse optimization. *ACM Transactions on Graphics (TOG)*, 24(3): 1071–1081, 2005. 3

[34] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding. In *2021 international conference on 3D vision (3DV)*, pages 930–939. IEEE, 2021. 2

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. *SMPL: A Skinned Multi-Person Linear Model*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 2

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3, 4

[37] Weng Fei Low, Zhi Gao, Cheng Xiang, and Bharath Ramesh. Sofea: A non-iterative and robust optical flow estimation algorithm for dynamic vision sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 82–83, 2020. 2

[38] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 5

[39] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 3

[40] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[42] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[43] Manasi Muglikar, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza. How to calibrate your event camera. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1403–1409, 2021. 2

[44] Liyuan Pan, Miaomiao Liu, and Richard Hartley. Single image optical flow estimation with an event camera. in 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1669–1678. 2

[45] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3

[46] Priyanka Patel and Michael J. Black. Camerahmr: Aligning people with perspective, 2024. 2

[47] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 2

[48] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[49] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pages 250–266. Springer, 2022. 8

[50] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11488–11499, 2021. 3

[51] Charles Rose, Michael F Cohen, and Bobby Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics and Applications*, 18(5):32–40, 1998. 3

[52] Nitin Saini, Chun-Hao P Huang, Michael J Black, and Aamir Ahmad. Smartmocap: Joint estimation of human and camera motion using uncalibrated rgb cameras. *IEEE Robotics and Automation Letters*, 2023. 2

[53] Gianluca Scarpellini, Pietro Morerio, and Alessio Del Bue. Lifting monocular events to 3d human poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1358–1368, 2021. 2

[54] Zhanpeng Shao, Xueping Wang, Wen Zhou, Wuzhen Wang, Jianyu Yang, and Youfu Li. A temporal densely connected recurrent network for event-based human pose estimation. *Pattern Recognition*, 147:110048, 2024. 5

[55] Yunzhou Song, Jiahui Lei, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Track everything everywhere fast and robustly. In *European Conference on Computer Vision*, pages 343–359. Springer, 2024. 5

[56] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7244–7253, 2019. 2

[57] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[58] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 2

[59] Bugra Tekin, Pablo Marquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[60] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[61] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2024. 2

[62] Ziyun Wang, Kenneth Chaney, and Kostas Daniilidis. EvAC3D: From event-based apparent contours to 3D models via continuous visual hulls. In *ECCV*, pages 284–299, 2022. 3

[63] Ziyun Wang, Fernando Cladera, Anthony Bisulco, Daewon Lee, Camillo J Taylor, Kostas Daniilidis, M Ani Hsieh, Daniel D Lee, and Volkan Isler. EV-Catcher: High-speed object catching using low-latency event-based neural networks. 7(4):8737–8744, 2022. 2

[64] Ziyun Wang, Jinyuan Guo, and Kostas Daniilidis. UnEVIMO: Unsupervised event-based independent motion segmentation. In *ECCV*, pages 228–245, 2024.

[65] Ziyun Wang, Friedhelm Hamann, Kenneth Chaney, Wen Jiang, Guillermo Gallego, and Kostas Daniilidis. Event-based continuous color video decompression from single frames. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4968–4978, 2025. 2

[66] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4968–4978, 2020. 1, 2, 3, 5, 7, 6

[67] Chengxi Ye, Anton Mitrokhin, Cornelia Fermüller, James A Yorke, and Yiannis Aloimonos. Unsupervised learning of dense optical flow, depth and egomotion from sparse event data. *arXiv preprint arXiv:1809.08625*, 2018. 2, 4

[68] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21222–21232, 2023. 2

[69] Zelin Zhang, Anthony J Yezzi, and Guillermo Gallego. Formulating event-based image reconstruction as a linear inverse problem with deep regularization using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8372–8389, 2022. 2

[70] Alex Zihao Zhu, Yibo Chen, and Kostas Daniilidis. Real-time time synchronized event-based stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 433–447, 2018. 2

[71] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018. 2, 3, 5

[72] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 2, 3, 4

[73] Alex Zihao Zhu, Ziyun Wang, Kaung Khant, and Kostas Daniilidis. Eventgan: Leveraging large scale image datasets for event cameras. In *2021 IEEE international conference on computational photography (ICCP)*, pages 1–11. IEEE, 2021. 1

[74] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10996–11005, 2021. 1, 2, 3, 5, 6, 7, 8

[75] Shihao Zou, Yuxuan Mu, Xinxin Zuo, Sen Wang, and Li Cheng. Event-based human pose tracking by spiking spatiotemporal transformer. *arXiv preprint arXiv:2303.09681*, 2023. 1, 2