3D Human Scan With A Moving Event Camera

Kai Kohyama
kaikohyama@keio.jp

Shintaro Shiba
sshiba@keio.jp

Yoshimitsu Aoki
aoki@elec.keio.ac.jp

Keio University

Abstract

Capturing the 3D human body is one of the important tasks in computer vision with a wide range of applications such as virtual reality and sports analysis. However, conventional frame cameras are limited by their temporal resolution and dynamic range, which imposes constraints in real-world application setups. Event cameras have the advantages of high temporal resolution and high dynamic range (HDR), but the development of event-based methods is necessary to handle data with different characteristics. This paper proposes a novel event-based method for 3D pose estimation and human mesh recovery. Prior work on event-based human mesh recovery require frames (images) as well as event data. The proposed method solely relies on events; it carves 3D voxels by moving the event camera around a stationary body, reconstructs the human pose and mesh by attenuated rays, and fit statistical body models, preserving high-frequency details. The experimental results show that the proposed method outperforms conventional frame-based methods in the estimation accuracy of both pose and body mesh. We also demonstrate results in challenging situations where other frame-based methods suffer from motion blur. This is the first-of-its-kind to demonstrate event-only human mesh recovery, and we hope that it is the first step toward achieving robust and accurate 3D human body scanning from vision sensors.

1. Introduction

Estimating human pose from cameras is one of the key computer vision challenges with a wide range of applications such as virtual reality (VR), sports analysis, and abnormal behavior detection. Recently, many pose estimation methods using deep neural networks (DNNs) have been proposed and developed for images from conventional frame cameras. However, such methods have limitations in the applicable scenes that inherits the constraints of frame cameras by nature: the temporal resolution is insufficient for intense motion such as during sports (i.e., motion blur), and the shutter speed must be adjusted to obtain data in dark scenes (i.e., limited dynamic range). To address these challenges, event cameras have recently gained attention from both research and industry. Unlike conventional frame cameras where all pixels record data synchronously, event cameras asynchronously respond only to brightness changes, resulting in high temporal resolution (μsec order) and high dynamic range (HDR). However, since event cameras produce different data from frames [10, 21], it is paramount to develop event-based methods for pose estimation and mesh recovery using event cameras. Previous research on human body mesh reconstruction using event cameras [51, 55] require frame images as complementary information, and could not achieve reconstruction from events alone. This is because event cameras do not generate data for stationary parts of the body if the camera is static (i.e., no event data observed). Frame cameras are better for capturing such static scene information, however, using frame images is problematic since it imposes the limitations of frames (e.g., dynamic range, motion blur). We summarize existing frame-based and event-based methods in Tab. 1.

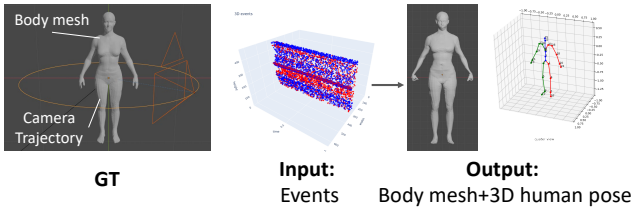

Figure 1. Summary of the proposed method. Our method reconstructs the human body mesh and estimates the pose only from an event camera that moves around the body.

In this work, we propose a 3D human body scanning method capable of estimating the 3D voxel representation, mesh, joints, and body model parameters of a stationary human body using only event data (Fig. 1). Our proposed method enables data acquisition from static bodies by moving the camera itself. Furthermore, we propose a ray attenuation to better preserve high-frequency detail information, by extending the existing event-based voxel carving method [50]. Finally, by fitting the statistical human body models,

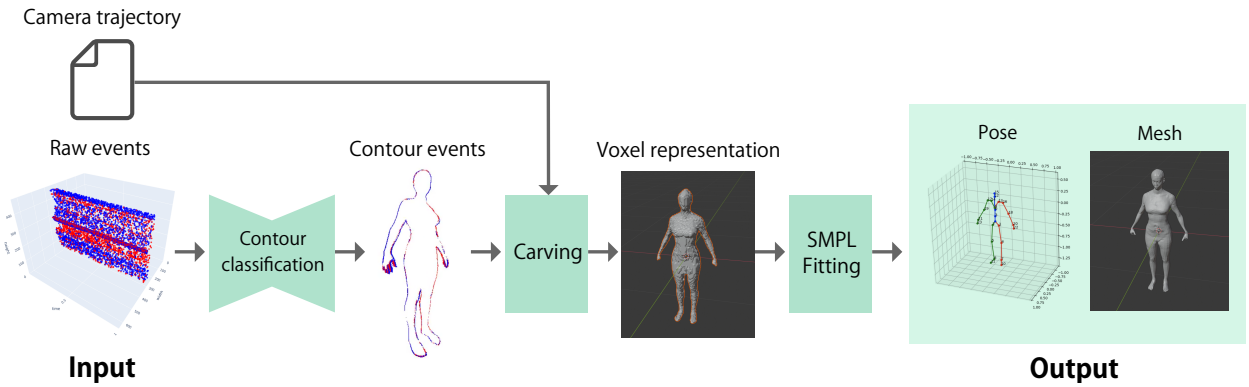

Figure 2. Overview of the proposed method.

Table 1. Comparison of (some) existing methods for frame-based and event-based human mesh reconstruction. Scene and Camera can be either dynamic ("D") or static ("S"). Existing event-based methods require images ("I") as well as events ("E"), resulting in mitigating motion blur to some extent (denoted with †).

	Scene	Camera	Data	Motion blur
PyMAF [53]	D	S	I	Yes
PyMAF-X [54]	D	S	I	Yes
EasyMoCap [53]	D	S	I	Yes
EventCap [51]	D	S	E + I	No†
EventHPE [55]	D	S	E + I	No†
Ours	S	D	**E**	**No**

such as SMPL [27] and SKEL [15], we demonstrate accurate estimation of the human body mesh, body parameters, and joint positions from the voxels.

To summarize, our contributions are as follows:

- The first-of-its-kind method to estimate human pose and body mesh from only events.
- A ray attenuation for preserving details better, compared with existing event-based carving.
- A thorough benchmarking with frame-based methods and ablation studies on different illuminations, camera speeds, hyper-parameters, and statistic body models.

The experimental results indicate that the proposed method achieves accurate human body mesh reconstruction and 3D pose estimation from events alone, without the need for additional frames. Leveraging the high temporal resolution of event cameras, our method achieves precise carving surpassing frame-based pose estimation methods.

2. Related Work

2.1. Frame-based Human Pose and Mesh Estimation

3D human pose estimation from frame-based cameras, which involves inferring the positions of human joints, is one of the popular tasks in computer vision. Most of the recent work utilize deep learning in various manners, such as direct regression of 3D joints from images [24, 35, 46], triangulating two-dimensional pose estimation results into the 3D space [5, 23, 48], and estimating joint positions using heatmaps after converting the human body into a 3D representation [32, 36, 47]. These approaches also consist of a wide range of problem settings, from using a single camera [5, 35, 36] to employing multiple cameras [1, 7, 44, 47], and utilizing depth sensors as additional information [32, 52].

While traditional pose estimation focuses solely on the joint positions, recent work have addressed human body *mesh* reconstruction [14, 20, 53, 54]. In the mesh reconstruction, not only the pose parameters but also the ones related to the body shape are estimated simultaneously to obtain the body mesh corresponding to the input image. Meshes are generated by parametric statistical human body models, such as SMPL [27], SMPL-X that extends SMPL for hands and faces [37], SKEL that considers anatomical structures [15], and CAPE that includes clothes [29].

2.2. Event-based Human Pose and Mesh Estimation

Human pose and mesh estimation using event cameras are relatively recent research fields due to the novelty of event cameras [10]. Common approaches accumulate events over a certain time interval, convert them into images, and feed them into convolutional neural networks (CNNs) to obtain the joint heatmaps [4, 39, 55]. Also, Chen et al. [6] utilize point-cloud neural networks to directly process events, reducing memory consumption and computational complexity.

Event-based human mesh reconstruction methods have also been studied, such as the approach by Xu et al. [51] that achieves 1000 fps human body capture in low-light conditions. However, it requires inputs from both events and frames, as well as prior scanning of the subject to realize accurate reconstruction. Zou et al. [55] propose an approach using optical flow from events to estimate both pose and body mesh, yet rely on frames for initialization. Event-based hand mesh reconstruction method [13] also uses the complementary frame data, and to the best of the authors' knowledge, no work has shown the mesh recovery solely from events. Our proposed method utilizes only event data (see Tab. 1) to simultaneously estimate the 3D joint positions and reconstruct parametric human body models, achieving precise estimation.

2.3. Visual Hull

Visual Hull, or carving, is a technique to reconstruct three-dimensional shapes by finely divided voxels, which are cubes carved from various angles based on the contours of objects extracted from images [22]. In carving, a smoother reconstruction is achieved by increasing viewpoints. Increasing the number of voxel subdivisions allows for finer detail representation, however, it requires more rays for sufficient carving, leading to larger data size for the three-dimensional representation. Moreover, due to its principle, high-frequency detailed information is difficult to extract contours, and concave surfaces cannot be reconstructed. Event cameras, due to their higher temporal resolution compared to traditional frame cameras, enable smoother voxel carving with more continuous viewpoint changes [50]. It demonstrates reconstructing smooth voxels for 3D reconstruction of simple objects like cans using an event camera.

3. Method

In this section, we propose a method to estimate the 3D voxel representation, mesh, and parameters of joints (pose) and body shape for a stationary human body from a moving event camera. Figure 2 shows the overview of the proposed method. It consists of three steps: (i) classifying "contour" events among the raw events (Sec. 3.2), (ii) carving voxels based on the contour events (Sec. 3.3), and (iii) fitting statistical body models on the carved voxels (Sec. 3.4). Our proposed pipeline is model-agnostic, as shown in Sec. 4.3, and achieves accurate estimation thanks to the high temporal resolution.

3.1. Event Cameras

Event cameras (e.g., the Dynamic Vision Sensor (DVS) [8, 25, 45]) are novel vision sensors that respond to intensity changes. There have been emerging computer vision research for event cameras, such as object detection [11, 31, 33], ego-motion estimation [9, 40], optical flow [3, 41, 42], SLAM [12, 17, 18], and many applications [2, 26, 34, 43]. They have independent pixels that operate continuously and generate "events" $e_k \doteq (x_k, y_k, t_k, p_k)$ whenever the logarithmic brightness at the pixel increases or decreases by a predefined sensitivity C:

$$L\left(x_k, y_k, t_k\right) - L\left(x_k, y_k, t_k - \Delta t_k\right) = p_k C. \quad (1)$$

Each event contains the space-time coordinates (x_k, y_k, t_k) of the brightness change and its polarity $p_k = \{+1, -1\}$, with the elapsed time Δt_k since the previous event.

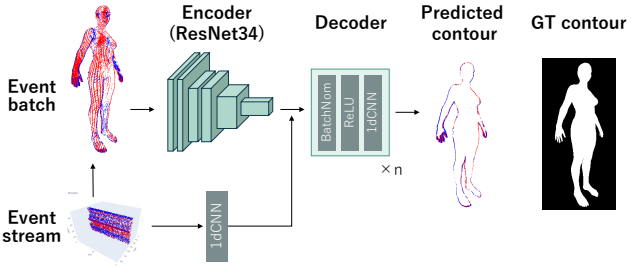

Figure 3. Contour classification network.

3.2. Contour Classification

Since events occur due to brightness changes, they do so not only at object contours but at any edges in the image plane, such as textures, assuming the scene brightness is constant. Hence, it is necessary to classify events derived from contours rather than other events for carving objects. For the contour classification, we utilize a convolutional neural network (CNN) encoder-decoder model (Fig. 3) [49]. As the model input, we collect the most recent $10,000$ events and use a voxel representation (three-dimensional tensor) by discretizing position and time.

$$\mathrm{C}_i(x, y, t) = p_i k_b(x - x_i) k_b(y - y_i) k_b(t - t_i), \quad (2)$$

where k_b is a bilinear kernel, allocating event coordinates (x, y, t) to each bin (x_i, y_i, t_i). The bilinear voting enables the model to learn based on the temporal and spatial relationships between events.

The model is trained in a supervised manner, minimizing binary cross-entropy loss between the ground truth (GT) labels $\hat{q}_i$ from mask and the contour inference result $f_\theta(e_i, C_i)$ for the i-th event:

$$\mathcal{L}_c = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{bce}(f_\theta(e_i, C_i), \hat{q}_i). \quad (3)$$

As results of the training, as shown in Fig. 3, we can extract human body contours from the original event data (i.e., event stream).

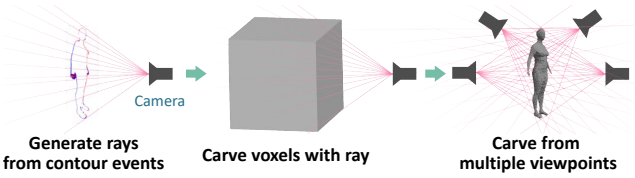

Figure 4. Carving.

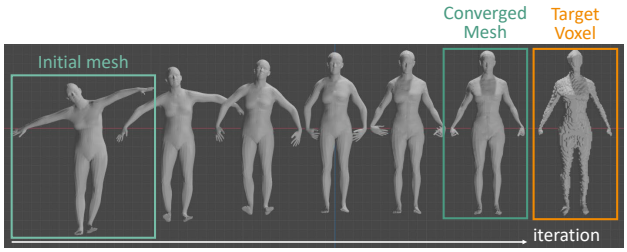

Figure 5. SMPL fitting.

3.3. Voxel Carving

We carve voxels using the contour events extracted in the previous step (Fig. 4). First, we convert the image coordinates (x, y) of individual events to the world coordinates (X_W, Y_W, Z_W) via the camera coordinate system (X_C, Y_C, Z_C), using the camera intrinsic parameters K, the rotation matrix R, and the translation T obtained from the input camera trajectory.

$$\begin{bmatrix} X_W \\ Y_W \\ Z_W \end{bmatrix} = R^{-1} \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} - T = sR^{-1}K^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} - T. \quad (4)$$

The conversion from the two-dimensional coordinates (x, y) to the three-dimensional coordinates (X_W, Y_W, Z_W) is a back-projection, and the contours in the world coordinates are represented as rays that contain an unknown scale variable s (4), i.e., lines from the camera origin.

Furthermore, we propose a ray attenuation that is inversely proportional to the distance from the camera to mitigate pixel quantization errors. The accuracy of the contour rays depends on the pixel pitch, and the error is proportional to the distance of the camera rays (the error increases by n when the distance from the camera is increased by n). We find that attenuating the rays inversely proportional to the distance enables preserving the details of the carved voxels. Now the rays' influence r' during carving becomes

$$\mathrm{r}' = \frac{\mathrm{r}}{d+1}, \quad (5)$$

with the distance from the camera d and the original influence r. The number of rays (intensity) passing through each voxel is weighted using r', and voxels above a threshold are pruned. We find that event-based carving leverages the high temporal resolution of event data, enabling smooth 3D reconstruction.

3.4. SMPL Fitting

The voxels obtained through carving do not reproduce fine structures, such as fingertips or concave surfaces like faces or navels. Therefore, we perform mesh reconstruction by fitting statistical body models, SMPL [27]. Notice that the proposed framework can also be used for different models, such as SKEL [15] as shown in Sec. 5.2.

The SMPL model defines a function W that outputs a human body mesh $M(\beta, \theta)$ with 6890 vertices:

$$\begin{aligned} M(\beta, \theta) &= W(T_P(\beta, \theta), J(\beta), \theta, \mathcal{W}), \\ T_P(\beta, \theta) &= \bar{T} + B_S(\beta) + B_P(\theta). \end{aligned} \quad (6)$$

The inputs of the model are the body shape parameter $\beta \in \mathbb{R}^{10}$ (v1.0) or $\beta \in \mathbb{R}^{300}$ (v1.1) and the pose parameters $\theta \in \mathbb{R}^{72}$ that consists of 3 degrees of freedom (rotational angles) for 24 joints. Here, $J(\beta)$ represents the joint positions considering the body shape parameters, and $\mathcal{W}$ represents the correspondence between vertices and joints. Additionally, $T_P(\beta, \theta)$ is a human body mesh that incorporates variations in body shape and deformation of the flesh according to posture, created from the template mesh $\bar{T}$.

We fit the SMPL parameters θ and β to the mesh that is obtained by applying the marching cubes method [28] on the carved voxels. The fitting becomes a process of minimizing the Chamfer loss, i.e.,

$$\begin{aligned} \hat{\theta}, \hat{\beta} &= \arg\min_{\theta, \beta} (\mathcal{L}_c), \\ \mathcal{L}_c &= \sum_p \min_q ||p - q||_2^2 + \sum_q \min_p ||p - q||_2^2. \end{aligned} \quad (7)$$

Here, p and q are $50,000$ points randomly sampled from the surfaces of the SMPL model and the carving result mesh, respectively, and we calculate the sum of squared Euclidean distances between the nearest q and p for all pairs. We use the Adam optimizer [19] with a learning rate of 0.01 and 1000 iterations, and finally obtain the estimation of the body parameters $(\hat{\theta}, \hat{\beta})$.

4. Experiments

In this section, we first present the dataset used in the experiments in Sec. 4.1, explain the evaluation metrics and the frame-based methods used to benchmark in Sec. 4.2, and discuss the results in Sec. 4.3 comparing other baselines. Finally, we discuss the advantages of the proposed method under motion blur in Sec. 4.4.

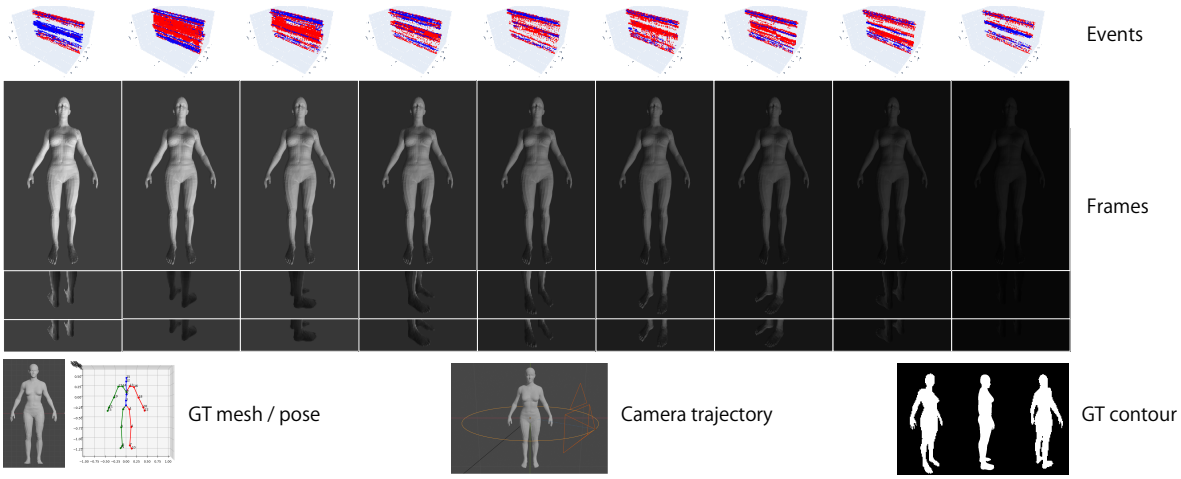

Figure 6. Dataset.

4.1. Dataset

The proposed problem setting of moving event cameras for human body mesh reconstruction is, to the best of our knowledge, the first attempt. As there are no datasets available for such problem settings, we create a new dataset using the event camera simulator (ESIM) [38]. ESIM generates raw event stream from object 3D model, camera intrinsic parameters, and camera trajectory. In our case, the human mesh is represented by a .obj file with the joint ground truth (GT) data created beforehand from the SMPL model, and the camera trajectory follows a path circling the subject twice at a radius of 1 m. The camera is modeled as a pinhole camera, with parameters of $(\text{width}, \text{height}, f_x, f_y, c_x, c_y) = (640, 480, 250, 250, 320, 240)$. We also generate 2082 contour GT data with timestamps for training the contour classification network, by binarizing the depth GT from ESIM. Additionally, we save frame images at 30 fps for the baselines. Example data in the dataset are shown in Fig. 6. In total, we prepare 27 sequences that consist of three conditions (two different poses and one motion-blur sequence) with nine illumination settings. Each sequence contains approximately 20 million events and 518 frame images.

4.2. Evaluation Metrics and Baselines

PEL-MPJPE. As quantitative evaluation metrics for joints, we use the Pelvis-Aligned Mean Per Joint Position Error (PEL-MPJPE), used in human pose estimation work [20, 55]. MPJPE represents the average 3D Euclidean distance between the estimated joint positions and the ground truth joint positions, such as

$$\mathbf{MPJPE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \|p_{\hat{x}}(i) - p_x(i)\|, \qquad (8)$$

where N denotes the total number of joints, p_x is the joint position of the GT mesh, and $p_{\hat{x}}$ is the estimated joint positions. PEL-MPJPE evaluates the accuracy of relative joint positions with respect to the root (pelvis) position after aligning them.

Chamfer Distance. For the evaluation of body meshes, we use the Chamfer Distance (CD):

$$\begin{aligned} \mathbf{CD}(X, \hat{X}) = & \frac{1}{|X|} \sum_{x \in X} \min_{\hat{x} \in \hat{X}} \|x - \hat{x}\|_2^2 \\ & + \frac{1}{\hat{X}} \sum_{\hat{x} \in \hat{X}} \min_{x \in X} \|x - \hat{x}\|_2^2. \end{aligned} \qquad (9)$$

Similar to the Chamfer Loss used in Sec. 3.4 (7), CD measures the similarity of meshes based on the 3D Euclidean distance of sampled points from the mesh surfaces. Here, X is the point cloud sampled from the GT mesh, and $\hat{X}$ is the point cloud sampled from the estimated mesh. While in Chamfer Loss (Sec. 3.4) during fitting the body model we sample 50k points, here we sample 1M points from the GT mesh and the recovered mesh surface for evaluation. Both MPJPE and CD are averaged over the results from the 9 sequences with different lighting conditions in the dataset.

Baselines. Existing methods of event-based human body mesh reconstruction track moving humans with stationary event cameras and require additional frame images. Hence, it is challenging to directly compare them with the proposed method, where the camera moves around stationary human bodies, and which does not rely on frames. Therefore, as baselines, we use three frame-based methods: PyMAF [53], its extension PyMAF-X [54], and another state-of-the-art EasyMoCap [1]. PyMAF and PyMAF-X regress the SMPL parameters that fit a single RGB image, and EasyMoCap fits the SMPL model from multiple cameras (viewpoints).

Table 2. Quantitative evaluation results.

		PEL-MPJPE [mm] ↓		CD [mm] ↓	
		Pose1	Pose2	Pose1	Pose2
Ours (Event-based)		**58.11**	**65.64**	**7.589**	**13.36**
EasyMocap [1]		59.54	97.29	36.18	30.09
PyMAF [53]	single	78.40	234.1	49.83	239.8
	multi	63.90	185.4	30.19	193.6
	all	64.09	195.2	29.82	203.5
PyMAF-X [54]	single	92.10	182.9	48.68	199.1
	multi	82.49	109.1	22.54	61.49
	all	81.95	108.6	22.16	60.12

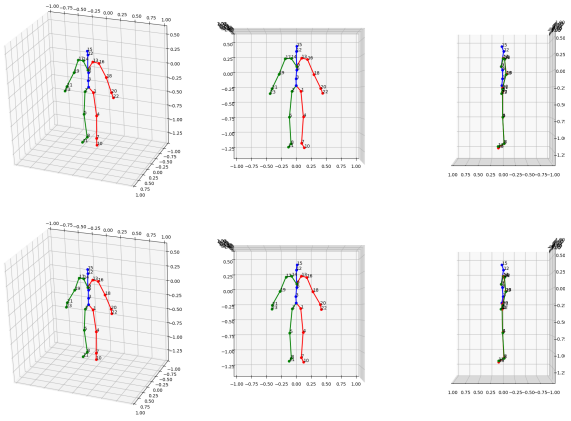

Figure 7. Qualitative results of the pose estimation for GT (top) and the proposed method (bottom).

For PyMAF and PyMAF-X, to fairly compare with the proposed method using multi-view events, we compare the averaged outputs when using 8 frames with viewpoints changing at 45° intervals (*multi-image*) and when using all 518 frame images output at 30 fps from ESIM (*all-image*).

4.3. Comparison with Frame-based Methods

Table 2 shows the quantitative results compared with the frame-based baselines. Our proposed method achieves the smallest errors in MPJPE and CD for both "Pose1" (A-pose), and "Pose2" (T-like pose). In particular, we observe significant improvements compared with PyMAF and PyMAF-X. Although these baselines are proposed as a single-view–estimation method, it is remarkable that the proposed event-based approach realizes 30–60% improvement in MPJPE and 80–90% improvement in CD. On the other hand, EasyMoCap results in competitive accuracy for "Pose1", while its results for "Pose2" are not as good. Nevertheless, our proposed method consistently achieves low errors (55–65 mm for MPJPE and 5–15 mm for CD) for

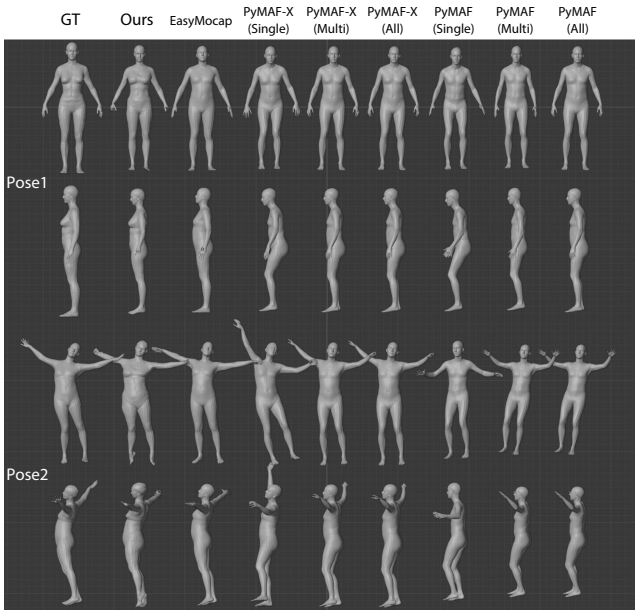

Figure 8. Qualitative comparison of the mesh reconstruction, from the front (top) and side (bottom).

both poses. These results demonstrate the effectiveness of the proposed carving and SMPL fitting combination for different poses and different illumination conditions.

The qualitative results are shown in Fig. 7 (poses) and in Fig. 8 (meshes). The proposed method can estimate 3D poses without major discrepancies, compared with the GT skeleton (Fig. 7). Furthermore, the reconstructed meshes (Fig. 8) have significantly better lateral views, compared to frame-based methods. However, the reconstructed meshes of the proposed method exhibit minor structural distortions in fingertips and toes, and the overall mesh is slightly smaller. This is attributed to missing detailed information during carving due to sampling errors and event noises. We discuss in detail the effect of such missing structure, together with the effect of ray attenuation in Sec. 5.

4.4. Results on Motion-Blur Sequences

One remarkable advantage of event cameras is their minimal motion blur. Therefore, we evaluate the accuracy by moving the camera at ten times faster, where the frames suffer from blurry images (Fig. 9). Table 3 reports the quantitative results. Compared with the results without any blur (Tab. 2), existing frame-based methods have worse results in both pose and mesh accuracy due to the blur. In contrast, the proposed method manages to mitigate the blur effect, which clearly shows the efficacy of event cameras for such rapid motion sequences.

Table 3. Results on the motion-blur sequence. The frame-based methods suffer from motion blur, resulting in the accuracy drop from Tab. 2.

	PEL-MPJPE [mm] ↓	CD [mm] ↓
Ours	**69.68**	**14.33**
EasyMocap [1]	223.6	121.0
PyMAF [53]	216.3	223.3
PyMAF-X [54]	148.6	109.2

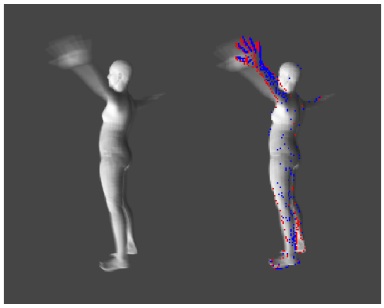

Figure 9. Visualization of the motion-blur sequence. (left) frames suffer from blur, while (right) events do not.

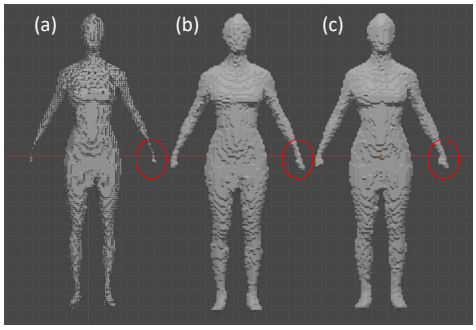

Figure 10. Effect of the ray attenuation. (a) No attenuation. (b) Linear attenuation. (c) Inverse attenuation (5).

Table 4. Effect of the ray attenuation.

	PEL-MPJPE [mm] ↓	CD [mm] ↓
No attenuation	56.90	144.6
Linear	38.75	7.601
Inverse (5)	**38.24**	**6.692**

5. Ablation

5.1. Effect of Ray Attenuation

To validate the effectiveness of the proposed ray attenuation (5), we compare the voxels of carving results with (a) no attenuation: $\mathrm{r}' = \mathrm{r}$, (b) linear: $\mathrm{r}' = \max(\mathrm{r} - d)$, and (c) inverse (5). Figure 10 shows the qualitative comparison among the three. In cases with ray attenuation (b),(c), carving loss of details in fingertips and toes is mitigated as expected. Quantitative evaluations of SMPL fitting on these voxels are presented in Tab. 4, demonstrating the efficacy of the proposed decay for both pose estimation and mesh reconstruction accuracies.

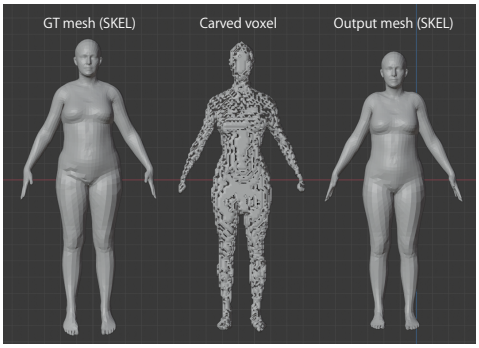

Figure 11. Qualitative results on the SKEL[15] model.

5.2. Results on Other Parametric Model

The proposed method is not limited to fitting the SMPL model. To this end, we conduct experiments using SKEL [15] for ground truth mesh and fitting. As shown in Fig. 11, the estimated meshes look reasonably similar to the GT, with the low error of MPJPE: 54.80 and CD: 7.943.

5.3. Comparison with GT masks

In addition to the frame-based baselines, we conduct another experiment that confirms the advantage of high temporal resolution of event cameras. To this end, we use the GT contour (Figs. 3 and 6) at 30 fps, replacing the contour classification network in the proposed pipeline. The contour edges obtained from GT are used for the same carving and SMPL fitting steps. As results, the mesh estimated from the GT contour achieves slightly better pose estimation (MPJPE: 40.54), while the proposed method achieves better mesh reconstruction (CD: 14.74). The quality of the carved voxel is also confirmed in Fig. 12.

6. Sensitivity Study

Finally, we conduct sensitivity studies for the proposed method. Here, we compare the results when (i) changing the dimensions of body shape parameters β in SMPL (300, 10), and (ii) changing the voxel size (i.e., the numbers of voxel divisions) during carving (512, 256, 128). Larger values of β can fit finer details, however, they are prone to overfitting on excessively carved voxels. Increasing the number of voxels allows for better representation of detailed shapes, while it requires more rays, which can result in uncarved

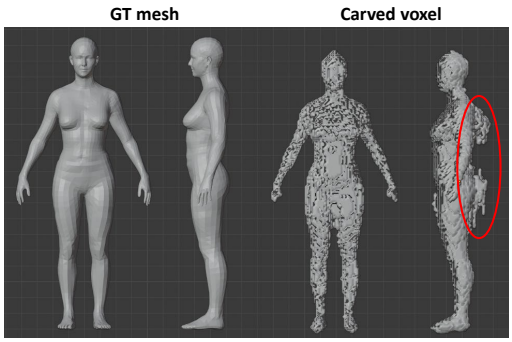

Figure 12. Qualitative results of carving with GT masks, from the front and side.

Table 5. Sensitivity study on the dimension of β and the voxel size. $*$ denotes the average without outliers (when the reconstruction fails), and † denotes the average with such outliers.

β dim	Voxel size	PEL-MPJPE [mm] ↓	CD [mm] ↓
300	512$*$	61.80	7.894
	512†	114.4	81.83
	256	71.90	10.38
	128	142.0	36.97
10	512$*$	55.42	6.800
	512†	111.1	80.93
	256	**58.11**	**7.589**
	128	151.3	36.38

voxels remaining. The quantitative results in Tab. 5 show that both joint estimation accuracy and mesh reconstruction accuracy surpass when $\beta \in \mathbb{R}^{10}$ compared to $\beta \in \mathbb{R}^{300}$. Being consistent with the qualitative results (Fig. 13), it suggests that expressive $\beta \in \mathbb{R}^{300}$ fits meshes to excessively trimmed terminal parts such as wrists, leading to increased deviation from GT (i.e., overfitting). Moreover, increasing the number of voxels improves estimation accuracy. However, mesh reconstruction tends to fail (voxel size $= 512$ in Tab. 5) due to insufficient rays. Hence, in the experimental setup tested, $\beta \in \mathbb{R}^{10}$, voxel size $= 256$ is validated as the optimal hyperparameters.

7. Limitations

In this work, we propose the first-of-its-kind method to estimate static human pose and mesh from only event stream, using a moving event camera. The proposed method combines the classical idea of carving and fitting statistical models of the human body with the ray attenuation to preserve fine structures of the subject, and utilizes the high temporal information of event data. The experimental results show that (i) the proposed event-based method achieves better accuracy than frame-based baselines both in pose and mesh estimation, (ii) the ray attenuation is effective for mesh recovery with high-frequency details, (iii) the high temporal resolution contributes to the precise carving, and (iv) the proposed method is robust against motion-blur scenes. We hope this work will serve as a baseline to be extended to various interesting directions in the future as followings.

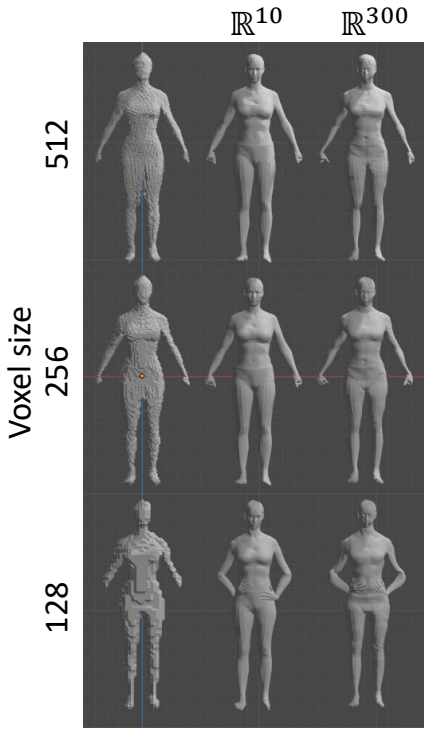

Figure 13. Qualitative result of sensitivity study.

The proposed method needs the camera trajectory, necessitating ego-motion estimation for the real-world applications. Since the scene (body) is static, one interesting direction could be the combination with SLAM (simultaneous localization and mapping) methods. Also, the contour classification utilizes supervised learning, while the other steps (carving and SMPL fitting) are optimization method. Thus, the contour classification performance may degrade for textures (e.g., clothes) or different poses that are not included in the training data. Creating a large-scale human body dataset with diverse textures and poses could be another important direction. Finally, the proposed method still may miss some fine structures as discussed in Sec. 6, which can be attributed to the limitation of the carving. Recently there have been many work for spatial understanding, such as Neural Radiance Field (NeRF) [30] and Gaussian Splatting [16], including dynamic scenes. Combining these ideas into the human scanning system could be a unique contribution in the future.

References

[1] Easymocap - make human motion capture easier. Github, 2021. 2, 5, 6, 7

[2] Anastasios N Angelopoulos, Julien NP Martel, Amit PS Kohli, Jorg Conradt, and Gordon Wetzstein. Event based, near eye gaze tracking beyond 10,000 hz. *IEEE Trans. Vis. Comput. Graphics*, 2020. 3

[3] Ryad Benosman, Sio-Hoi Ieng, Charles Clercq, Chiara Bartolozzi, and Mandyam Srinivasan. Asynchronous frameless event-based optical flow. *Neural Netw.*, 27:32–37, 2012. 3

[4] 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 *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 2019. 2

[5] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7035–7043, 2017. 2

[6] 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

[7] 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

[8] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch. A 1280x720 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 Intl. Solid-State Circuits Conf. (ISSCC)*, pages 112–114, 2020. 3

[9] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3867–3876, 2018. 3

[10] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2022. 1, 2

[11] Arren Glover and Chiara Bartolozzi. Event-driven ball detection and gaze fixation in clutter. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 2203–2208, 2016. 3

[12] Shuang Guo and Guillermo Gallego. Cmax-slam: Event-based rotational-motion bundle adjustment and slam system using contrast maximization. *IEEE Trans. Robot.*, pages 1–20, 2024. 3

[13] Jianping Jiang, Xinyu Zhou, Bingxuan Wang, Xiaoming Deng, Chao Xu, and Boxin Shi. Complementing event streams and rgb frames for hand mesh reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2024. 3

[14] 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. 2

[15] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, Liu C. Karen, and Michael J. Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. In *ACM ToG, Proc. SIGGRAPH Asia*, 2023. 2, 4, 7

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 8

[17] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J. Davison. Simultaneous mosaicing and tracking with an event camera. In *British Mach. Vis. Conf. (BMVC)*, 2014. 3

[18] Hanme Kim, Stefan Leutenegger, and Andrew J. Davison. Real-time 3D reconstruction and 6-DoF tracking with an event camera. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 349–364, 2016. 3

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5253–5263, 2020. 2, 5

[21] Xavier Lagorce, Garrick Orchard, Francesco Gallupi, Bertram E. Shi, and Ryad Benosman. HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(7):1346–1359, 2017. 1

[22] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2):150–162, 1994. 3

[23] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019. 2

[24] Sijin Li, Weichen Zhang, and Antoni B Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2848–2856, 2015. 2

[25] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128$\times$128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, 2008. 3

[26] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 695–710, 2020. 3

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 4

[28] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353, 1998. 4

[29] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020. 2

[30] 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. 8

[31] Anton Mitrokhin, Cornelia Fermuller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 1–9, 2018. 3

[32] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018. 2

[33] Elias Mueggler, Chiara Bartolozzi, and Davide Scaramuzza. Fast event-based corner detection. In *British Mach. Vis. Conf. (BMVC)*, 2017. 3

[34] Manasi Muglikar, Guillermo Gallego, and Davide Scaramuzza. ESL: Event-based structure light. In *Int. Conf. 3D Vision (3DV)*, pages 1165–1174, 2021. 3

[35] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 156–169. Springer, 2016. 2

[36] 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

[37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[38] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. In *Conf. on Robotics Learning (CoRL)*, pages 969–982. PMLR, 2018. 5

[39] 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

[40] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. A fast geometric regularizer to mitigate event collapse in the contrast maximization framework. *Adv. Intell. Syst.*, page 2200251, 2022. 3

[41] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 628–645, 2022. 3

[42] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Fast event-based optical flow estimation by triplet matching. *IEEE Signal Process. Lett.*, pages 1–5, 2023. 3

[43] Shintaro Shiba, Friedhelm Hamann, Yoshimitsu Aoki, and Guillermo Gallego. Event-based background-oriented schlieren. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 3

[44] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[45] Yunjae Suh, Seungnam Choi, Masamichi Ito, Jeongseok Kim, Youngho Lee, Jongseok Seo, Heejae Jung, Dong-Hee Yeo, Seol Namgung, Jongwoo Bong, Jun seok Kim, Paul K. J. Park, Joonseok Kim, Hyunsurk Ryu, and Yongin Park. A 1280x960 Dynamic Vision Sensor with a 4.95-μm pixel pitch and motion artifact minimization. In *IEEE Int. Symp. Circuits Syst. (ISCAS)*, pages 1–5, 2020. 3

[46] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE international conference on computer vision*, pages 2602–2611, 2017. 2

[47] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018. 2

[48] Keze Wang, Liang Lin, Chenhan Jiang, Chen Qian, and Pengxu Wei. 3d human pose machines with self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1069–1082, 2019. 2

[49] Yuanhao Wang, Ramzi Idoughi, and Wolfgang Heidrich. Stereo event-based particle tracking velocimetry for 3D fluid flow reconstruction. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 36–53, 2020. 3

[50] Ziyun Wang, Kenneth Chaney, and Kostas Daniilidis. Evac3d: From event-based apparent contours to 3d models via continuous visual hulls. In *European conference on computer vision*, 2022. 1, 3

[51] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 1, 2, 3

[52] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296, 2018. 2

[53] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human

pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2, 5, 6, 7

[54] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 5, 6, 7

[55] 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