RELI11D: A Comprehensive Multimodal Human Motion Dataset and Method

Ming Yan[1,2,3*] Yan Zhang[1,3*] Shuqiang Cai[1,3] Shuqi Fan[1,3] Xincheng Lin[1,3] Yudi Dai[1,3]
Siqi Shen[1,3†] Chenglu Wen[1,3] Lan Xu[4] Yuexin Ma[4] Cheng Wang[1,3]

[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University
[2]National Institute for Data Science in Health and Medicine, Xiamen University
[3]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, School of Informatics, Xiamen University
[4]Shanghai Engineering Research Center of Intelligent Vision and Imaging, ShanghaiTech University

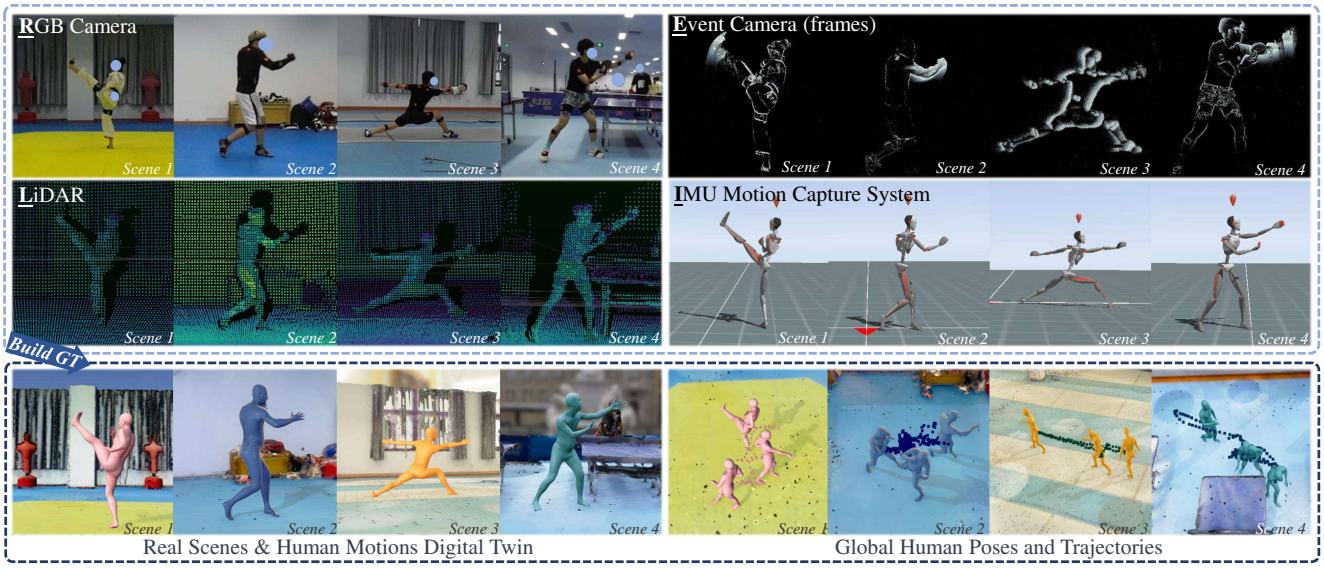

Figure 1. RELI11D is a high-quality dataset that provides four different modalities and records movement actions(first two rows). Our dataset's annotation pipeline can provide accurate global SMPL joints, poses as well as global human motion trajectories(last row).

Abstract

Comprehensive capturing of human motions requires both accurate captures of complex poses and precise localization of the human within scenes. Most of the HPE datasets and methods primarily rely on RGB, LiDAR, or IMU data. However, solely using these modalities or a combination of them may not be adequate for HPE, particularly for complex and fast movements. For holistic human motion understanding, we present ***RELI11D****, a high-quality multimodal human motion dataset involves* ***L****iDAR,* ***I****MU system,* ***R****GB camera, and* ***E****vent camera. It records the motions of 10 actors performing 5 sports in 7 scenes, including 3.32 hours of synchronized LiDAR point clouds, IMU measurement data, RGB videos and Event steams. Through extensive experiments, we demonstrate that the RELI11D presents considerable challenges and opportunities as it contains many rapid and complex motions that require precise location. To address the challenge of integrating different modalities, we propose* ***LEIR****, a multimodal baseline that effectively utilizes LiDAR Point Cloud, Event stream, and RGB through our cross-attention fusion strategy. We show that LEIR exhibits promising results for rapid motions and daily motions and that utilizing the characteristics of multiple modalities can indeed improve HPE performance. Both the dataset and source code release publicly in http://www.lidarhumanmotion.net/reli11d/, fostering collaboration and enabling further exploration in this field.*

1. Introduction

Human Pose Estimation (HPE) [1, 8, 10, 11, 42, 52, 54, 71, 97] is a challenging and long-standing research problem with significant potential for various applications, including AR/VR, autonomous driving, and sport analysis. Capturing complex and rapid human motions [16] is particularly chal-

* Equal contribution.
† Corresponding author.

lenging, which requires the accurate estimation of poses and precise localization of individuals within various scenes.

Researchers adopt RGB cameras [9, 25, 36, 63, 75, 76, 83] for HPE as they can capture appearance information, but they are light-sensitive and a monocular camera cannot provide depth information. Besides RGB imagery [28, 29, 31, 44, 50, 85], there are multiple types of sensors that excel in capturing various aspects of human motions [76]. RGBD sensors compensate for the absence of depth information, but their sensing range is limited. LiDAR [48] is light-insensitive and can provide 3D geometry. However, it suffers from sparsity and low frame rate issues. Inertial Measurement Units (IMUs) [33, 92] is occlusion-free. Nevertheless, they should be body-worn and are subject to the drifting issue. Event cameras [49] can capture motions with high temporal resolution and dynamic range by measuring intensity change asynchronously. However, they do not provide appearance information. In conclusion, these sensors have distinct characteristics. Therefore, to obtain the holistic understanding of human motions, using multiple types of sensors is important.

HPE methods are partially driven by the development of human motion datasets [53]. Most of them use several types of sensors. As far as we know, there does not exist a human motion dataset that contains the RGB, LiDAR, Event, and IMU modalities. Such a multimodal dataset is beneficial for the community to better understand human motions. To this end, we introduce a multimodal human motion dataset, **RELI11D**, which involves four types of sensors: RGB cameras, LiDAR, Event cameras, and IMU measurements. It comprises data from 10 actors (2 females, 8 males), encompassing motions of 5 different sports (table tennis, taekwondo, boxing, fencing, and badminton) in 7 scenes. RELI11D includes a diverse range of synchronized data, consisting of 199.26 minutes of RGB videos, event streams, IMU motion capture data, and point cloud frames.

The rich modalities and annotations provided in our dataset enable benchmarking on a series of 3D HPE tasks. We quantitatively and qualitatively evaluate multiple state-of-the-art methods for these tasks. Most of these methods cannot deal with rapid, coherent, and complex movements that require precise location. The experimental results show that our dataset brings new challenges to current computer vision algorithms.

To address these challenges, we propose **LEIR**, a baseline that estimates global human poses using **L**iDAR point clouds, **E**vent streams, and **R**GB images. It effectively utilizes the geometry information from LiDAR, the motion dynamics encoded in events, and the appearance features in RGB images through our multimodal cross-attention bases method. The advantages of LEIR have been thoroughly validated through experiments. We show that leveraging multiple modalities is necessary for a comprehensive understanding of human motions. In summary, our contributions are listed below:

- We present RELI11D, the *first* HPE dataset consisting of the RGB, IMU, LiDAR, and Event modalities.
- We provide a benchmark that enables the comparison of multiple methods using different modalities.
- We propose LEIR, a multi-modality baseline integrating the LiDAR point clouds, event streams, and RGB images for global human poses and trajectories estimation.

2. Related Work

2.1. Single modality Datasets and Methods

Many RGB motion datasets [8, 17, 30, 70] are collected using marker-based (e.g., Human3.6M [34], HumanEva[72]) or marker-less methods (e.g., MPI-INF-3DHP[55]). RGB-Based methods have flourished in recent years with a variety of approaches [7, 16, 18, 23, 43, 58, 60, 63, 67, 68, 73, 81–83, 89, 99], but most still focus solely on single RGB modality construction, with only a few considering the estimation of global trajectories [45, 47, 74, 94].

LiDAR can directly acquire three-dimensional spatial information. P4T [24] and STCrowd [20] use LiDAR point clouds to segment the human body. LiDARCap [48] performs 3D HPE through LiDAR point clouds.

An event camera generates a stream of events. Each pixel of the event camera responds asynchronously and independently to illumination change and generates an event if the change exceeds a threshold, which makes event cameras excel at capturing the local motions of objects [3, 51]. DHP19 [10] and [69] perform 2D HPE by treating event streams as image frames. EventCap [84] estimates 3D poses through a monocular event camera. EventHPE [101] performs HPE through a flow-based approach. EventPointPose [14] regresses poses through event point clouds.

IMU-based methods [33, 61, 80, 93] are environment-independent and occlusion-free. [57] fuses visual and inertial information for HPE; EgoLocate [91] estimates human poses based on sparse IMUs; [64] estimates poses based on physical contact. However, their necessity to be worn poses practical challenges and difficulties.

2.2. Multi-modality Datasets and Methods

TotalCapture[77] collects human poses in a studio through multi-view cameras and IMUs. 3DPW [79] collects subjects walking in a city through IMU and a hand-held camera. PedX[40] records pedestrian poses through stereo images and LiDAR point clouds. ImmFusion [12], FusionPose [19], and [96] use RGB and LiDAR body point clouds to reconstruct human poses. LIP [66] reconstructs poses using sparse IMUs with LiDAR. Besides using a monocular camera and IMUs, LiDARHuman26M [48],

Dataset	Sensor Modalities				Global Trajectory	Frames	3D Scene	Motion	Real/ Synthetic	Number of Sequences	Number of Subjects
	RGB	MoCap	LiDAR	Event							
LiDARHuman26M[48]	✔	IMU	✔	-	-	184k	-	Daliy	Real	20	13
HSC4D [22]	-	IMU	✔	-	✔	10k	✔	Daliy	Real	8	1
SLOPER4D [21]	✔	IMU	✔	-	✔	100k	✔	Daliy	Real	15	12
CIMI4D [87]	✔	IMU	✔	-	✔	180k	✔	Climbing	Real	42	12
LIPD [66]	✔	IMU	✔	-	✔	-	-	Obvious	Real	10	6
LiCamPose[19]	✔	IMU	✔	-	-	9k	-	Daliy	Real	-	-
EMDB[38]	✔	EM	-	-	✔	105k	-	Daliy	Real	81	10
SMART[16]	✔	-	-	-	-	110k	-	Sports	Real	640	-
X-Avatar[71]	✔	-	-	-	-	35k	✔	Daliy	Real	233	20
BEHAVE [5]	✔	-	-	-	✔	15k	-	Interactions	Real	-	8
RICH [32]	✔	-	-	-	✔	577k	✔	Interactions	Real	142	22
AGORA [59]	✔	-	-	-	✔	106.7K	✔	Daliy	Synthetic	-	-
3D-FRONT HUMAN [90]	-	-	-	-	✔	-	✔	Daliy	Synthetic	-	-
BEDLAM [6]	✔	-	-	-	-	1M	✔	Daliy	Synthetic	-	-
DHP19 [10]	-	-	-	✔	-	87k	-	Daliy	Real	-	17
MMHPSD [100]	✔	-	-	✔	-	240k	-	Daliy	Real	84	15
RELI11D(Ours)	✔	✔	✔	✔	✔	**239k**	✔**(7)**	**Sports**	**Real**	**48**	**10**

Table 1. Comparisons with related datasets. The ”-” symbol indicates that it is not included in the dataset.

HSC4D [22], SLOPER4D [21], CIMI4D [87] collect human poses through a static monocular LiDAR, a body-mounted LiDAR, a head-mounted LiDAR, and a LiDAR respectively. Researchers have also explored some other sensors for human motions, such as WIFI [35, 65, 98],mmWave [2, 12, 13], and electromagnetic sensor [38].

3. RELI11D: a multimodal motion dataset

RELI11D is a multimodal high-quality human movement dataset that contains five different categories of sports: table tennis, taekwondo (examination movements and free exercises), boxing (traditional boxing, free fighting, Muay Thai), fencing (saber, epee, foil), and badminton. It collects 4 modalities for a total of 48 sequences of 199.2 minutes (3.32 hours) of synchronized RGB camera video, Event camera streams, IMU Measurements Data, and LiDAR point clouds. In total, there are 239k frames of human body point clouds. In RELI11D, we invite 10 volunteers to collect sports in 7 scenes. All the participants agree that their recorded data may be used for scientific purposes. Fig. 2 describes the rich modalities and annotations the community can get from RELI11D. Tab. 1 provides statistics comparing with other publicly available human pose datasets. As far as we know, RELI11D is the *first* dataset that consists of RGB, LiDAR, IMU, and Event modalities. Moreover, it contains high-precision 3D scans of 7 scenes , global poses and trajectories of each actor, contributing to comprehensive human scene perception.

3.1. Hardware and Configuration

The motion collection device is composed of multiple sensors that can collect motions indoor and outdoor. As shown in Fig. 3, we use LiDAR (Ouster-OS1, 128-beam) to capture 3D dynamic point clouds at 20 frames-per-second (FPS), a monocular RGB camera (DJI Action 2, 4096×3072) to

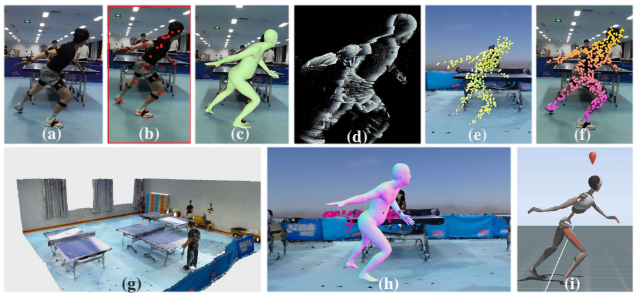

Figure 2. RELI11D provides rich data and annotations: (a) RGB Videos, (b) 2D Annotation, (c) 2D SMPL Poses, (d) Events, (e) 3D Point Clouds, (f) 2D Point Clouds, (g) High Precision Scene Meshes, (h) 3D SMPL Shape, Poses, and Trajectories, (i) IMUs Measurements.

Figure 3. Portable Human Motion Capturing system.

record RGB video at 60 FPS, and an event camera (CeleX-V [15], 1280x800) to record event streams. For each sports scene, we use the Trimble X7 3D laser scanning system to reconstruct a high-precision RGB 3D point cloud for it, totaling 80 million points for each scene.

Each volunteer wears an Xsens MVN inertial motion capture system. It contains 17 IMUs, which record poses at a speed of 60 FPS. To obtain their body shape (SMPL β), we scan their body using a handheld point cloud modeling equipment and obtain β through IPNet [4].

Human Pose Model and Label. A human motion is denoted by $M = (T, \theta, \beta)$, where T represents the $N \times 3$

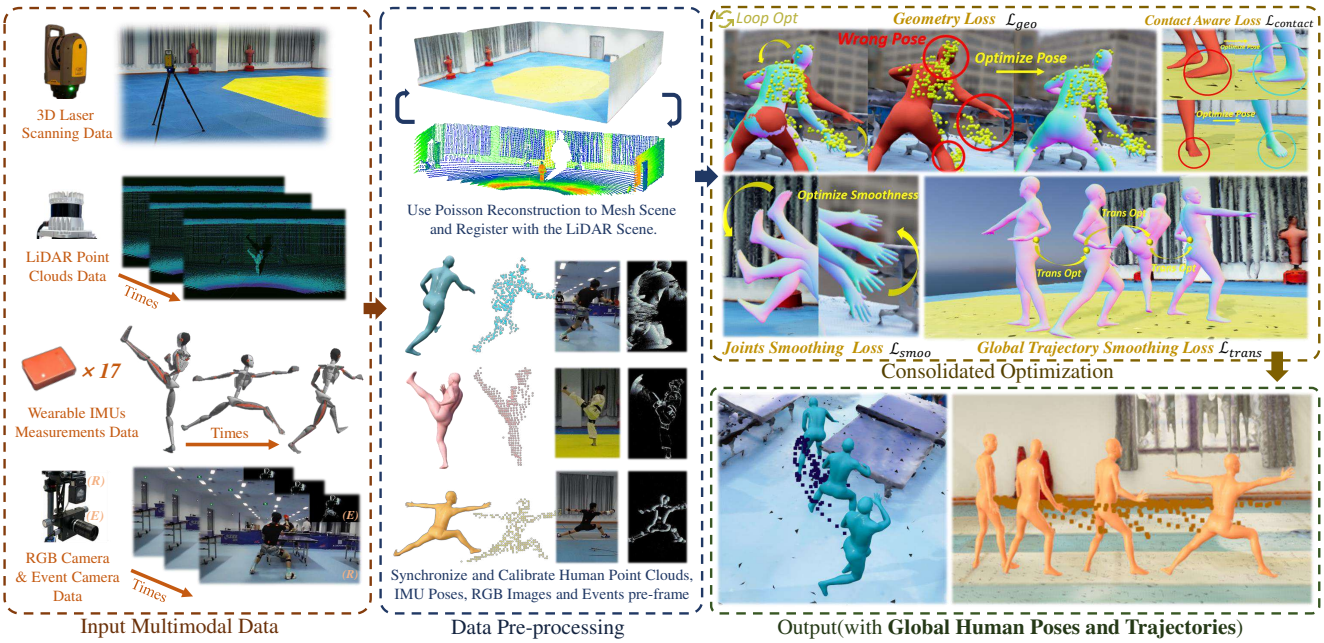

Figure 4. **Overview of main annotation pipeline.** The dotted boxes of different colors represent different data processing stages, and the arrows represent the data flow direction. **Dotted box:** The input of each scene sequence consists of RGB videos, point cloud sequences, IMU measurements, events flow(times axis), and 3D laser scanning data. The data pre-processing stage calibrates and synchronizes different modalities. The consolidated optimization includes the global pose and translation based on multiple constraint losses.

translation parameters, $N \times \theta$ is the 24×3 pose parameter, and β is the 10 dimensions shape parameter following SMPL [52], N is the frame count. As IMUs suffer severe drifting for long-period capturing, we seek to find the precise T and θ for RELI11D as annotation labels.

3.2. Data Annotation Pipeline

The data annotation pipeline consists of 3 stages: pre-processing, consolidated optimization, and manual annotation. Fig. 4 depicts the main annotation pipeline.

3.2.1 Multimodal Data Pre-processing Stage

Scene reconstruction. For each RGB high-precision static 3D point cloud scene, we convert it into a mesh scene composed of triangular patches through implicit surface Poisson reconstruction [39]. Data in this format can more accurately calculate the interaction between the human body and the environment is convenient.

Time synchronization. Synchronization between the IMU, LiDAR, RGB video, and event streams is achieved by detecting spikes in jump events. In each motion sequence, the subject jumps in place, and we design a peak detection algorithm to automatically find the height peaks in the IMU and LiDAR trajectories. RGB videos and IMU data are downsampled to 20/fps, consistent with LiDAR's frame rate. The event stream is divided into multiple event frames $\mathbb{E} = \{\mathbb{E}_{t_i}\}_{i=1}^{N}$. $\mathbb{E}_{t_i}$ is the set of events whose time stamp t satisfies $t_{i-1} < t \leq t_{i-1}$.

Calibration. Initially, we register the LiDAR sparse point scene and the high-precision scene of each sequence to the same coordinate system. Next, for each frame, we isolate the human body point clouds, and derive the human poses based on it. The movement sequence of a person in world coordinates $\{W\}$ is represented by $\boldsymbol{M^W} = (T^W, \theta^W, \beta)$. T^I and θ^I in $M^I = (T^I, \theta^I, \beta)$ are provided by IMUs. $\theta^W = R_{WI}\theta^I$ is used as the initial poses, where R_{WI} is a rough calibration matrix from the IMU coordinate system to the world coordinate system. As the translation measured by IMUs is not accurate [87], we use the position of the center of the human hip in the point clouds as T^W. Lastly, we execute frame-level temporal synchronization and spatial calibration for the scenes and all the modalities.

3.2.2 Consolidated Optimization

We utilize contact aware loss $\mathcal{L}_{contact}$, smoothness loss $\mathcal{L}_{smoo}$ and geometry loss $\mathcal{L}_{geo}$ to perform consolidated optimization of global poses and trajectories to obtain accurate and scene-natural human motion. Please refer to the supplementary for a detailed formulation of these losses. We minimize the overall loss which is defined as follows.

$$\mathcal{L} = \lambda_c \mathcal{L}_{contact} + \lambda_s \mathcal{L}_{smoo} + \lambda_g \mathcal{L}_{geo} \tag{1}$$

where λ_c, λ_s, λ_g are loss coefficients.

Contact Aware Loss. The $\mathcal{L}_{contact}$ term combines scene constraints $\mathcal{L}_{sceneC}$ and self-penetration constraints $\mathcal{L}_{selfC}$ to improve the quality of local human poses. First, we use

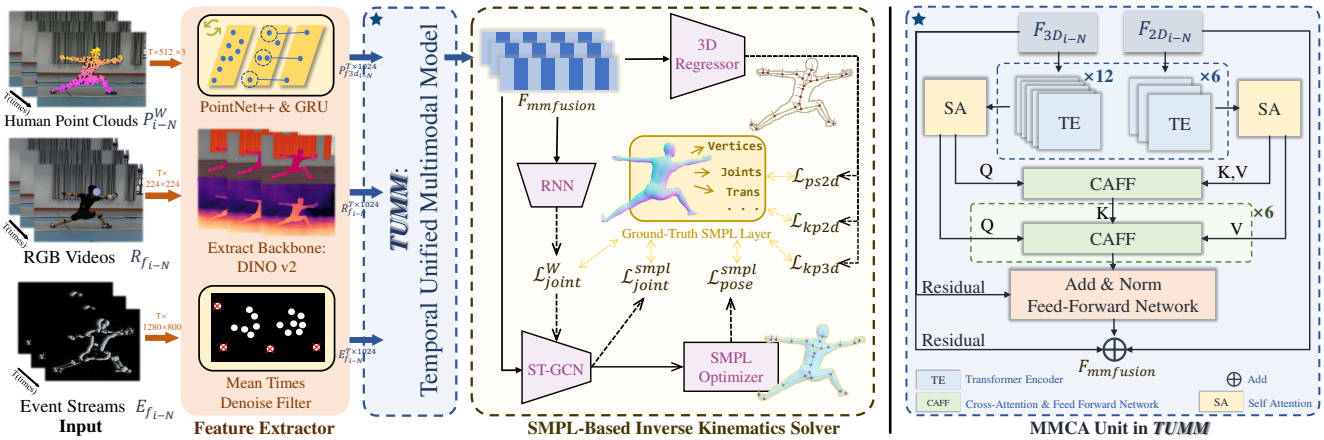

Figure 5. **Overview of LEIR method (Left) and Multimodal Cross-Attention Unit (Right).** Orange arrows represent different modalities of data input. Dark blue arrows represent the inputs and outputs data flows of the TUMM model. Dotted arrows represent the predicted data and calculation loss with ground truth.

λ_{sceneC} to penalize the vertices in the human SMPL mesh that penetrate the scene mesh. To avoid the self-penetration problem, we use the self-penetration constraint $\mathcal{L}_{selfC}$.

Smoothness Loss. We introduce $\mathcal{L}_{smoo}$ to ensure the motion smoothness. It includes (1) the global trajectory smoothing term $\mathcal{L}_{trans}$, which smooths the human body motion by minimizing pelvis acceleration. (2) The body posture smoothing term $\mathcal{L}_{poses}$ maintains the stability of the entire human body motion by minimizing the angular velocity of each pelvis-related joint. (3) The human joints smoothing term $\mathcal{L}_{joints}$ smooths SMPL joint acceleration.

Geometry Loss. Point clouds contain the geometry of human motion, we use $\mathcal{L}_{geo}$ to make visible SMPL vertices approximate the geometric relationship. Following [22], for each SMPL mesh, we use [37] to remove invisible mesh vertices from the LiDAR perspective. $\mathcal{L}_{geo}$ is the 3D Chamfer distance between body point and visible SMPL vertices.

3.2.3 Manual Annotation and Verification Stage

We manually correct the pose and translation parameters of a subject's motions for some artifacts. Further, an external person has examined our dataset. We have adjusted the imprecise annotations pointed out by this person.

4. LEIR: A multimodal HPE baseline

We propose LEIR, a multi-modality baseline for human motion estimation. Given the synchronized LiDAR point clouds, RGB images, and event streams that are captured by multiple sensors, the task of the baseline is to predict the 3D pose of the human in the world coordinate system.

Many existing two-modal-based methods [26, 41, 46] adopt a two-tower architecture, where each tower processes only one modality. LEIR aims to fully model three (rather than two) modalities, which is non-trivial because each modality contains different information.

As is depicted in Fig. 5, LEIR consists of three major modules: feature extractors, the temporal unified multimodal model (TUMM), and SMPL-based inverse kinematics solver. For each modality, feature extractors are used to extract its features, which are fused through the TUMM modules to fully utilize the 3D geometric information of point clouds, the appearance information of RGB images, and the temporal dynamics of event streams. In the end, the fused features are fed into SMPL solver to obtain the estimated poses. Please see the appendix for more details.

4.1. Feature Extraction

RGB Feature Extraction. For each RGB frame, we specifically target the human body by applying a bounding box and extracting its corresponding feature $R_{f_{i-N}}$ using an RGB encoder (DINOv2 [56]).

LiDAR Feature Extraction. For the human body point clouds P^W_{i-N}, we extract its features $P_{f3d_{i-N}}$ by feeding the point clouds into a PointNet++[62] and a GRU network.

Event Feature Extraction. For each event frame $\mathbb{E}_{t_i}$, we employ an average time sampling filter with adjacent point denoising [86] to effectively process the noise in the frame. This filtering technique enhances the visibility of changes in human body movement between frames. Subsequently, we aggregate all the events in a frame based on their pixel location and polarity, thus generating an image-like event frame. The features $E_{f_{i-N}}$ of this frame are then extracted using an RGB encoder (DINOv2).

4.2. Temporal unified multimodal model (TUMM)

Previous methods, such as [96], fuse the LiDAR and RGB modalities through projection rely on accurate calibration, which may not always available. Moreover, it is unclear how to effectively fuse LiDAR and Event modalities.

To automatically learn correspondence among three

modalities and eliminate calibration sensitivity, TUMM uses the cross-attention strategy. The TUMM module consists of two steps. In the first step, the LiDAR point clouds and the RGB images are fused using the multimodal cross-attention unit (MMCA), LiDAR point clouds and the event frames are fused using MMCA as well. This step aims to effectively integrate the geometry information with appearance information, and integrate the geometry information with motion dynamics. In the second step, the features obtained from the first step are further fused using MMCA, which allows a comprehensive integration of the features from different modalities.

The design of the MMCA unit is depicted in the right part of Fig. 5. The LiDAR features $F_{3D_{i-N}}$ and the RGB/event features $F_{2D_{i-N}}$ are processed through a series of transformer encoders [78] and self-attention mechanisms. MMCA employs a 2-layer cross-attention structure, using the fused keys as intermediaries to match and align two sources of information. In the first layer, the features from the LiDAR act as queries, while the features from the RGB/events serve as keys and values. In the second layer, the output from the last layer serves as the keys; the LiDAR feature and RGB feature serve as query and value, respectively. The output features are obtained by element-wise addition of the input features and the results of the cross-attention structures. For the second step of TUMM, the 2D (right) branch of MMCA is replaced by a 3D branch. The output results of the previous step are fed into MMCA for finding correspondence among three modalities.

4.3. SMPL-Based inverse motion solver

The fused features $F_{mmfusion}$ are utilized in three branches within the network. In the first branch, these features are fed into a 3D regressor to estimate 3D joints and camera intrinsic parameters. This branch involves three different loss functions. $\mathcal{L}_{ps2d}$ serves as a projection loss, ensuring that the 2D appearance of the SMPL model aligns with the human body in pixel coordinates. $\mathcal{L}_{kp2d}$ and $\mathcal{L}_{kp3d}$ are used to respectively constrain the 2D and 3D joints of the human body. In the second branch, the features are fed into an RNN network that predicts 3D human joints in the world coordinate system. $\mathcal{L}^{W}_{joint}$ is employed to encourage alignment with the labels. The third branch employs an ST-GCN [88], where the fused features are used to predict 3D human joints. In addition, we apply $\mathcal{L}^{smpl}_{joint}$ to ensure accurate joint orientation. Finally, the outputs of the third branch are passed through an SMPL optimizer to obtain the human pose in axis-angle form, and $\mathcal{L}^{smpl}_{pose}$ is employed to enforce alignment with the ground truth poses. The overall loss function used in LEIR combines all these individual losses, which is defined as follows.

$$\mathcal{L}_{ps2d} + \mathcal{L}_{kp2d} + \mathcal{L}_{kp3d} + \mathcal{L}^{W}_{joint} + \mathcal{L}^{smpl}_{pose} + \mathcal{L}^{smpl}_{joint} \quad (2)$$

Sport sequences	ACCEL↓	MPJPE↓	PA-MPJPE↓	PVE↓	PCK0.3↑
Pingpong	0.72/3.70	22.31/58.42	21.90/58.62	31.89/85.55	0.98
Badminton	0.84/1.97	27.91/67.57	26.19/64.11	28.45/63.20	0.95
Taekwondo	1.25/2.89	29.28/68.23	25.02/63.13	36.83/66.23	0.95
Boxing	1.79/5.22	31.42/67.56	26.43/50.83	38.45/75.45	0.96
Fencing	0.40/5.00	25.66/62.73	20.54/52.09	27.73/53.68	0.98

Table 2. Quality of the optimization process. Each cell reports the mean and maximal error metrics which are separated by "/".

Constraint term			PingPong1			Boxing1		
$\mathcal{L}_{contact}$	$\mathcal{L}_{smoo}$	$\mathcal{L}_{geo}$	ACCEL↓	MPJPE↓	PA-MPJPE↓	ACCEL↓	MPJPE↓	PA-MPJPE↓
✗	✗	✗	5.29	29.14	20.7	6.79	31.20	25.65
✓	✓	✗	1.96	13.75	10.11	1.52	11.89	14.96
✓	✗	✓	4.03	15.19	12.91	5.50	18.82	12.13
✗	✓	✓	1.29	18.14	13.80	1.10	12.23	9.66
✓	✓	✓	**0.92**	**12.25**	**9.31**	**0.85**	**10.26**	**8.35**

Table 3. Evaluation of Consolidated Optimization for different constraints. Unit: mm

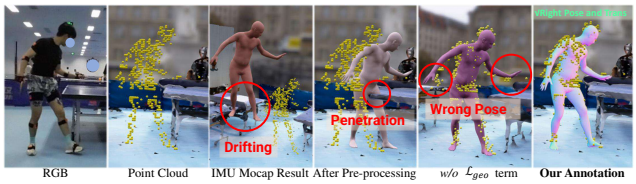

Figure 6. **Qualitative evaluation.** From left to right: RGB image, LiDAR point clouds, initial IMU motion capture result, after pre-processing stage, after optimization without the $\mathcal{L}_{geo}$ loss, after optimization stage.

To evaluate the performance of different input modalities using the same network, we employ the same training strategy as [95], which freezes unnecessary model parameters based on different input modalities, thus enabling multimodal or single-modal training and inference.

5. Experimental Results

Evaluation metrics. We report Mean Per Joint Position Error (MPJPE), Procrustes Aligned Mean Per Joint Position Error (PA-MPJPE), Percentage of Correct Keypoints (PCK0.3), Per Vertex Error (PVE), Acceleration Error(mm/s^2) (ACCEL). Regarding the evaluation of trajectory errors, we use Global MPJPE (GMPJPE) to calculate the mean per joint position error of the SMPL model in global coordinates, global human body root node Translation Error(T-Error). The PCK0.3 is calculated as a percentage, while other indicators are in mm.

5.1. Dataset Evaluations

We study the quality of the RELI11D dataset through qualitative and quantitative evaluations.

Qualitative evaluation. Fig. 6 depicts a frame of the RELI11D dataset. The global translation of IMUs may be in-precise as it is floating in the air (column 3 of Fig. 6). Through the pre-processing stage, the quality of the dataset is improved, the position of the SMPL model is on the ground (column 4 of Fig. 6). However, it may cause penetration. In the consolidated optimization stage, RELI11D

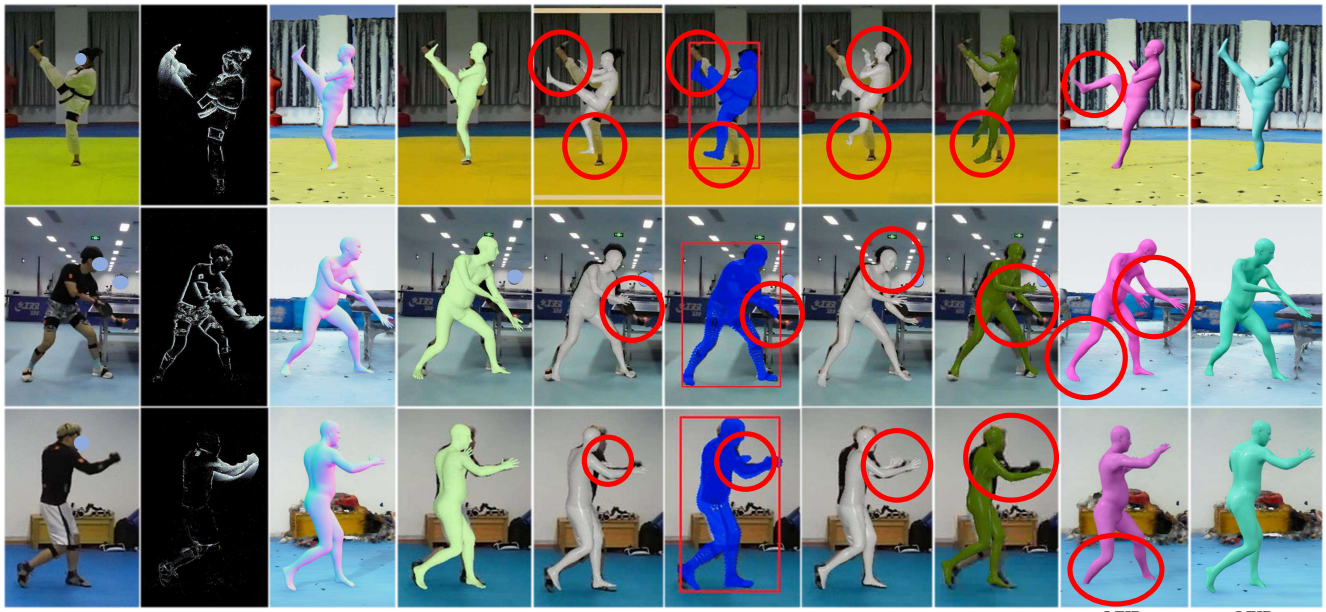

Figure 7. **Benchmark experiment and RELI11D demonstration.** Image-based methods (left col 5 to 9) can produce erroneous results due to occlusions and rapid movements (red circles). Our multimodal method (first right col) performs best qualitatively by comparison.

uses the contact aware loss to avoid penetration so that the quality is further improved. (column 5 of Fig. 6). Nevertheless, limb poses may be wrong. Through using the $\mathcal{L}_{geo}$ loss, we obtain the accurate poses and translations (the last column of Fig. 6).

Quantitative evaluation. To quantitatively evaluate the annotation quality of RELI11D, we select annotated motion sequences, and evaluate the performance of the consolidated optimization stage (in Sec. 3.2.2) by comparing the generated annotations against optimized annotations. Tab. 2 depicts the error metrics for the annotations generated without/with the optimization stage in mean/max. The error metrics are small, which demonstrates the effectiveness of the annotation pipeline and the high quality of RELI11D.

In order to understand the impact of different constraints used in the consolidated optimization stage, we conduct a ablation study on 3 different losses: $\mathcal{L}_{contact}$, $\mathcal{L}_{smoo}$ and $\mathcal{L}_{geo}$. Tab. 3 shows the error metrics of using different combinations of losses for two scenes. Without using any losses (row 1), the error metrics are the largest. If we remove any loss from the optimization stage (row 2 to 4), the error metrics increase, indicating all the losses are useful to improve the quality of our dataset.

5.2. Benchmark

The rich modalities provided in RELI11D allow us to conduct a systematic benchmark on HPE methods. We consider 5 HPE tasks: LiDAR-based, RGB-based, RGB+LiDAR-based, Event-based, and HPE with global trajectory.

Input Modality	Method	ACCEL↓	MPJPE↓	PA-MPJPE↓	PVE↓	PCK0.3↑
LiDAR	P4Tranformer [24]	66.33	172.04	150.65	206.75	0.51
	PCT [27]	59.19	144.40	116.99	174.33	0.67
	LiDARCap [48]	54.42	144.51	106.20	176.98	0.67
RGB	HybrIK [47]	58.39	249.34	163.91	255.98	0.53
	NIKI [45]	55.62	196.68	142.48	198.10	0.61
	SMPLer-X [9]	50.15	171.97	128.02	185.83	0.66
Event	EventHPE [101]	-	193.7	115.72	224.59	0.52
	EventPointPose [14]	-	16.24(2D)	10.91(2D)	-	0.69(2D)
RGB+ LiDAR	ImmFusion [12]	49.19	175.00	159.62	187.31	0.67
	FusionPose [19]	44.89	136.15	110.19	166.94	0.75

Table 4. Comparison of SOTA HPE methods on different modals in RELI11D. Unit: mm

Input Modality	Method	ACCEL↓	MPJPE↓	PA-MPJPE↓	GMPJPE↓	T-Error↓	PCK0.3↑
RGB	GLAMR [94]	47.83	202.66	179.59	495.40	590.46	0.65
	TRACE [74]	50.09	197.41	165.18	488.91	581.67	0.68

Table 5. Comparison of SOTA 3D HPE methods **with global translation** in RELI11D. Unit: mm

5.2.1 Human Pose Estimation (HPE)

We evaluate the performance of multiple state-of-the-art 3D HPE methods in RELI11D. Based on their input modality, we categorize these methods into four categories. For LiDAR-based input, LiDARCap [48], P4Transformer [24], and PCT [27] are used. For RGB-based input, NIKI [45] and SMPLer-X [9] are used. For event-based input, EventHPE [101] and EventPointPose [14] are compared. For the RGB and LiDAR inputs, FusionPose [19] and ImmFusion [12] are tested.

Tab. 4 shows the HPE experimental results. We find that all the methods perform poorly on the RELI11D dataset, even for methods trained on multiple datasets. Fig. 7 shows that some methods do not work well for certain poses in the RELI11D dataset. The motions in taekwondo sports are rapid, their range is rarely seen in daily activities. As is depicted in the first row of Fig. 7, all the RGB-based meth-

Input Modality	ACCEL↓	MPJPE↓	PA-MPJPE↓	G-MPJPE↓	T-Error↓	PCK0.3↑
LiDAR	31.26	59.93	48.23	125.77	195.72	0.89
RGB	28.43	62.71	54.11	557.81	710.84	0.88
Event	34.45	107.78	83.64	605.45	743.71	0.59
LiDAR+ RGB	27.07	55.36	45.72	122.32	168.61	0.90
LiDAR+ Event	25.41	57.79	46.70	123.75	178.97	0.89
LiDAR+ RGB+Event	**23.90**	**49.19**	**40.87**	**115.36**	**146.13**	**0.92**

Table 6. The performance of LEIR input with different modalities based on RELI11D. Unit: mm

Input Modality	Method	ACCEL↓	MPJPE↓	PA-MPJPE↓	PVE↓	PCK0.3↑
LiDAR	LiDARCap [48]	45.89	80.08	67.50	102.24	0.85
	LEIR(Ours)	**45.60**	**79.00**	**67.45**	**100.87**	**0.85**
LiDAR+ RGB	ImmFusion [12]	46.45	96.93	81.16	107.29	0.75
	FusionPose [19]	**44.51**	78.18	66.70	99.66	0.85
	LEIR(Ours)	44.52	**75.09**	**62.94**	**95.96**	**0.87**

Table 7. Performance evaluation of LEIR in the LiDARHuman26M dataset [48]. Unit: mm

ods do not model the limbs well. For Table Tennies, besides modeling the rapid movement of hands and feet, their global position should be captured to avoid penetration. As shown in the second row of Fig. 7, all the RGB methods have prediction artifacts. This indicates that RELI11D is a *challenging dataset* for existing methods.

As it is shown in Tab. 4, using two modalities (RGB+LiDAR) leads to better performance than using these two modalities separately. The rich modalities provided by RELI11D enable the comparison of different combinations of modality-based methods.

5.2.2 Global HPE

Capturing certain actions, such as playing table tennis, requires the precise locations of humans. We evaluate two RGB-based global 3D pose estimation methods (GLAMR [94] and TRACE [74]) on the RELI11D dataset. Their results are depicted in Tab. 5, which shows that their performance is unsatisfactory. This suggests that despite the monocular camera methods can match local motions in 2D RGB well, they do not excel at 3D global motions.

5.3. Baseline Evaluation

We evaluate the proposed baseline, LEIR, based on the RELI11D and the LiDARHuman26M [48] datasets, with different combinations of modalities.

LEIR on RELI11D. Tab. 6 presents the results of LEIR with different combinations of modalities as input. For single modality (rows 2 to 4), LiDAR-based input achieves the best performance due to its ability to provide detailed geometric information. RGB-based input achieves the second-best performance, benefiting from its appearance information. Event-based input yields the worst results. As the number of modalities increases, the performance of LEIR improves. This emphasizes that correctly fusing the information of each modality feature makes the method robust. When combining LiDAR with RGB/Event inputs (rows 5 and 6), LEIR outperforms the use of LiDAR alone. For HPE, the best performance is achieved when utilizing all the modalities (LiDAR+RGB+Event), and it is significantly better than all the studied state-of-the-art methods.

Regarding global HPE, compared with the two RGB-based methods (shown in Tab. 5), the results shown in Tab. 6 demonstrate that using the LiDAR modality is necessary for global pose estimation. Combining all three modalities can achieve the best global HPE results.

These results highlight the effective utilization of information from different modalities in LEIR, and they are visualized in Fig. 7, intuitively showing that LEIR (RGB+PC+Events) works well on the RELI11D dataset.

We conduct a study on another dataset LiDARHuman26M [48], which contains both RGB and LiDAR modalities. In this experiment, we train all the methods from scratch based on the methodology described in [48] and follow the same evaluation as [48]. As is shown in Tab. 7, when considering LiDAR input alone, LEIR demonstrates a slight improvement over LiDARCap. When combining LiDAR and RGB inputs, LEIR out-performs ImmFusion [12] and FusionPose [19]. This indicates that our proposed method, LEIR, performs well on other dataset.

Global trajectory prediction. The T-Error measures the translation error is depicted in Tab. 6. And the predicted trajectory is plotted in Supp Fig.3. It shows that incorporating the LiDAR point clouds with global trajectory information improves the global motion indicators (low T-Error, similarity between the curve and ground truth). This observation indicates a promising trend of multimodal methods that fuse global information.

6. Conclusion

Different motion sensors have distinct characteristics (e.g., geometry) that excel at capturing challenging motions (e.g., complex and fast motion). We introduce RELI11D, the first human motion dataset with the LiDAR, RGB, IMU, and Event modalities for a holistic understanding of human motions. It records the motions of 10 actors performing 5 sports in 7 different scenes. The rich annotations in RELI11D enable benchmarking a series of HPE tasks. We demonstrate that RELI11D is challenging due to its fast and complex motions. To address this challenge, we propose LEIR, a multimodal HPE baseline that utilizes the LiDAR points cloud, event streams, and RGB videos through cross-attention strategy. We show through extensive experiments that LEIR can obtain the most competitive results.

Acknowledgements. This work was partially supported by the National Natural Science Foundation of China (No.62171393), by the Fundamental Research Funds for the Central Universities (No.20720230033, No.20720220064), by PDL (2022-PDL-12).

References

[1] Thiemo Alldieck, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor. Optical flow-based 3d human motion estimation from monocular video. In *German Conference on Pattern Recognition*, pages 347–360. Springer, 2017. 1

[2] Sizhe An, Yin Li, and Umit Ogras. mRI: Multi-modal 3d human pose estimation dataset using mmwave, RGB-d, and inertial sensors. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[3] R. Wes Baldwin, Ruixu Liu, Mohammed Almatrafi, Vijayan K. Asari, and Keigo Hirakawa. Time-ordered recent event (TORE) volumes for event cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(2):2519–2532, 2023. 2

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 311–329. Springer, 2020. 3

[5] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 3

[6] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 3

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2

[8] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577. Springer, 2022. 1, 2

[9] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *NeurIPS Dataset and Benchmark Track*, June 2023. 2, 7

[10] 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*, 2019. 1, 2, 3

[11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[12] Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2752–2758. IEEE, 2023. 2, 3, 7, 8

[13] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3501–3510, 2022. 3

[14] Jiaan Chen, Hao Shi, Yaozu Ye, Kailun Yang, Lei Sun, and Kaiwei Wang. Efficient human pose estimation via 3d event point cloud. In *3DV*, 2022. 2, 7

[15] Shoushun Chen and Menghan Guo. Live demonstration: Celex-v: A 1m pixel multi-mode event-based sensor. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1682–1683, 2019. 3

[16] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129:2846–2864, 2021. 1, 2, 3

[17] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023. 2

[18] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *CVPR*, pages 4800–4810, 2023. 2

[19] Peishan Cong, Yiteng Xu, Yiming Ren, Juze Zhang, Lan Xu, Jingya Wang, Jingyi Yu, and Yuexin Ma. Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In *AAAI*, pages 461–469. AAAI Press, 2023. 2, 3, 7, 8

[20] Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19608–19617, 2022. 2

[21] Yudi Dai, Yitai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 682–692, 2023. 3

[22] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6792–6802, June 2022. 3, 5

[23] Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, and Wenping

Wang. TORE: token reduction for efficient human mesh recovery with transformer. In *ICCV*, 2023. 2

[24] Hehe Fan, Yi Yang, and Mohan S. Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14199–14208, 2021. 2, 7

[25] Michael Fürst, Shriya T. P. Gupta, René Schuster, Oliver Wasenmüller, and Didier Stricker. HPERL: 3d human pose estimation from RGB and lidar. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 7321–7327. IEEE, 2020. 2

[26] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 5

[27] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021. 7

[28] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 38(2):14:1–14:17, 2019. 2

[29] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[30] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2282–2292. IEEE, 2019. 2

[31] Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuexin Ma, and Lan Xu. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11400–11411, 2021. 2

[32] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 3

[33] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37(6):185:1–185:15, nov 2018. 2

[34] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:1325–1339, 2014. 2

[35] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. Towards 3d human pose construction using wifi. In *MobiCom '20: The 26th Annual International Conference on Mobile Computing and Networking, London, United Kingdom, September 21-25, 2020*, pages 23:1–23:14. ACM, 2020. 3

[36] 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 Regognition (CVPR)*, 2018. 2

[37] Sagi Katz, Ayellet Tal, and Ronen Basri. Direct visibility of point sets. In *ACM SIGGRAPH 2007 papers*, pages 24–es. 2007. 5

[38] 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*, pages 14632–14643, 2023. 3

[39] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 4

[40] Wonhui Kim, Manikandasriram Srinivasan Ramanagopal, Charlie Barto, Ming-Yuan Yu, Karl Rosaen, Nicholas Goumas, Ram Vasudevan, and Matthew Johnson-Roberson. Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 4:1940–1947, 2019. 2

[41] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 5

[42] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2020. 1

[43] 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 (ICCV)*, pages 11127–11137, October 2021. 2

[44] 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

[45] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12933–12942, 2023. 2, 7

[46] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare,

Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 5

[47] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 2, 7

[48] Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20502–20512, 2022. 2, 3, 7, 8

[49] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbrück. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid State Circuits*, 43(2):566–576, 2008. 2

[50] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pages 21159–21168. IEEE, 2023. 2

[51] Xiuhong Lin, Changejie Qiu, Zhipeng Cai, Siqi Shen, Yu Zhang, Weiquan Liu, Xuesheng Bian, Matthias MGüller, and Cheng Wang. E2PNet: event to point cloud registration with spatio-temporal representation learning. In *NeurIPS*, 2023. 2

[52] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015. 1, 4

[53] 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 (ICCV)*, October 2019. 2

[54] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 1

[55] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal V. Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 2

[56] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[57] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2

[58] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7363–7373, 2020. 2

[59] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3

[60] 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, June 2019. 2

[61] Gerard Pons-Moll, Andreas Baak, Juergen Gall, Laura Leal-Taixe, Meinard Mueller, Hans-Peter Seidel, and Bodo Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *2011 International Conference on Computer Vision*, pages 1243–1250. IEEE, 2011. 2

[62] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5

[63] 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, October 2021. 2

[64] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 71–87. Springer, 2020. 2

[65] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. Winect: 3d human pose tracking for free-form activity using commodity wifi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(4):176:1–176:29, 2021. 3

[66] Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu, and Yuexin Ma. Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2337–2347, 2023. 2, 3

[67] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2304–2314. IEEE, 2019. 2

[68] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 81–90. Computer Vision Foundation / IEEE, 2020. 2

[69] Gianluca Scarpellini, Pietro Morerio, and Alessio Del Bue. Lifting monocular events to 3d human poses. In *CVPRW*, pages 1358–1368, 2021. 2

[70] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *European Conference on Computer Vision*, pages 702–720. Springer, 2022. 2

[71] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16911–16921, 2023. 1, 3

[72] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 2009. 2

[73] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *ECCV*, 2020. 2

[74] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8866, 2023. 2, 7, 8

[75] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, Rui Zhao, and Wanli Ouyang. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *CVPR*, pages 21970–21982, 2023. 2

[76] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *CoRR*, abs/2203.01923, 2022. 2

[77] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John P. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *BMVC*, 2017. 2

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[79] 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 *ECCV*, 2018. 2

[80] Timo von Marcard, Bodo Rosenhahn, Michael Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*, pages 349–360, 2017. 2

[81] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13013–13022. IEEE, 2021. 2

[82] Kuan-Chieh Wang, Zhenzhen Weng, Maria Xenochristou, João Pedro Araújo, Jeffrey Gu, Karen Liu, and Serena Yeung. Nemo: Learning 3d neural motion fields from multiple video instances of the same action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22129–22138, 2023.

[83] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: implicit clothed humans obtained from normals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13286–13296. IEEE, 2022. 2

[84] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *CVPR*, pages 4967–4977, 2020. 2

[85] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 37(2):27:1–27:15, 2018. 2

[86] Ming Yan, Yewang Chen, Yi Chen, Guoyao Zeng, Xiaoliang Hu, and Jixiang Du. A lightweight weakly supervised learning segmentation algorithm for imbalanced image based on rotation density peaks. *Knowledge-Based Systems*, 244:108513, 2022. 5

[87] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12988, 2023. 3, 4

[88] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 6

[89] Ze Yang, Shenlong Wang, Siva Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *CVPR*, 2021. 2

[90] Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. Mime: Human-aware 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12965–12976, 2023. 3

[91] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM Transactions on Graphics (TOG)*, 42(4):1–17, 2023. 2

[92] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (PIP): physics-aware real-time human motion tracking from sparse inertial sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13157–13168. IEEE, 2022. 2

[93] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-

time 3d human translation and pose estimation with six inertial sensors. *ACM Trans. Graph.*, 40:86:1–86:13, 2021. 2

[94] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: global occlusion-aware human mesh recovery with dynamic cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11028–11039. IEEE, 2022. 2, 7, 8

[95] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 6

[96] Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4478–4487, 2022. 2, 5

[97] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[98] Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet Things J.*, 10(16):14128–14136, 2023. 3

[99] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: Unified pretraining for human motion analysis. In *ICCV*, 2023. 2

[100] 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. 3

[101] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. Eventhpe: Event-based 3d human pose and shape estimation. In *ICCV*, pages 10976–10985, 2021. 2, 7