EgoHumans: An Egocentric 3D Multi-Human Benchmark

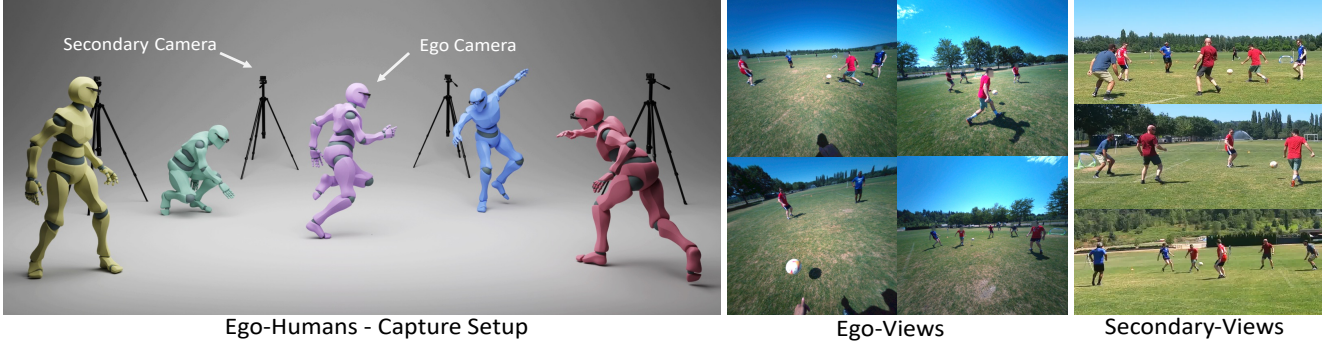Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, Kris Kitani

Figure 1: *Left*: Our proposed capture setup consists of multiple egocentric cameras from wearable glasses and stationary secondary cameras. This setup is flexible and mobile, allowing us to generate high-quality multi-human 3D annotations for diverse in-the-wild settings. *Center*: Multiple synchronized egocentric views while playing soccer. *Right*: Synchronized secondary views (cropped) from the stationary cameras. All cameras are spatiotemporally localized in the world coordinate.

Abstract

*We present **EgoHumans**, a new multi-view multi-human video benchmark to advance the state-of-the-art of egocentric human 3D pose estimation and tracking. Existing egocentric benchmarks either capture single subject or indoor-only scenarios, which limit the generalization of computer vision algorithms for real-world applications. We propose a novel 3D capture setup to construct a comprehensive egocentric multi-human benchmark in the wild with annotations to support diverse tasks such as human detection, tracking, 2D/3D pose estimation, and mesh recovery. We leverage consumer-grade wearable camera-equipped glasses for the egocentric view, which enables us to capture dynamic activities like playing soccer, fencing, volleyball, etc. Furthermore, our multi-view setup generates accurate 3D ground truth even under severe or complete occlusion. The dataset consists of more than 125k egocentric images, spanning diverse scenes with a particular focus on challenging and unchoreographed multi-human activities and fast-moving egocentric views. We rigorously evaluate existing state-of-the-art methods and highlight their limitations in the egocentric scenario, specifically on multi-human tracking. To address such limitations, we propose **EgoFormer**, a novel approach with a multi-stream transformer architecture and explicit 3D spatial reasoning to estimate and track the human pose. EgoFormer significantly outperforms prior art by 13.6% IDF1 and 9.3 HOTA on the EgoHumans dataset.*

1. Introduction

Understanding humans in 3D from the egocentric view is key to building immersive social telepresence [4, 62, 73, 77], assistive humanoid robots [28, 31, 91], and augmented reality systems [1, 10, 13]. A crucial step in this direction is to obtain 3D supervision at scale for deep learning models to generalize to the real world. However, unlike the large-scale 2D benchmarks [18, 23, 46, 64, 72], the diversity of the 3D benchmarks [48] is severely limited - primarily because manual annotation in the 3D space is impractical. As a result, existing popular 3D benchmarks [37, 43, 48, 66, 82, 110] are constrained to indoor environments or, at most, two human subjects if outdoors, stationary/slow camera motion, with limited occlusion. Furthermore, the majority of these benchmarks only portray the third-person view. Recent progress has been made in constructing egocentric benchmarks [35, 86, 115, 123]. However, they suffer from the same diversity pitfalls, making it difficult to evaluate how close the field is to fully robust and general solutions. To drive advances in the field, we propose a benchmark, *EgoHumans*, that includes challenging scenarios ignored in previous studies and a novel method, *EgoFormer*, that outperforms prior art as a starting point for the evaluations.

EgoHumans is a new egocentric benchmark consisting of high-resolution videos and comprehensive ground truth annotations such as camera parameters, 2D bounding boxes, human tracking ids [22], 2D/3D human poses, and 3D human meshes [74]. EgoHumans goes beyond previous bench-
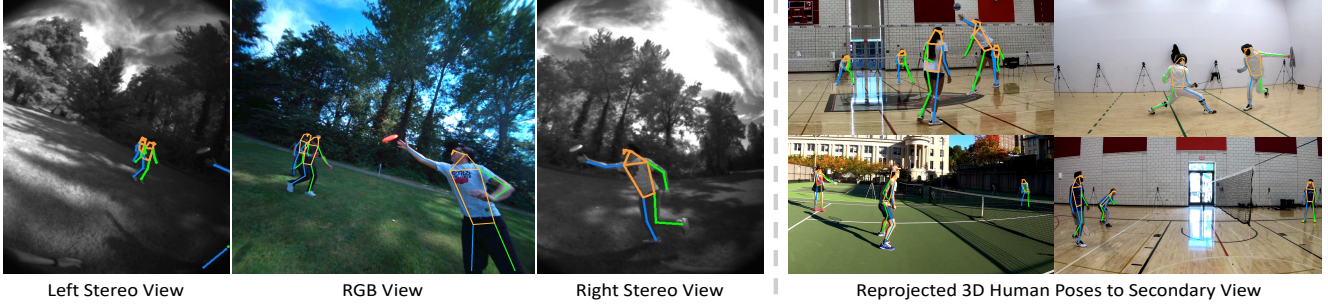
| Left Stereo View | RGB View | Right Stereo View | Reprojected 3D Human Poses to Secondary View |

Figure 2: *Left*: Our egocentric view consists of RGB, left and right stereo cameras, closely capturing the human eye's perspective. *Right*: EgoHumans consists of ground-truth 3D poses and person ids for multiple dynamic activities.

marks in important ways. First, it captures outdoor videos of unconstrained environments and dynamic human activities, including challenging sporting events such as fencing, badminton, volleyball, etc. Second, the activities are unchoreographed to truly capture the *in-the-wild* philosophy of our work. Our video sequences include fast ego-camera motion, human-human occlusion, truncation, and humans appearing at a wide range of spatial scales. We leverage a flexible multi-camera setup consisting of Meta's Aria glasses [76], with an RGB and two greyscale cameras, for the egocentric view and stationary secondary RGB cameras for the auxiliary views (see Fig. 2). Such camera combination allows us to accurately track and triangulate human poses in 3D for a long duration without using visual markers [43] or additional sensors [110]. The natural form factor of glasses [79] coupled with the RGB and stereo cameras closely resembles the human vision [81]. Last, as a by-product of our capture setup, we provide 3D annotations for the multi-view secondary cameras. We hope these annotations allow the ability to move fluidly between the egocentric and secondary perspectives [68] and inspire new research for holistic human understanding. To our knowledge, EgoHumans is the only multi-human 3D egocentric benchmark with these attributes.

We generate high-quality 3D ground truth by leveraging state-of-the-art visual-inertial odometry algorithm (VIO) [76], which is robust to fast head motion and sudden changes in the eye gaze - frequently observed in natural human behavior [121]. All the cameras in our multi-view capture are aligned to a single world coordinate system using procrustes alignment [75] of the camera poses. EgoHumans consists of 125k egocentric RGB images and 410k human instance annotations (Tab. 1) capturing high-energy activities in various locations, clothing, and lighting conditions with

severe occlusion. We annotate the tracking ids, bounding boxes, and 2D/3D human poses for all views using off-shelf estimators [45, 112] and manual supervision. With carefully calibrated camera parameters and the multi-view 2D poses for a video, we optimize for 3D skeletons using triangulation [44] and refinement constraints like constant limb length, joint symmetry, and temporal consistency [109]. Finally, we build an efficient multi-stage motion capture pipeline to fit the SMPL [74] body model to the 3D human skeletons.

The scale and diversity of the EgoHumans dataset allow unprecedented opportunities to evaluate and improve egocentric methods. Specifically, we evaluate existing methods for multi-human tracking. Our results show that prior art is susceptible to common failures like person-id switching due to rapid camera motion, occlusion, and unconstrained human activities. Inspired by this, we present *EgoFormer*, a novel 3D human tracking approach with multi-stream transformer architecture that effectively performs human depth reasoning in a camera-agnostic frame of reference. Our proposed method uses self-attention to aggregate multi-view spatial information from the RGB, left, and right stereo cameras simultaneously. EgoFormer significantly outperforms existing state-of-the-art tracking methods [12, 94, 125] by 13.6% IDF1 score on EgoHumans.

Our contributions are summarized as follows.

- *EgoHumans* is the first multi-human 3D egocentric dataset capturing unconstrained human activities in the wild. We provide high-quality 3D ground truth from egocentric and secondary views for all humans.

- We benchmark existing state-of-the-art methods for multi-human tracking and highlight their fundamental limitations on egocentric views.

- We propose *EgoFormer*, a 3D tracking method that uses a multi-stream spatial transformer encoder for depth reasoning from the ego view. Our method consistently outperforms the prior art on the EgoHumans test set.

Ego-Datasets	Location	Ego-Views	Sec-Views	Images	Instances	Mesh	World Co.
Mo2Cap2 [115]	indoor	1	0	15k	15k	✗	✗
You2Me [86]	indoor	1	0	150k	150k	✗	✗
HPS [35]	indoor	1	0	300k	320k	✓	✓
EgoBody [123]	indoor	1	5	199k	374k	✓	✓
EgoHumans	in/outdoor	6	15	125k	410k	✓	✓

Table 1: Comparison with 3D ego datasets. *Ego-Views* and *Sec-Views* are number of ego-views and secondary views. *Images* and *Instances* are number of ego-images and self + other visible human instances. *World Co.* refers to world translation and rotation.

2. Related Works

Biased 3D human benchmarks. Throughout the history of computer vision research, benchmarks [18, 23, 30, 40, 43,

64, 72, 114, 124] have played a critical role. However, unlike 2D benchmarks, 3D benchmarks [43, 48, 49, 82, 103, 110, 116] are limited in diversity which significantly hampers the ability of deep models to generalize to the real world [120]. In addition, 3D human poses are challenging for humans to annotate since the task requires metric precision. As a result, existing datasets rely on wearable sensors [110] or calibrated camera setups [43, 48, 66, 123] limited to indoor settings or are entirely synthetic [2, 89, 99, 107]. Popular datasets like Human3.6M [43], AMASS [78], HumanEva [103], AIST++ [66], HUMBI [119], PROX [37], and TotalCapture [49] only contain single human sequences. Multi-human datasets like PanopticStudio [48], MuCo-3DHP [83], TUM Shelf [14] are limited to indoor lab conditions. Outdoor multi-human datasets like 3DPW [110] and MuPoTS [83] have constrained human activities and lack egocentric annotations [5, 108], or are limited in diversity [109]. Existing egocentric datasets primarily focus on hand-object interactions and action recognition [3, 19, 20, 27, 53, 54, 56, 61, 67, 85, 87, 92, 101, 104, 118, 128]. Recent datasets like Mo2Cap2 [115], You2Me [86], HPS [35] and EgoBody [123] focus on 3D human pose annotations - but are limited to one or two human subjects and indoor settings. We showcase various statistics of EgoHumans against existing ego benchmarks in Tab.1 and highlight the key differences.

Monocular 3D human reconstruction. Among the recent approaches [15, 25, 34, 50, 55, 58, 65, 70, 71, 106, 122], many rely on the SMPL model [74], which offers a low dimensional parametrization of the human body. HMR [51] uses a neural network to regress the parameters of an SMPL body from a single image. Follow-up works like SPIN [60], ROMP [106], METRO [70], PARE [58], OCHMR [55], SPEC [59], and CLIFF [69] have improved the robustness of the original method in various ways by using additional information like body centers, camera parameters, segmentation mask, 2D pose, etc. Further, methods like VIBE [57], HMMR [52], MAED [111], and DynaBOA [33] predict 3D body parameters from videos. However, most methods require "full-body" images [90] and therefore lack robustness when body parts are occluded or truncated, as is often the case in egocentric videos. We also show that existing methods do not exhibit temporal consistency under fast ego-camera motion present in our EgoHumans benchmark.

Multi-object tracking. Multi-object tracking is a well-studied area, and we refer the readers to [17, 21, 26, 117] for a comprehensive summary. In this work, we focus on human tracking methods. Modern tracking methods [6, 125, 127] are primarily driven by bounding-box detections [29, 39, 97], motion models [8, 16, 63, 113], association algorithms [32, 84] or clustering [109]. Fundamentally, these methods use 2D representations like body centers [130], keypoints [93], and appearance [105] and

lack 3D reasoning crucial to resolving ambiguities posed by severe/complete object occlusion. Recent methods like [94, 126, 132] incorporate 3D pose information relative to the camera frame into tracking. However, these methods assume stationary/slow camera motion and are unsuitable for rapid ego-camera movement. Our proposed approach EgoFormer addresses this limitation and performs 3D association in a static global reference frame for tracking.

3. EgoHumans Dataset

In this section, we describe the data collection setup and annotation algorithms using spatially localized and synchronized multi-view videos. The goal is to design a semi-automatic pipeline to provide ground truth 3D human shapes and poses for egocentric videos. We propose solutions to associate the identities of the subjects consistently across time, compute the body poses, and recover the 3D trajectories of each person in a common frame of coordinates.

Data collection. For in-the-wild capture, we design a flexible and simple multi-view system with heterogeneous sensors, including multiple Aria glasses and GoPros for the egocentric and secondary views, respectively (c.f., Fig. 3a). The glass camera provides a natural human eye's perspective during the capture. The large number of secondary cameras at unique viewpoints ensures robust human pose estimation under occlusions and removes the restriction on the view direction of the subjects. Importantly, the volume created by our cameras is portable and can be moved across locations. Our captures typically consist of 2 to 6 egocentric views and 8 to 15 secondary views. For Aria glasses, we use 1408×1408 pixel resolution RGB images and 480×640 greyscale images. For GoPro cameras, we set the resolution to be 3840×2160. All cameras are synchronized.

Camera calibration and 3D localization. For the egocentric cameras, we obtain the intrinsic parameters of the custom lens from the factory calibration and the per-timestamp extrinsic parameters using state-of-the-art visual-inertial odometry (VIO) [76]. As the VIO algorithm only provides individual egocentric camera trajectories in an arbitrary coordinate system, we merge multiple egocentric camera trajectories together with the stationary secondary cameras into a single frame of reference by using procrustes-alignment [75] and structure-from-motion [102] (c.f.,Fig. 3b).

BBox, identity association and 2D human pose. In contrast to previous single-human datasets, for our multi-human sequences, solving for consistent person identity throughout the video and across views is a crucial task. We observed that existing state-of-the-art tracking algorithms [6, 12, 125] are prone to failure under occlusion and fast motion. To this end, we obtain an initial 3D region proposal for each subject using the egocentric camera's 3D location and approximating the subject by a 3D cylinder. Further, we obtain per view 2D
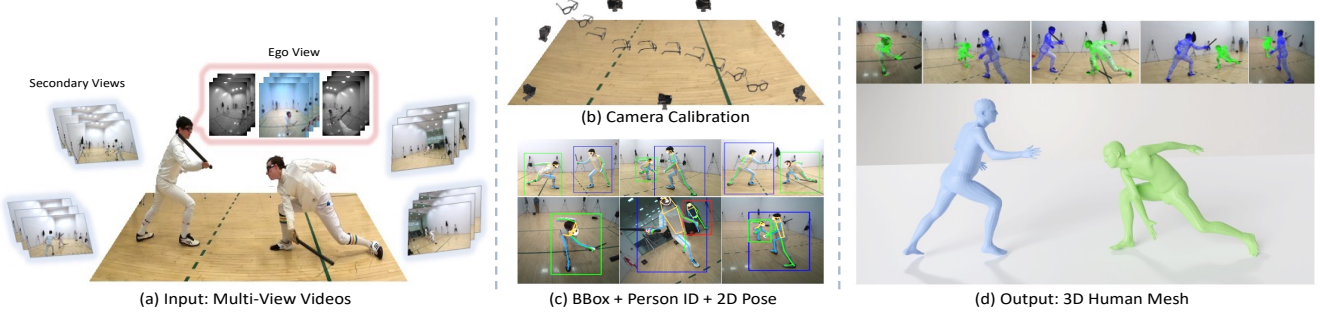
Figure 3: **Overview of EgoHumans data processing setup.** (*a*) Multiple synchronized secondary and ego cameras capture the sequence from multiple views. (*b*) We align secondary and ego cameras for all time steps into the world coordinate system. (*c*) For all views, we obtain bboxes, person ids, and 2D poses for all humans. (*d*) Reconstructed ground-truth meshes overlaid on multiple secondary and ego views.

bounding box (bbox) proposals and ground-truth person ids by reprojecting the 3D proposals (cylinders) to all ego and secondary views. We further refine these 2D bbox proposals using FasterRCNN [97] and manual supervision. Lastly, we annotate the 2D human poses in a top-down fashion for all the views using HRNet-WholeBody [45, 112] along with manual error fixes (c.f., Fig. 3c).

3D human pose. Let C be all synchronized video streams from egocentric and secondary cameras with known projection matrices P_c. We aim at estimating the global 3D pose $\mathbf{y}_{j,t} \in \mathbb{R}^3$ of a fixed set of human keypoints with indices $j \in (1..J)$ at timestamp $t \in (1..T)$ for all humans in the scene (we omit the human index for simplicity since each subject is processed independently). Let $\mathbf{x}_{c,j,t} \in \mathbb{R}^2$ be the jth 2D keypoint at time t from camera c.

To infer the 3D poses from their 2D estimates, we use a linear algebraic multi-view triangulation approach [36]. A naïve triangulation algorithm assumes that the 2D keypoints $\mathbf{x}_{c,j,t}$ from each view are independent and, thus, make equal contributions to the triangulation. However, in some views, the 2D keypoints cannot be estimated reliably (*e.g.*, due to occlusions or being out of frame), leading to unnecessary degradation of the final triangulation result. This is addressed by applying RANSAC; specifically, for time step t, we solve the over-determined system of equations on the homogeneous 3D coordinate vector of the 3D keypoint $\tilde{\mathbf{y}}_{j,t}$, $A_{j,t}\tilde{\mathbf{y}}_{j,t} = 0$, where $A_{j,t} \in \mathbb{R}^{2C' \times 4}$ matrix is composed of the components from the full projection matrices and $\mathbf{x}_{c,j,t}$ and C' is the cardinality of the camera inlier set after RANSAC. We further refine the per time step 3D pose estimates globally $\mathbf{y}_{\{1..T\}}$ by leveraging human pose priors like constant limb length, joint symmetry, and temporal smoothing [109]. Our cost function is given as

$$\mathcal{L}_{\text{pose3d}}(\mathbf{y}) = w_l\mathcal{L}_{\text{limb}}(\mathbf{y}) + w_s\mathcal{L}_{\text{symm}}(\mathbf{y})$$
$$+ w_t\mathcal{L}_{\text{temporal}}(\mathbf{y}) + w_i\mathcal{L}_{\text{reg}}(\mathbf{y}) \quad (1)$$

where $\mathbf{y} = \mathbf{y}_{\{1..T\}}$ and $\mathcal{L}_{\text{limb}}$, $\mathcal{L}_{\text{symm}}$, $\mathcal{L}_{\text{temporal}}$, $\mathcal{L}_{\text{reg}}$ denotes constant limb length, left-right joint symmetry, temporal smoothing, and regularization losses respectively.

w_l, w_s, w_t, w_i are scalar weights. Please refer to the supplemental for the loss definitions.

Mesh recovery. We represent the human mesh using the body pose and shape, $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\text{pose}}, \boldsymbol{\theta}_{\text{shape}}, \boldsymbol{\theta}_{\text{global}}]$, where $\boldsymbol{\theta}_{\text{pose}} \in \mathbb{R}^{23 \times 6}, \boldsymbol{\theta}_{\text{shape}} \in \mathbb{R}^{10}, \boldsymbol{\theta}_{\text{global}} \in \mathbb{R}^6$. The pose parameters $\boldsymbol{\theta}_{\text{pose}}$ are the 6D representation of the joint rotations [131] of the 23 body joints of the SMPL [74] body. The shape parameters $\boldsymbol{\theta}_{\text{shape}}$ represent the first 10 coefficients of the PCA shape space learnt from a corpus of registered scans. $\boldsymbol{\theta}_{\text{global}}$ consists of the global root orientation and translation of the body. Similar to [41, 109], we fit $\boldsymbol{\theta}$ to the entire 3D pose trajectory in a three-stage optimization scheme. In addition, we use the gender-specific $\boldsymbol{\theta}_{\text{shape}}$ latent space for a better fit to the 3D poses.

Note, as the mesh fitting procedure is highly underconstrained, obtaining a good initialization for $\boldsymbol{\theta}$ plays an important role in avoiding local minima. To this regard, we run CLIFF [69] on all the camera views and pick the mesh estimate with the lowest joint-reprojection-error (MPJPE) [110] as the initialization. Let $\Phi : \boldsymbol{\theta} \to \mathbf{y}$ be a differentiable mapping function that projects SMPL parameters $\boldsymbol{\theta}$ to corresponding 3D keypoints $\mathbf{y}$. We define the mesh fitting loss $\mathcal{L}_{\text{mesh}}(\boldsymbol{\theta})$ as follows,

$$\mathcal{L}_{\text{mesh}}(\boldsymbol{\theta}) = w_1||\mathbf{y} - \Phi(\boldsymbol{\theta})||_2 + w_2||\boldsymbol{\theta}_{\text{pose}}||_2$$
$$+ w_3\mathcal{L}_{\text{limb}}(\Phi(\boldsymbol{\theta})) + w_4\mathcal{L}_{\text{symm}}(\Phi(\boldsymbol{\theta}))$$
$$+ w_5\mathcal{L}_{\text{temporal}}(\Phi(\boldsymbol{\theta})) + w_6\mathcal{L}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_{\text{shape}}) \quad (2)$$

where $\mathcal{L}_{\boldsymbol{\beta}}$ is the Gaussian mixture shape prior loss[11], the term $||\boldsymbol{\theta}_{\text{pose}}||_2$ penalizes hyper-extensions of joints and $w_1..w_6$ are scalar weights. Other losses are the same as in eq.1. We optimize $\mathcal{L}_{\text{mesh}}$ iteratively in three stages. The first stage consists of optimizing $\boldsymbol{\theta}_{\text{global}}$, followed by $\boldsymbol{\theta}_{\text{shape}}$ in the second stage and lastly $\boldsymbol{\theta}_{\text{pose}}$ and $\boldsymbol{\theta}_{\text{global}}$ in the third stage. Our recovered meshes are pixel aligned across all egocentric as well as the secondary views (c.f., Fig. 3c).
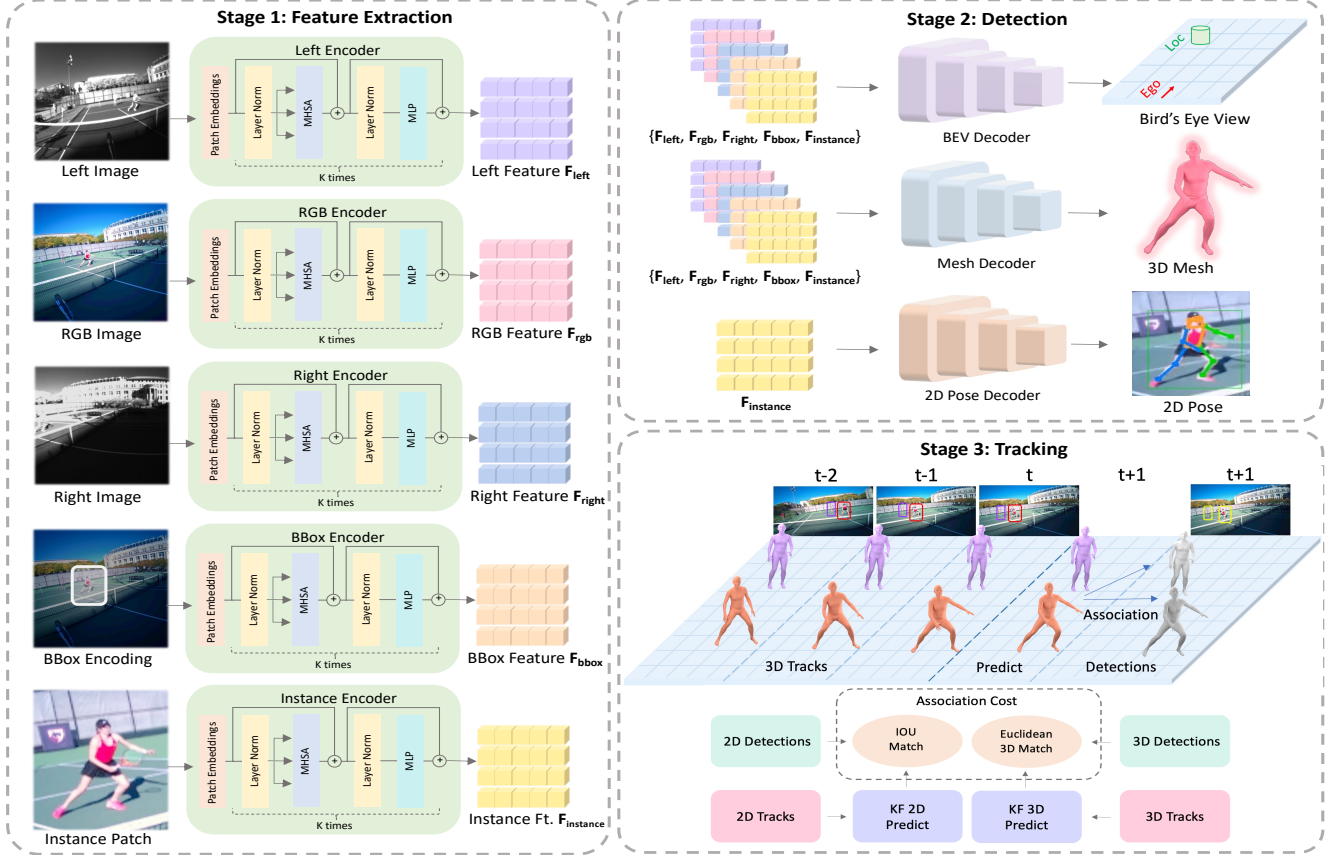
Figure 4: **Overall architecture of EgoFormer**. Stage 1 extracts multi-view features from the ego images for a human instance. Stage 2 decodes the 2D pose, bird's eye view heatmap of the 3D root location, and the mesh parameters from the features. Stage 3 tracks detections over time steps using two Kalman filters for 2D bounding boxes and 3D root locations.

4. EgoFormer Tracking

In this section, we present EgoFormer – a simple yet effective multi-stream transformer baseline to track multiple humans from an egocentric camera setup. The goal is to estimate the 3D shape and poses of each observed person over time using the RGB and two greyscale cameras. This task is challenging due to rapid head motion and frequent occlusions. The input to EgoFormer is three images, and the output includes the 2D detection, 2D/3D human poses, human shape per person, and identity associations across time. We assume the 3D poses of the camera have been reliably estimated from VIO [129]. Fig. 4 illustrates the three-stage algorithm design, which performs 3D bird's-eye view (BEV) reconstruction in the local camera coordinate and tracks human associations in the global world coordinates. Ego-Former can be trained end-to-end, and we experimentally found that it outperforms modular design variations [44]. We now explain each module in detail.

Stage 1 – feature extraction. As shown in Fig. 4, this stage extracts view-dependent features using multiple encoders that share the network architecture. First, the RGB image I_{rgb}, left/right greyscale images $I_{\text{left}}, I_{\text{right}} \in$

$\mathbb{R}^{H \times W \times 3}$ are encoded to the corresponding feature maps $\mathbf{F}_{\text{rgb}}, \mathbf{F}_{\text{left}}, \mathbf{F}_{\text{right}}$, respectively. We resize the images to the same resolution $H \times W$, and the greyscale images are concatenated three times channel-wise to standardize the number of channels. Following ViT [24], we first embed the images into tokens via a patch embedding layer. Then the tokens are processed by several transformer layers, where each contains a multi-head self-attention (MHSA) layer and a multi-layer perceptron (MLP) with residual connections. To help the later stage reasoning about the instance, we also perform 2D bbox detection using the off-shelf YOLOX [29] on the RGB image I_{rgb}. For each detected bbox, we further obtained two feature encodings. The first $\mathbf{F}_{\text{bbox}}$ is computed from a boolean representation of the bbox pixels I_{bbox}. The second $\mathbf{F}_{\text{instance}}$ is encoded from the cropped patch. Note, all the features $\mathbf{F} \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times K}$ where d is the downsampling ratio of the patch embeddings (*e.g.*, 16 by default), and K is the channel dimension.

Stage 2 – detection. We adopt three lightweight decoders to process the extracted features. Every decoder is composed of two deconvolution blocks, where each block contains one deconvolution layer followed by batch normalization [42]. The first decoder – BEV, predicts the target human instance's
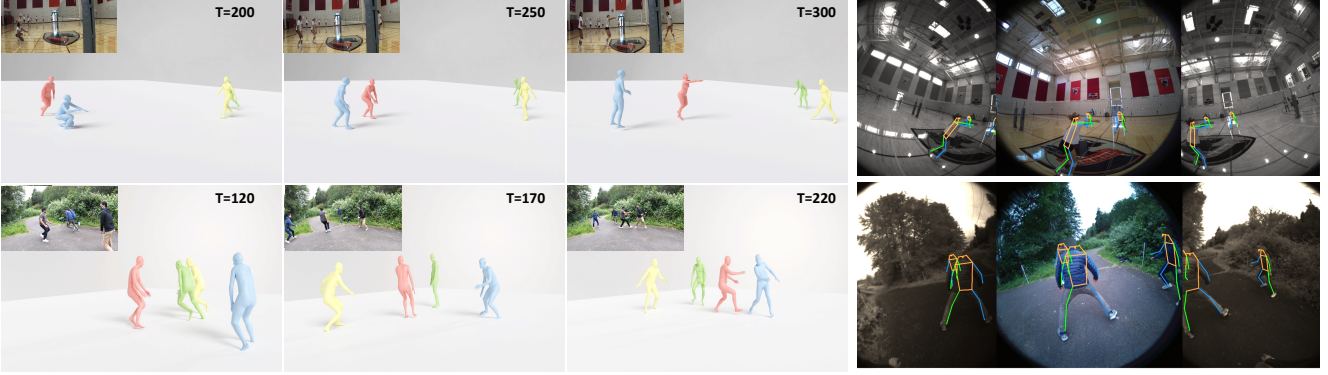
Figure 5: **Qualitative results for the EgoHumans data processing method.** *Left*: Estimated 3D human meshes with person ids (in color) for playing volleyball and tagging over a long duration of time. *Right:* We visualize the estimated 3D human poses for tennis and badminton by reprojection to the egocentric view. As shown, our multi-view system is robust to occlusion.

3D root location in an *unseen* bird's eye view in the local camera coordinate as a heatmap of $P \times Q$ spatial resolution. We use the log-polar (ρ, ϕ) to parameterize the root for the BEV heatmap, with a total of P bins for the $\log \rho$ and Q bins for the ϕ. The second decoder is used to regress the 3D SMPL shape and pose $\boldsymbol{\theta}$ using two fully-connected layers. The third pose decoder predicts the 2D keypoints, which regulates the learning by leveraging large-scale 2D annotations from datasets like COCO [72]. Please refer to the supplemental for the formal losses.

Stage 3 – tracking. We design a motion model-based tracker using Kalman Filters (KF) [9]. First, we transform the predicted 3D root locations from the local camera coordinates to the common world reference using the camera poses. Next, we use two filters, KF-2D and KF-3D, to temporally aggregate predictions for the bounding boxes and 3D root locations, respectively. The states $\mathbf{x}_{2D} = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]$ of KF-2D include the 2D coordinates of the bbox center (u, v), the bbox area s, a constant aspect ratio r, and the first derivative $\dot{u}, \dot{v}, \dot{s}$ with respect to time. The states $\mathbf{x}_{3D} = [x, y, z, \dot{x}, \dot{y}, \dot{z}]$ of KF-3D, include the 3D root location (x, y, z) and its velocity $(\dot{x}, \dot{y}, \dot{z})$. The association cost between detection and KF prediction is a weighted sum of IoU match between 2D bboxes and the Euclidean distances between the 3D root locations as illustrated in Fig. 4.

5. Experiments

In this section, we first describe the statistics and quantitative evaluations of the EgoHumans dataset. Then, we perform extensive benchmarking of multi-human tracking algorithms on our dataset. Further evaluations of 2D detection, 2D pose estimation, and 3D mesh reconstruction are presented in the supplementary.

5.1. EgoHumans Dataset

Data statistics. We collect 11 sequences from 24 subjects across 8 diverse locations (4 indoors and 4 outdoors). Two

sequences focus on casual social interactions like walking and running, whereas nine sequences focus on sports activities like basketball, fencing, soccer, badminton, tennis, volleyball, tagging, castle-build, and frisbee (c.f., Fig. 5). Each sequence has minimum 2 and maximum 6 subjects wearing Aria glasses and 8 – 15 secondary cameras, depending on the scenarios. In total, there are 125k RGB images, 250k greyscale images from egocentric glasses, and 446k images from GoPros. We divide each sequence into shorter clips of 30 seconds on average at 20 FPS. The annotation per time step includes the calibration and poses per camera, the bounding boxes, person ids, 2D/3D human poses, and 3D shapes per subject. Overall, EgoHumans captures 410k visible 2D instances. We split the dataset into 70130 egocentric images for `train`, 7130 for `val`, and 47740 for `test`. The split ensures non-overlapping locations between sets.

Annotation accuracy. We evaluate the end-to-end accuracy by comparing the output 3D human meshes against a dynamic *ground-truth point cloud* obtained by a Kaarta stencil LiDAR [80]. We register the point cloud to the scene using ICP [100] and manual correction. Tab. 3 reports the bidirectional Chamfer distance between the recovered 3D human meshes and the point cloud. We analyze the impact of the number of secondary cameras and different losses. As expected, increasing the number of cameras improves annotation. We found $\mathcal{L}_{\text{temp}}$ is the most impactful loss term.

5.2. EgoFormer Tracking

Implementation details. We follow the common practice to detect instances with YOLOX [29], and our network predicts the 3D root location, 2D pose and SMPL parameters per instance. The number of keypoints J is set to 17 [72]. The encoders are initialized with MAE-Base[38] pretrained weights. For the tracking stage, we adopt ByteTrack's association settings. For the tracking baselines, we use their MOT17 [21] configuration as default. The input resolution is set to 256×192. The model is trained for 210 epochs

Tracker	IDF1↑	HOTA↑	MOTA↑	MOTP↑	FP(10^4)↓	FN(10^4)↓	IDs↓	Rcll↑	Prcn↑
SORT [9]	25.2	18.5	20.3	73.8	4.35	1.12	7,996	85.7	60.8
DeepSORT [113]	38.3	28.1	22.7	73.8	4.35	1.12	6,087	85.7	60.8
CenterTrack [130]	39.8	29.9	35.7	74.0	3.25	1.17	4,862	85.8	63.7
FairMOT [127]	41.2	32.0	38.0	74.1	3.73	1.21	3,915	84.5	65.1
QDTrack [88]	45.5	32.6	43.8	74.3	3.11	1.21	1,074	84.6	68.2
Tracktor [6]	43.6	31.5	55.7	71.5	2.32	1.18	1,872	83.8	66.5
PHALP [94]	40.1	29.7	52.9	70.2	2.48	1.20	1,750	84.4	67.7
OCSORT [12]	46.4	36.7	54.6	78.9	2.44	0.82	3,486	89.2	74.4
ByteTrack [125]	49.7	38.8	59.5	**78.9**	**2.10**	0.82	2,696	89.5	**77.1**
SimpleBaseline (Ours)	60.9 (+11.2)	45.3 (+6.5)	59.1	78.8	2.30	0.79	1,203	89.9	75.5
EgoFormer (Ours)	**63.1 (+13.4)**	**48.1 (+9.3)**	**59.8**	78.8	2.30	**0.79**	**741**	**89.9**	75.5

Table 2: Results on the EgoHumans `test` set. We use the publicly released bounding box detector accompanying the respective methods on MOT17 for evaluation. Our proposed methods SimpleBaseline and EgoFormer, use the same detections as ByteTrack.

Cameras	$\mathcal{L}_{limb}$	$\mathcal{L}_{symm}$	$\mathcal{L}_{temp}$	$\mathcal{L}_{reg}$	Error (cm)
4	✓	✓	✓	✓	10.4
8	✓	✓	✓	✓	8.3
12	✗	✓	✓	✓	6.6
12	✓	✗	✓	✓	7.1
12	✓	✓	✗	✓	7.9
12	✓	✓	✓	✗	7.6
12	✓	✓	✓	✓	**5.8**

Table 3: 3D error (in cm) of our localization with a varying number of secondary cameras and 3D pose refinement losses. Increasing the number of cameras reduces the error due to better coverage. The relative importance of $\mathcal{L}_{temp}$ is greater than $\mathcal{L}_{limb}$ and $\mathcal{L}_{symm}$. with AdamW [96] with $5e-4$ learning rate, decayed by 10 at the 170th and 200th epoch on 8 A6000 GPUs on a combination of EgoHumans and COCO dataset. Note only *feature-extraction* and *detection* stages of EgoFormer have learnable parameters.

Metrics. To evaluate the 3D human tracking performance, we use the CLEAR metrics [7], including MOTA, FP, FN, IDs, *etc*. along with IDF1[98] and HOTA [47]. MOTA focuses on bbox detection accuracy. IDF1 evaluates the instance identity preservation and focuses on the association performance. Recently, HOTA is proposed, which explicitly

Tracker	IDF1↑	HOTA↑	MOTA↑	IDs↓
DeepSORT [113]	40.1	30.5	23.6	5,971
QDTrack [88]	46.1	34.1	44.6	1,342
Tracktor [6]	42.8	33.4	53.2	3,312
OCSORT [12]	47.2	37.9	56.1	2,430
ByteTrack [125]	51.5	40.6	59.7	1,203
EgoFormer (Ours)	**63.1**	**48.1**	**59.8**	**741**

Table 4: Comparison of EgoFormer with state-of-the-art tracking baselines. All methods are fine-tuned on the EgoHumans train set.

balances the effect of accurate detection and consistent association. As we predominantly use an off-shelf bbox detector in our experiments, HOTA and IDF1 are the primary metrics.

Baselines. We benchmark a wide set of algorithms on EgoHumans to analyze the state-of-the-art performance for egocentric multi-human tracking. These algorithms include SORT [9], DeepSORT [113], CenterTrack [130], FairMOT [127], QDTrack [88], Tracktor [6], PHALP [94], OCSORT [12] and ByteTrack [125]. For fair analysis, we used the published models. Since these algorithms only consider single RGB input without training from our data, we design a baseline – *SimpleBaseline* to be directly comparable. Specifically, we use the monocular depth estimator MiDaS [95] to predict a dense depth map for the RGB image. The 3D root location of each detection is obtained by averaging the depth of pixels inside the bbox. From here, the SimpleBaseline shares the same *tracking* stage as the EgoFormer, where the 3D root locations in the camera coordinates are transformed to the world coordinate via the camera poses to compute the KF-3D's state. Note the SimpleBaseline does not need to be trained on the EgoHumans dataset since the tracking stage contains no network parameters.

Results. Tab. 2 compares the performance of EgoFormer, SimpleBaseline, and other methods on the EgoHumans `test` set. EgoFormer significantly outperforms ByteTrack by 13.4% IDF1 and 9.3 HOTA highlighting the superior instance association. We observe that MOTA is comparable since the same detections are used. Our method also drastically reduces the identity switches to 741 compared to the prior art. EgoFormer implicitly learns to solve for person-id across views leveraging a larger field of view. We observe similar performance gains for SimpleBaseline over previous methods showcasing that tracking with the 3D association is crucial for the egocentric view. We show the qualitative tracking results in Fig. 6.
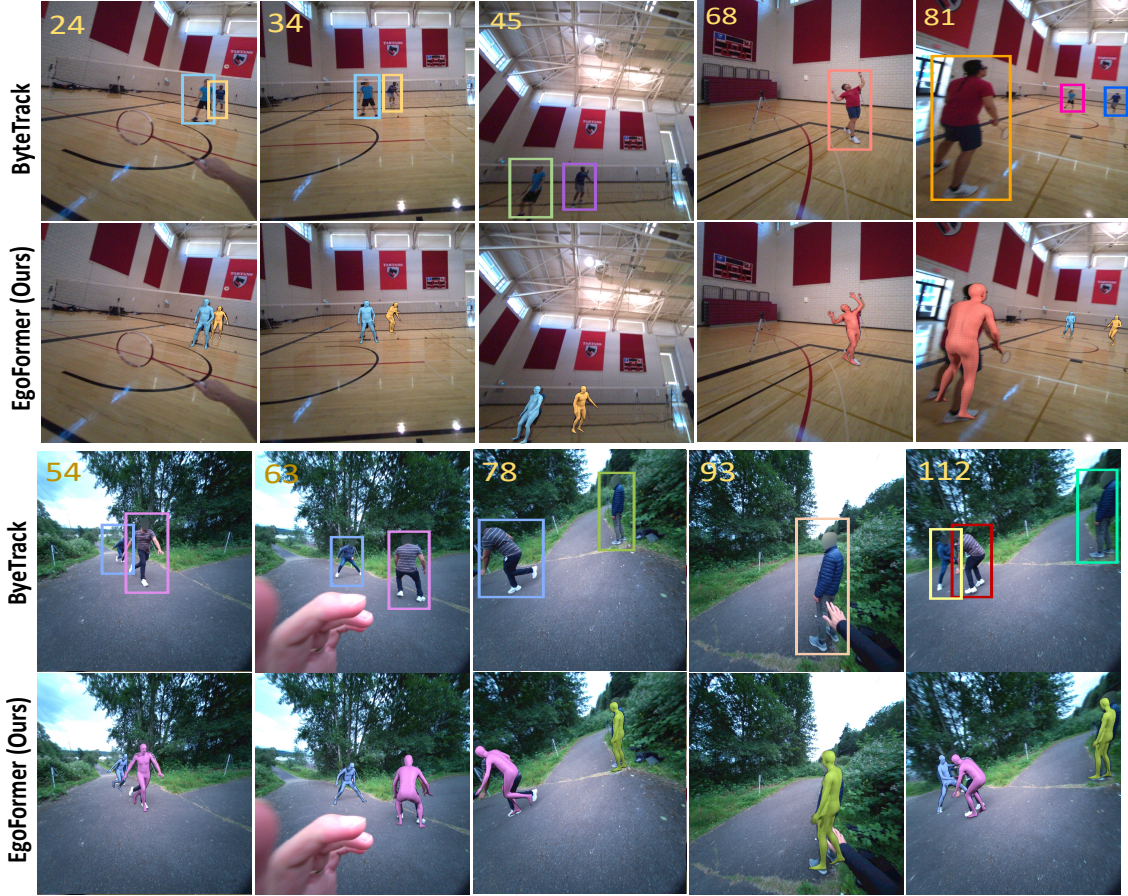
Figure 6: **Qualitative results of EgoFormer for tracking**. In comparison to a 2D approach like ByteTrack [125], our proposed method performs identity association in 3D and suffers from fewer identity switches. Predicted person ids are shown in color.

Finetuned baselines. To evaluate the effectiveness of our `train` set, we finetune tracking baselines on EgoHumans. We use the same hyperparameters provided by the authors for the MOT17 dataset for all the methods and combine MOT17 and EgoHumans `train` sets for training. As shown in Tab. 4, we observe the average gain of 2.1% IDF1 for all methods. However, baselines are still limited by their 2D nature, hence unsuitable for egocentric reasoning.

Ablations. Tab. 5 compares the performance of EgoFormer with a varying set of input images and decoders. Using all three images gives the best results due to a wider *field-of-view* aiding in better depth-reasoning. In addition, we find the 2D pose decoder with COCO annotations crucial for generalization to unseen scenes.

6. Discussion

We introduced a new in-the-wild 3D benchmark for detection, tracking, pose estimation, and mesh recovery of humans from egocentric captures. Emphasis was placed on capturing unchoreographed, dynamic activities in the real world. Our evaluations show that existing state-of-the-art methods are not suited for rapid camera motion present in wearable

Method	IDF1↑	HOTA↑	MOTA↑	IDs↓
RGB, $FOV = 110°$	58.1	40.5	59.5	2,139
RGB + Left, $FOV = 130°$	61.6	45.8	59.5	1,102
RGB + Right, $FOV = 130°$	61.2	45.2	59.4	1,164
w/o 2D Pose Decoder	56.8	38.2	59.4	2,877
w/o 3D Mesh Decoder	62.5	46.8	59.6	983
Full system, $FOV = 150°$	**63.1**	**48.1**	**59.8**	**741**

Table 5: **Ablation of the main components of EgoFormer**. We observe that increasing field-of-view and regularization using the 2D pose decoder reduces identity switches.

ego cameras. We believe that EgoHumans is a significant conceptual change for 3D datasets and will inspire a new research direction for egocentric methods. We also present EgoFormer, a simple 3D human tracker with multi-stream transformer architecture and explicit 3D spatial reasoning which outperforms existing methods by a significant margin. **Limitations.** At present, we trade off 3D keypoint localization accuracy in favor of an in-the-wild capture. There is still a particular gap between the performance of static indoor wired 3D capture systems and our capture setup due to errors in camera synchronization and calibration.

References

[1] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997. 1

[2] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 59–64, 2011. 3

[3] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE international conference on computer vision*, pages 1949–1957, 2015. 3

[4] Oluleke Bamodu and Xu Ming Ye. Virtual reality and virtual reality system components. In *Advanced materials research*, volume 765, pages 1169–1172. Trans Tech Publ, 2013. 1

[5] Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 4d visualization of dynamic events from unconstrained multi-view videos. In *CVPR*, 2020. 3

[6] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 3, 7

[7] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 7

[8] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 3

[9] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 6, 7

[10] Mark Billinghurst, Adrian Clark, Gun Lee, et al. A survey of augmented reality. *Foundations and Trends® in Human–Computer Interaction*, 8(2-3):73–272, 2015. 1

[11] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 4

[12] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*, 2022. 2, 3, 7

[13] Julie Carmigniani and Borko Furht. Augmented reality: an overview. *Handbook of augmented reality*, pages 3–46, 2011. 1

[14] Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, and Shuang Liu. Cross-view tracking for multi-human 3d pose estimation at over 100 fps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3279–3288, 2020. 3

[15] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, pages 769–787. Springer, 2020. 3

[16] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE international conference on computer vision*, pages 3029–3037, 2015. 3

[17] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020. 3

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2

[19] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 3

[20] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. 3

[21] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129(4):845–881, 2021. 3, 6

[22] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[25] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J Black. Learning to regress bodies from images using differentiable semantic rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11250–11259, 2021. 3

[26] Litong Fan, Zhongli Wang, Baigen Cail, Chuanqi Tao, Zhiyi Zhang, Yinling Wang, Shanwen Li, Fengtian Huang, Shuangfu Fu, and Feng Zhang. A survey on multiple object tracking algorithm. In *2016 IEEE International Conference on Information and Automation (ICIA)*, pages 1855–1862. IEEE, 2016. 3

[27] Alireza Fathi, Ali Farhadi, and James M Rehg. Understand-

ing egocentric activities. In *2011 international conference on computer vision*, pages 407–414. IEEE, 2011. 3

[28] Marina Fridin and Mark Belokopytov. Acceptance of socially assistive humanoid robot by preschool and elementary school teachers. *Computers in Human Behavior*, 33:23–31, 2014. 1

[29] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3, 5, 6

[30] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[31] Michael A Goodrich, Jacob W Crandall, and Emilia Barakova. Teleoperation and beyond for assistive humanoid robots. *Reviews of Human factors and ergonomics*, 9(1):175–226, 2013. 1

[32] Martin Grötschel and Olaf Holland. Solving matching problems with linear programming. *Mathematical Programming*, 33(3):243–259, 1985. 3

[33] Shanyan Guan, Jingwei Xu, Michelle Z He, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[34] Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10472–10481, 2021. 3

[35] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 1, 2, 3

[36] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[37] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2282–2292, 2019. 1, 3

[38] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 6

[39] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[40] Derek Hoiem, Santosh K Divvala, and James H Hays. Pascal voc 2008 challenge. *World Literature Today*, 24, 2009. 2

[41] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017. 4

[42] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 5

[43] 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(7):1325–1339, 2013. 1, 2, 3

[44] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. 2, 5

[45] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pages 196–214. Springer, 2020. 2, 4

[46] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Aberystwyth, UK, 2010. 1

[47] Luiten Jonathon, Osep Aljosa, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Leibe Bastian. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, 2021. 7

[48] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 1, 3

[49] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 3

[50] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. corr abs/1712.06584 (2017). *arXiv preprint arXiv:1712.06584*, 2017. 3

[51] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 3

[52] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 3

[53] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[54] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 3

[55] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1715–1725, 2022. 3

[56] Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro

Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR 2011*, pages 3241–3248. IEEE, 2011. 3

[57] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 3

[58] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. *arXiv preprint arXiv:2104.08527*, 2021. 3

[59] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11035–11045, 2021. 3

[60] 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. 3

[61] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 3

[62] Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G Desloge, Tommy Fortes, Eric M Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, et al. Project starline: A high-fidelity telepresence system. *CVPR*, 2021. 1

[63] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006. 3

[64] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019. 1, 3

[65] 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. 3

[66] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1, 3

[67] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 3

[68] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 2

[69] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. *arXiv preprint arXiv:2208.00571*, 2022. 3, 4

[70] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. 3

[71] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 3

[72] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 3, 6

[73] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1

[74] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 2, 3, 4

[75] Bin Luo and Edwin R Hancock. Iterative procrustes alignment with the em algorithm. *Image and Vision Computing*, 20(5-6):377–396, 2002. 2, 3

[76] Zhaoyang Lv, Edward Miller, Jeff Meissner, Luis Pesqueira, Chris Sweeney, Jing Dong, Lingni Ma, Pratik Patel, Pierre Moulon, Kiran Somasundaram, Omkar Parkhi, Yuyang Zou, Nikhil Raina, Steve Saarinen, Yusuf M Mansour, Po-Kang Huang, Zijian Wang, Anton Troynikov, Raul Mur Artal, Daniel DeTone, Daniel Barnes, Elizabeth Argall, Andrey Lobanovskiy, David Jaeyun Kim, Philippe Bouttefroy, Julian Straub, Jakob Julian Engel, Prince Gupta, Mingfei Yan, Renzo De Nardi, and Richard Newcombe. Aria pilot dataset. https://about.facebook.com/realitylabs/projectaria/datasets, 2022. 2, 3

[77] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. 1

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

[79] Andrew Maimone and Henry Fuchs. Computational augmented reality eyeglasses. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 29–38. IEEE, 2013. 2

[80] Miguel Ángel Maté-González, Vincenzo Di Pietra, and Marco Piras. Evaluation of different lidar technologies for the documentation of forgotten cultural heritage under forest environments. *Sensors*, 22(16):6314, 2022. 6

[81] Larry Henry Matthies. *Dynamic stereo vision*. Carnegie Mellon University, 1989. 2

[82] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild

using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 1, 3

[83] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 3

[84] G Ayorkor Mills-Tettey, Anthony Stentz, and M Bernardine Dias. The dynamic hungarian algorithm for the assignment problem with changing costs. *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27*, 2007. 3

[85] Sanath Narayan, Mohan S Kankanhalli, and Kalpathi R Ramakrishnan. Action and interaction recognition in first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–518, 2014. 3

[86] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9890–9900, 2020. 1, 2, 3

[87] Keisuke Ogaki, Kris M Kitani, Yusuke Sugano, and Yoichi Sato. Coupling eye-motion and ego-motion features for first-person activity recognition. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–7. IEEE, 2012. 3

[88] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 164–173, 2021. 7

[89] 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 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 3

[90] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1485–1495, 2022. 3

[91] Chiara Piezzo and Kenji Suzuki. Feasibility study of a socially assistive humanoid robot for guiding elderly individuals during walking. *Future Internet*, 9(3):30, 2017. 1

[92] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012. 3

[93] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4620–4628, 2019. 3

[94] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition, pages 2740–2749, 2022. 2, 3, 7

[95] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 7

[96] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. 7

[97] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 3, 4

[98] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 17–35. Springer, 2016. 7

[99] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 3

[100] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001. 6

[101] Michael S Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2730–2737, 2013. 3

[102] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3

[103] 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(1-2):4, 2010. 3

[104] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7396–7404, 2018. 3

[105] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022. 3

[106] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021. 3

[107] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the

IEEE conference on computer vision and pattern recognition, pages 109–117, 2017. 3

[108] Minh Vo, Yaser Sheikh, and Srinivasa G Narasimhan. Spatiotemporal bundle adjustment for dynamic 3d human reconstruction in the wild. *IEEE TPAMI*, 2020. 3

[109] Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa G Narasimhan. Self-supervised multi-view person association and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2794–2808, 2020. 2, 3, 4

[110] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 1, 2, 3, 4

[111] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13033–13042, 2021. 3

[112] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 2, 4

[113] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3, 7

[114] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 3

[115] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019. 1, 2, 3

[116] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, June 2022. 3

[117] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13–es, 2006. 3

[118] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2629–2638, 2016. 3

[119] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2990–3000, 2020. 3

[120] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *European Conference on Computer Vision*, pages 507–523. Springer, 2020. 3

[121] Han Zhang, Kevin F Miller, Xin Sun, and Kai S Cortina. Wandering eyes: Eye movements during mind wandering in video lectures. *Applied Cognitive Psychology*, 34(2):449–464, 2020. 2

[122] 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/CVF International Conference on Computer Vision*, pages 11446–11456, 2021. 3

[123] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision (ECCV)(Oct 2022)*, 2022. 1, 2, 3

[124] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2019. 3

[125] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 2, 3, 7, 8

[126] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenyu Liu, and Wenjun Zeng. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[127] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021. 3, 7

[128] Zehua Zhang, David Crandall, Michael Proulx, Sachin Talathi, and Abhishek Sharma. Can gaze inform egocentric action recognition? In *2022 Symposium on Eye Tracking Research and Applications*, pages 1–7, 2022. 3

[129] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7244–7251. IEEE, 2018. 5

[130] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 3, 7

[131] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 4

[132] Shihao Zou, Yuanlu Xu, Chao Li, Lingni Ma, Li Cheng, and Minh Vo. Snipper: A spatiotemporal transformer for simultaneous multi-person 3d pose estimation tracking and forecasting on a video snippet. *arXiv preprint arXiv:2207.04320*, 2022. 3