A Survey on 3D Egocentric Human Pose Estimation

Md Mushfiqur Azam, Kevin Desai
The University of Texas at San Antonio
{mdmushfiqur.azam, kevin.desai}@utsa.edu

Abstract

Egocentric human pose estimation aims to estimate human body poses and develop body representations from a first-person camera perspective. It has gained vast popularity in recent years because of its wide range of applications in sectors like XR-technologies, human-computer interaction, and fitness tracking. However, to the best of our knowledge, there is no systematic literature review based on the proposed solutions regarding egocentric 3D human pose estimation. To that end, the aim of this survey paper is to provide an extensive overview of the current state of egocentric pose estimation research. In this paper, we categorize and discuss the popular datasets and the different pose estimation models, highlighting the strengths and weaknesses of different methods by comparative analysis. This survey can be a valuable resource for both researchers and practitioners in the field, offering insights into key concepts and cutting-edge solutions in egocentric pose estimation, its wide-ranging applications, as well as the open problems with future scope.

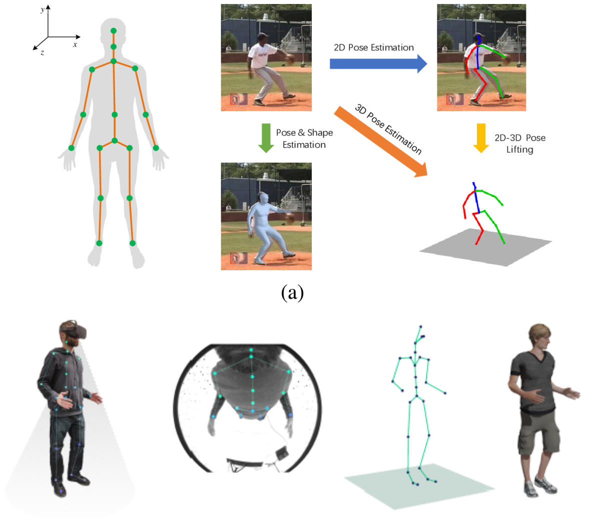
1. Introduction

Human pose estimation [10, 19, 55, 62] has gained prominence due to its relevance in numerous applications, ranging from animation and gaming to surveillance, healthcare, and human-computer interaction. The rise of wearable technology, including smart glasses, body-mounted cameras, and head-mounted displays has significantly fueled interest in egocentric pose estimation, where the focus is on estimating the pose of the person from the point of view of a wearable camera or device worn by the person (first person perspective). Egocentric pose estimation plays a crucial role across various domains, such as in human computer interaction for gesture recognition, augmented and virtual reality experiences by tracking body movements, healthcare for precise therapy monitoring, biomechanical analysis in sports training, hand-object interaction for contextual understanding, and enhancing realism in professional simulations through accurate movement replication. Unlike traditional pose estimation, which relies on external cameras or sensors, egocentric pose estimation offers a unique and immersive perspective on human body representation. Real-time processing, adaptability to different environments, user interaction mechanisms, including gestures, and semantic scene understanding contribute to the effectiveness of egocentric pose estimation systems. Figure 1 shows the difference between traditional and egocentric 3D human pose estimation.

Challenges for Egocentric 3D Human Pose Estimation stem from the complexity of accurately capturing and interpreting human movements from the first-person perspective. Some of the key challenges include:

- *Viewpoint Variations:* The use of egocentric cameras, attached to the body, introduces challenges in pose estimation as body parts may be occluded, particularly when hidden from view. The wide range of possible viewpoints in egocentric settings, involving varying camera angles, heights, and orientations, demands robust models to ensure accurate pose estimation across diverse scenarios.
- *Limited Depth Information:* Egocentric cameras, commonly mounted on wearable devices, capture scenes in 2D, lacking explicit depth details. This absence complicates the accurate determination of the distance of body parts from the camera, as 2D images may project objects at different distances onto the same plane.
- *Dataset Constraints:* In-the-wild datasets are essential for capturing real-world complexity, including variations in lighting, backgrounds, activities, and environments. However, their scarcity hinders model generalization, especially in dynamic environments with unpredictable situations. Limited availability of diverse samples, often from motion capture systems, poses challenges for models aiming at real-world outdoor applications.

Scope of the Survey: Currently, there are numerous systematic surveys related to 2D and 3D human pose estimation on traditional and deep learning based approaches [16, 45, 65, 78] as well shape recovery based approaches [40, 59]. While comprehensive reviews on hand pose [7] and action recognition [47] from egocentric vision are present, it is noteworthy that, to the best of our knowledge, no comprehensive survey on full body egocentric 3D pose



(a)

(b)

Figure 1. Difference between (a) traditional human pose estimation [27] and (b) egocentric human pose estimation [61]

estimation methods has been published to date. This absence underscores a notable gap in existing research, despite the increasing interest and advancement in this domain.

In this survey, we aim to explore the multifaceted aspects of 3D egocentric human pose estimation, by first describing the widely used datasets in Section 2. Next, in Section 3, we explore the different egocentric estimation methods by dividing them into two categories on the basis of output generation: skeletal based methods and human body shape based methods. Skeletal methods explore different methods which are mostly regression based (estimation of 3D joint co-ordinates) and heatmap based (estimation of 2D heatmaps). On the other hand, body shape based methods mainly generate human models using different shape recovery methods. Additionally, we present a comprehensive evaluation of egocentric pose estimation models, showcasing various evaluation metrics in Section 4 and a detailed performance analysis of state-of-the-art approaches on prominent datasets in Section 5. Lastly, we conclude the survey in Section 6 with some future research scopes for egocentric 3D human pose estimation.

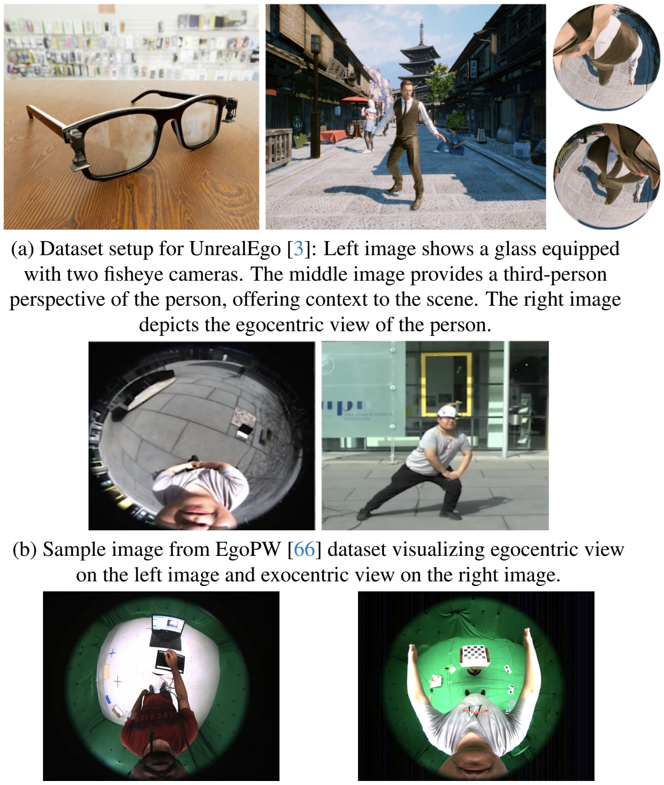
2. Datasets

Large scale dataset is one of the key factors in visualizing and analysing a computer vision problem. While benchmark datasets like MPII [6] and Human3.6M [26] exist for traditional human pose estimation, there's a notable gap for egocentric pose estimation benchmark datasets. Figure 2 showcases sample images from 4 different datasets. Table 1 summarizes the key features of 9 egocentric pose estimation datasets, with more details provided in the text below.

EgoCap [51] proposed a method for creating large training datasets using a marker-less motion capture system. They leveraged eight fixed cameras to estimate 3D skeleton motion. They projected it onto fisheye images from a head-mounted camera setup, enhancing the dataset with background replacement, clothing color variations, and simulated lighting changes. The training set includes 75,000 annotated fisheye images from six subjects and 25,000 images from two additional subjects for validation.

The ***Mo^{2}Cap^2*** dataset [68] tackles the challenge of obtaining annotated 3D pose data and introduces a marker-less multi-view motion capture. To address the time-consuming nature of obtaining diverse egocentric training examples, the dataset includes a synthetic training corpus generated from egocentric fisheye views. Built upon the SURREAL dataset [63], it offers 530,000 realistic training images with ground truth annotations of 2D and 3D joint positions.

xr-EgoPose [60] provides an extensive collection of 383,000 frames featuring individuals showcasing a rich diversity of skin tones, body shapes, clothing styles, set with various backgrounds and lighting scenarios. Scenes are randomly generated from mocap data, featuring realistic body types like skinny short to full tall versions and skin tones from white to black. Prioritizing photorealism, the synthetic dataset is created through Maya animation with mocap data and V-Ray's physically based rendering setup.



(a) Dataset setup for UnrealEgo [3]: Left image shows a glass equipped with two fisheye cameras. The middle image provides a third-person perspective of the person, offering context to the scene. The right image depicts the egocentric view of the person.

(b) Sample image from EgoPW [66] dataset visualizing egocentric view on the left image and exocentric view on the right image.

(c) Sample image from EgoGTA [67] dataset.

(d) Sample image from Wang et al.'s [64] dataset.

Figure 2. Sample images from different datasets used for egocentric human pose estimation.

Dataset	Year	No. of Images	No. of Subjects / Actions	Characteristics	Dataset Website
EgoCap [51]	2016	100,000	8 subjects	marker-less motion capture system; annotated.	Link
Mo^{2}Cap^2[68]	2019	530,000	3000 actions	annotated; 700 different body textures.	Link
xr-EgoPose [60]	2019	383,000	23 male and 23 female subjects; 9 actions	synthetic; scene is generated from randomized characters, environments, lighting rigs and animation.	Link
EgoBody [74]	2022	219,731	15 indoor scenes; 36 subjects	two subjects (camera wearer and interactee) involved in different interaction scenarios.	Link
EgoPW [66]	2022	318,000	10 subjects; 20 actions	in-the-wild real data; 20 different clothing styles.	Link
UnrealEgo [3]	2022	900,000	17 subjects; 30 actions	450k in-the-wild stereo views; Motions, 3D environments, spawning human characters.	Link
EgoGTA [67]	2023	320,000	101 different actions	synthetic; based on GTA-IM containing different daily motions and scene geometry.	Link
ECHP [39]	2023	30 video sequences; 75,000 frames	9 subjects; 10 daily actions	indoor and outdoor; real-world data.	Link
Ego-Exo4D [18]	2023	5625 video sequences; 1422 hours	839 subjects; 43 actions	131 different scenes in 13 different cities; comprises skilled human activities (e.g., sports, music, dance, bike repair).	Link

Table 1. Popular datasets for egocentric 3D human pose estimation.

EgoBody [74] captures 2-person interactions using a Microsoft HoloLens2 headset. It provides synchronized multimodal data, including RGB, depth, head, hand, and eye gaze tracking. With 125 sequences from 36 subjects in 15 scenes, it offers accurate 3D human shape, pose, and motion ground-truth. The dataset aims to explore the relationships between human attention, interactions, and motions, overcoming limitations of prior datasets, and advancing sociological and human-computer interaction research.

The ***EgoPW*** [66] dataset is the first in-the-wild human performance dataset captured by synchronized egocentric and external cameras. It features 10 actors, 20 clothing styles, and 20 actions from 318,000 frames organized into 97 sequences, along with the 3D poses as pseudo labels.

The ***UnrealEgo*** dataset [3] introduces robust egocentric 3D human motion capture with 17 diverse 3D models and over 45,000 motions in 14 environments. It has stereo fisheye images and depth maps capturing complex activities like breakdance and backflips. With metadata including 3D joint positions and camera details, it comprises 450,000 in-the-wild stereo views, showcasing wider joint position distributions compared to xR-EgoPose [60]. They followed up with the **UnrealEgo2** and **UnrealEgo-RW** datasets [4], which provide more views with diverse human motions.

The ***EgoGTA*** [67] dataset comprises of 320,000 frames across 101 sequences with distinct human body textures, by leveraging the diverse daily motions and ground truth scene geometry of GTA-IM [9]. The methodology involves fitting the SMPL-X [50] model to 3D joint trajectories from GTA-IM [9], followed by attaching a virtual fisheye camera to the forehead for generating synthetic images, semantic labels, and depth maps with and without the human body.

The ***ECHP*** [39] dataset consists of 65,000 training images, 10,000 validation images, and a test set with egocentric images and 3D ground truth from VICON Mocap. Egocentric poses are extracted using OpenPose [10] and human segmentation. Calibration and Aruco markers [17] aid in obtaining egocentric camera pose. The dataset has 30 sequences with 9 subjects, 20 textures, and 10 actions in various indoor/outdoor scenes. The test set provides generalization with 4 unseen subjects and 17,000 ground truth frames.

Ego-Exo4D [18] is a groundbreaking multimodal dataset and benchmark suite, offering the largest public collection of time-synchronized first and third-person videos captured by 839 individuals across 131 scenes in 13 cities. It is comprised of 1,422 hours of video, featuring both egocentric and multiple synchronized exocentric views. The **EgoPose** benchmark focuses on recovering 3D body and hand movements from egocentric videos. The task is to estimate 17 3D body joint positions and 21 3D joint positions per hand, following the MS COCO convention.

3. 3D Egocentric Pose Estimation Methods

After performing an extensive literature search for egocentric pose estimation, in this section, we discuss around 35 popular techniques by classifying them into two categories, namely skeletal and body shape based approaches. Skeletal-based 3D pose estimation methods [49, 56, 57] leverage the human skeleton representation to accurately track and infer 3D joint position and body movements. Human body shape based human body pose estimation methods [8, 23, 73] utilize a parametric model, such as SMPL [41] and SMPL-X [50], to accurately estimate 3D joint locations and body shapes. The retrieval of human body meshes is pivotal in supporting subsequent tasks like reconstructing clothed humans [70, 79], rendering [22], and modeling avatars [24, 80]. The sub-sections below expand on the different methods in each category. We have further subcategorized the methods based on some significant features, as highlighted in bold.

3.1. Skeletal Based Methods

In this section, we have provided details on the skeletal based egocentric pose estimation methods. Table 2 provides a brief overview of 17 such skeletal based methods.

Rhodin et al. [51] introduced a **marker-less** egocentric motion capture system using fisheye cameras embedded in a helmet or VR headset. The method employs a generative pose estimation framework with a ConvNet-based body part detector, ideal for VR applications needing natural move-

Skeletal based Methods	Year	Highlighted Characteristics	Dataset	Limitations	Code/Project Website Link
Egocap [51]	2016	First marker-less motion capture system; utilized pose estimation framework for fisheye views with a ConvNet based body-part detector.	EgoCap	No real-time prototype.	Project
Jiang et al. [28]	2017	Leveraged dynamic motion signatures and static scene structures to infer the invisible pose efficiently.	custom Kinect V2 dataset	Ambiguity in egocentric inputs due to unpredictable arm poses.	–
Mo^{2}Cap^2 [68]	2019	Real-time; disentangled 3D pose estimation, addressed 2D joint detection, camera-to-joint distances, and joint position recovery for accurate results and a precise 2D overlay.	Mo^{2}Cap^2	Scenes with severe occlusions.	Project
xr-EgoPose [60]	2019	Encoder-decoder model for VR headset images, addressing resolution differences in upper and lower body poses, with a dual-branch decoder preserving uncertainty information.	xR-EgoPose	Scenes with extreme occlusions and out-of-field view.	Code
You2Me [46]	2020	Inferred robust poses by incorporating static scene features, explicit second-person body interactions and utilizing dyadic interactions and dynamic first-person motion features.	You2Me	Scenerios where camera wearer is crouched and camera points towards the floor.	Code
EgoGlass [77]	2021	Utilized body part information for low-visible joints and tackling self-occlusion by preserving uncertainty information.	EgoGlass	Lower body estimation produces larger errors.	–
Zhang et al. [76]	2021	Implemented auto-calibration module with self-correction for fisheye cameras to rectify image distortions, ensuring alignment between 3D predictions and distorted 2D poses.	xR-EgoPose	Not evaluated in real-world setting.	–
Wang et al. [64]	2021	Spatio-temporal optimization framework that combines 2D and 3D keypoints, VAE-based motion priors and SLAM-based camera pose estimation for stable global body pose estimation.	Mo^{2}Cap^2, AMASS	Not evaluated in real-world setting.	Project
Wang et al. [66]	2022	Implemented weak supervision with spatio-temporal optimization and synthetic data with domain adaptation for better egocentric pose estimation.	EgoPW	Accuracy of pseudo labels constrained by in-the-wild capture system.	–
Akada et al. [3]	2022	Enhances 3D pose estimation by integrating a stereo-based 2D joint location estimation module with weight-sharing encoders and a multi-branch autoencoder for uncertainty capture.	UnrealEgo	Occlusions and complex motions scenerios.	Project
Ego+X [38]	2022	Dual-camera framework for 3D global pose estimation and social interaction characterization, leveraging visual SLAM and a Pose Refine Module (PRM) for spatial and temporal accuracy and characterizes social interactions based on global 3D poses.	ECHA	Camera localization robustness limited; temporal smoothing effectiveness not fully evaluated.	–
Wang et al. [67]	2023	First egocentric pose estimation framework, integrating depth estimation for occlusion handling in close interactions.	EgoGTA, EgoPW-Scene	Accuracy is constrained by depth estimation where scene is occluded by body.	Project
EgoFish3D [39]	2023	A self-supervised framework for egocentric 3D pose estimation, utilizing real-world data with three key modules: third person view, egocentric, and interactive modules, achieving accurate results without the need for ground truth annotations.	ECHP	Overlooked the significance of the perspective factor, which can convey valuable information about the 3D effect intensity.	–
Ego3DPose [33]	2023	A stereo matcher network and perspective embedding heatmap representation, independent learning of stereo correspondences and leveraging 3D perspective information.	UnrealEgo	Scenes with occlusions, distortions and real-world setting.	Code
Ego-STAN [48]	2023	Tackles fisheye distortion and self-occlusions in egocentric human pose estimation through a domain-guided spatio-temporal transformer, using 2D image representations, feature map tokens, and 3D pose estimation for accurate joint localization and uncertainty management.	xr-EgoPose	Scenes in real-world setting.	–
Dhamanaskar et al. [13]	2023	Utilized third-person view information, creating a self-supervised neural network that establishes a shared space for consistent 3D body pose detection across diverse video settings, ensuring adaptability to real-world scenarios with unknown camera configurations.	First2Third-Pose	Evaluation limited to two datasets; broader assessment needed for generalization.	Code
EgoFormer [36]	2023	Leveraged video context and establishing long-term temporal relationships. It addresses ambiguity in first-person videos, surpassing dynamic features, and introduces a novel motion clue representation for enhanced accuracy.	CMU Mocap [1]	Lack of real-world testing and limited model comparisons.	–

Table 2. Popular skeletal based egocentric 3D pose estimation methods.

ment and interaction. However, it was not able to attain **real-time** performance. To solve which, Mo^{2}Cap^2 [68] uses a two-scale location invariant convolutional network to detect 2D joints, accommodating perspective and radial distortions. It uses a location-sensitive distance module for estimating absolute camera-to-joint distances, and then recovers actual joint positions by back-projecting 2D detections. However, it struggles in scenes with severe occlusions.

EgoGlass [77] solves the **occlusion** problem by leveraging body part information for improved pose detection. The 2D module incorporates branches for heatmap and body part prediction, while the 3D module employs a pseudo-limb mask approach to handle occlusion in real-world images. This module also functions as an autoencoder for joint heatmaps, enhancing 3D body pose estimation and capturing uncertainty in 2D predictions across multiple views. [67] introduces an egocentric depth estimation network for predicting scene depth maps behind the human body using a wide-view egocentric fisheye camera, addressing occlusion caused by the human body through a depth-inpainting network. Additionally, a scene-aware pose estimation network was presented for 3D pose regression. [25] used a Vector Quantized-Variational AutoEncoder (VQ-VAE) to predict and optimize human pose, addressing the challenge of obscured lower body appearance. *xR-EgoPose* [60] and *Self-Pose* [61] uses an encoder-decoder architecture designed to improve accuracy in capturing upper and lower body poses from monocular images obtained via VR headset cameras. [60] employs a dual-branch decoder to address resolution discrepancies between the upper and lower body. It handles uncertainties in 2D joint locations by initially generating 2D heatmaps and subsequently using an autoencoder for 3D pose regression.

To solve the problem of **out-of-field-view**, [28] developed a method aiming to infer the invisible pose of a person in egocentric videos using dynamic motion signatures and static scene structures. By combining short-term and longer-term pose dynamics, the method utilizes classifiers to estimate pose probabilities and performs joint inference for a longer sequence. They extended the idea [29] by using both dynamic motion information from camera SLAM and occasionally visible body parts for robust ego pose estimation ensuring geometrical consistency. *EgoTAP* [32] addresses **out-of-view limbs** and self-occlusion issues in stereo egocentric 3D pose estimation by introducing a Grid ViT Heatmap Encoder and Propagation Network. The Grid ViT efficiently summarizes joint heatmaps, preserving spatial relationships. The Propagation Network utilizes skeletal information to predict 3D poses, improving accuracy for both visible and less visible joints.

To reduce the scarcity of **real-world datasets** from egocentric view, [66] proposed the use of weak supervision from an external viewpoint. The approach utilizes spatio-temporal optimization to generate accurate 3D poses for frames in the *EgoPW* dataset, using them as labels for training an egocentric pose estimation network. It also incorporates a synthetic dataset and employs domain adaptation to bridge the gap between synthetic and real data. [3] proposed a solution for egocentric pose estimation in an **unconstrained environment**. It uses a 2D joint location estimation module for stereo inputs by utilizing weight-sharing encoders and a decoder leveraging stereo information to boost performance. The 3D module comprises a multi-branch autoencoder, predicting 2D heatmaps to generate 3D pose and reconstructing heatmaps to capture uncertainty.

Perspective distortion can cause issues like scale variation, depth ambiguity and limited field of view. To tackle this problem, *Ego3DPose* [33] introduces a Stereo Matcher network that independently learns stereo correspondences and predicts explicit 3D orientation for each limb, avoiding dependence on full-body information. Additionally, a Perspective Embedding Heatmap representation is introduced, allowing the 2D module to extract and utilize 3D perspective information. [48] addressed the challenges of **fisheye distortion** and **self-occlusions** by leveraging a domain-guided spatio-temporal transformer model, *Ego-STAN*. It utilizes 2D image representations and spatiotemporal attention to mitigate distortions and accurately estimate the location of heavily occluded joints. [76] employed an **automatic calibration** module with self-correction to mitigate the impact of image distortions on 3D pose estimation. Unlike traditional post-processing steps, this module ensures consistency between 3D predictions and distorted 2D poses.

When the predicted poses are in the fisheye camera's **local coordinate system** instead of the global coordinate system, it can cause issues like **temporal instability**. To solve this issue, [64] proposed a method for precise and stable global body pose estimation in egocentric videos. It utilizes CNN-detected 2D and 3D keypoints, VAE-based motion priors, and SLAM-based camera pose estimation. This approach effectively tackles challenges like temporal jitters and tracking failures, significantly enhancing accuracy and stability in obtaining coherent body poses. *Ego+X* [38] proposed a framework with two cameras for 3D **global pose estimation** and **social interaction characterization**. The *Ego-Glo* module solves spatial and temporal errors using a dual-branch network and visual SLAM. Whereas, the *Ego-Soc* module performs egocentric social interaction characterization, including object detection and human-human interaction, based on the global 3D human poses.

Generating **3D ground truth** data using motion capture system is a cumbersome task. To alleviate this problem, *EgoFish3D* [39] proposed three modules: a third-person view module generating accurate 3D poses from external camera images, an egocentric module predicting 3D poses from a single fisheye image via self-supervised learning, and an interactive module estimating rotation differences between third-person and egocentric views. This method achieves self-supervised egocentric 3D pose estimation without ground truth annotations, leveraging a real-world dataset (ECHP) with synchronized third-person and egocentric images. **Linking first-person and third-person view** [37, 53, 69] plays a crucial role for better understanding wearer's action and poses. [12] used visual information from paired third-person videos to create a shared space where different views of the same pose are close together. They trained a special neural network to learn this shared space in a self-supervised manner, teaching it to distinguish if two views show the same 3D skeleton.

EgoFormer [36], a tansformer-based model for ego-pose estimation in AR and VR applications, addresses the ambiguity in first-person videos by leveraging video context and establishing long-term temporal relationships. It extracts effective temporal features, dynamic features, and introduces a novel representation for motion clues. *You2Me* [46] addresses the challenge of estimating the 3D body pose of the camera wearer by leveraging interactions with a **visible second person**. The key insight is that dyadic interactions between individuals help to learn temporal models for interlinked poses even when one person is largely **out**

Body Shape based Methods	Year	Highlighted Characteristics	Dataset	Limitations	Code/Project Website Link
Yuan et al. [71]	2018	Integrates control-based modeling, physics simulation, and imitation learning for ego-pose estimation, enabling domain adaptation by considering underlying physics dynamics.	CMU Mocap [1]	Indirect 2D evaluation may not capture full 3D accuracy; limited behaviors may hinder complex motion generalization.	–
Dittadi et al. [14]	2021	Variational autoencoders for generating human body poses from limited head and hand pose data; addressing challenges through specialized inference models.	AMASS [44]	Incomplete utilization of temporal history, constraints on body shape variation and reliance on assumed availability of hand signals.	–
CoolMoves [2]	2021	Achieves real-time, expressive full-body motion synthesis for avatars using limited input cues, dynamically fusing stylized examples from skilled performers, excelling in activities like dancing and fighting.	CMU MoCap	Limited sensing of legs and feet, resulting in lower body reconstruction jitters and reduced accuracy in foot-driven motions.	–
EgoRenderer [21]	2021	Renders full-body neural avatars from egocentric fisheye images - texture synthesis, pose construction, and neural image translation; addresses challenges of top-down view and distortions.	EgoRenderer	Incomplete joint estimation; unnatural motions in SMPL model animations and temporal instability in frame predictions.	Project
HPS [20]	2021	Integrates wearable sensors, IMUs and a head-mounted camera for precise 3D pose tracking in pre-scanned environments, eliminating drift with localization and scene constraints.	HPS	Lack of features and scene changes between static 3D scans and real images.	Project
Avatarposer [30]	2022	First learning-based method predicting full-body poses in world coordinates, leveraging transformer encoder and motion input from head and hands	CMU Mocap, AMASS	Sensitivity to inaccuracies and occlusions in hand tracking data.	Code
FLAG [5]	2022	Flow-based model for realistic 3D human body pose prediction with uncertainty estimates, enhancing prior work through high-quality pose generation and efficient latent variable sampling for optimization.	AMASS	Difficulty in generating complex lower-body poses due to sparse training data and lack of temporal data integration.	Project
Su et al. [54]	2022	A data framework transforms raw video into 3D pose, enriched by a lightweight Self-Perception Excitation (SPE) module for egocentric self-awareness.	Mocap dataset [72]	Dependency on MoCap data and synchronized third-person view videos may limit the method's real-world applicability.	–
EgoEgo [35]	2023	Ego body pose estimation using ego head pose estimation leveraging SLAM, and conditional diffusion for disentangled head and body pose estimation.	ARES	Evaluation on synthetic and relatively small real-world datasets.	Code
EgoHMR [75]	2023	Scene-conditioned diffusion approach using a physics-based collision score, realistic human-scene interaction, accurate estimation for visible body parts while enhancing diversity.	EgoBody	Limited temporal context for reconstructing egocentric human motions.	Code
EgoPoser [31]	2023	Used sparse motion sensor; mitigates overfitting with position-invariant prediction, adaptable to diverse body sizes, robust with hands out of view and reduces motion artifacts.	AMASS	Limited evaluation on diverse real-world scenarios.	–
SimpleEgo [11]	2024	Directly predicts joint rotations as matrix Fisher distributions, providing robust uncertainty estimation and realistic deployment prospects.	SynthEgo	Accuracy could be limited when large portions of the body are occluded in the image.	Project

Table 3. Popular body shape based egocentric 3D pose estimation models.

of the field view. The method incorporates dynamic first-person motion features, static first-person scene features, and second-person's body pose interaction features to explicitly account for the body pose of the camera wearer.

3.2. Body Shape Based Methods

In this section, we expand on the different human body shape-based egocentric pose estimation methods found in the literature. Out of them, 12 are highlighted in Table 3.

Dittadi et al. [14] used variational autoencoders to generate human body poses from **noisy head and hand pose data**. It addresses the challenge of predicting full body poses with limited information by training specialized inference models. Yuan et al. [71] employed control-based modeling with physics simulation and used imitation learning to acquire a video-conditioned control policy for ego-pose estimation. Traditional computer vision methods focus solely on motion kinematics [42] neglecting the underlying physics of dynamics [43]. Taking this into account, this framework allows domain adaptation, transferring the policy from simulation to real-world data. *CoolMoves* [2] is a VR system that has achieved real-time, expressive full-body motion synthesis for a user's avatar using limited input cues from VR systems. It delineates the prominent movements through dynamic fusion with stylized examples from skilled performers. The system excels in synthesizing upper and lower-body motions without explicit tracking cues, addressing challenges in activities like dancing and fighting.

To solve the problem of **top-down view distortions**, *EgoRenderer* [21] renderes full-body avatars from egocentric images by decomposing the process into texture synthesis, pose construction, and neural image translation [52, 58]. *The Human POSEitioning System (HPS)* [20] combines wearable sensors, IMUs, and a head-mounted camera to accurately track and integrate 3D human poses within pre-scanned environments. By fusing camera-based localization with IMU-based tracking and scene constraints, HPS achieves physically plausible motion estimation.

To address challenges like **body truncation** and **pose ambiguities**, [75] introduced a scene-conditioned diffusion model guided by a physics-based collision score, facilitating the generation of realistic human-scene interactions. It uses classifier-free training for flexibility in sampling, providing accurate estimations for visible body parts and diverse plausible results for unseen parts. [30] predicted full-body poses in world coordinates solely from motion input derived from the user's head and hands. Leveraging a transformer encoder, the method extracts deep features, distinguishing global motion from local joint orientations to facilitate pose estimation. *FLAG* [5] uses sparse input signals from head mounted devices and a flow-based generative model to predict full-body poses and provide uncertainty estimates for joints. [54] estimated 3D wearer poses from egocentric video, overcoming challenges of **body invisibility and complex motion**. They convert raw video to 3D pose, incorporating Self-Perception Excitation module for self understanding from egocentric view.

EgoEgo [35] uses monocular egocentric videos to estimate ego-head pose and generate ego-body pose, allowing **independent learning without paired datasets**. It combines monocular SLAM and transformer-based models for accurate ego-head pose estimation, employing a conditional diffusion model for full-body pose generation based on the predicted head pose. *SimpleEgo* [11] performs regression of probabilistic full-body pose parameters from head-mounted camera images. It directly predicts joint rotations, eliminating the need for iterative fitting processes or manual tuning. By representing joint rotations as matrix Fisher distributions, the model predicts confidence scores, allowing for robust uncertainty estimation. *AGRoL* [15] proposes a lightweight MLP-based diffusion model for realistic full-body motion synthesis from sparse tracking signals. [34] introduces affordable motion capture using smartwatches and head-mounted camera, integrating head poses for sparsity, tracking floor levels for outdoor settings, and optimizing motion with visual cues. *EgoPoser* [31] generates full-body pose estimation using sparse motion sensors, focusing on HMD-based ego-body pose estimation in large scenes. It addresses overfitting issues by emphasizing position-invariant prediction with a Global-in-Local motion decomposition strategy. Notably, it adapts to diverse body sizes and remains robust when hands are out of view.

4. Evaluation Metrics

In this section, we briefly describe the different metrics used to assess 3D egocentric human pose estimation methods.

MPJPE (Mean Per Joint Position Error) is a widely utilized metric which measures the mean error between all the predicted 3D joint positions and the ground truth positions, by calculating the Euclidean distance between them.

PA-MPJPE focuses on the individual pose accuracy by checking the alignment between the estimated pose and the ground truth pose of each frame using Procrustes analysis.

BA-MPJPE first resizes the bone lengths to a standard skeleton and then calculates the PA-MPJPE, providing a comprehensive evaluation by considering structural consistency in bone lengths and eliminating body scale influence.

Global MPJPE evaluates global joint position accuracy by aligning all poses within a batch to the ground truth, considering translation and rotation.

MPJRE (Mean Per Joint Rotation Error) and ***MPJVE (Mean Per Joint Velocity Error)*** compares the predicted and ground truth joints by calculating the average rotational and velocity disparity respectively.

Percentage of Correct Key-points (PCK) is a measure of accuracy that checks if the predicted keypoint and the actual joint are close within a specific distance limit. Typically, this distance threshold is set based on the size of the subject.

Head Translation & Orientation Error focuses on translation and rotational accuracy in head pose estimation respectively. The translation error is quantified using the mean Euclidean distance between predicted and ground truth head trajectories. Whereas, the orientation error is calculated using the Frobenius norm of the difference between the predicted and ground truth head rotation matrices.

Methods	Walking	Sitting	Crawling	Crouching	Boxing	Dancing	Stretching	Waving	Average
EgoFish3D [39]	60.9	42.1	65.0	82.7	79.0	55.5	59.1	94.5	66.8
Zhang et al. [76]	41.16	76.58	73.04	89.67	52.96	58.90	92.21	71.55	62.13
Mo^{2}Cap^2 [68]	38.41	70.94	94.31	81.90	48.55	55.19	99.34	60.92	61.40
xR-EgoPose [60]	38.39	61.59	69.53	51.14	37.67	42.10	58.32	44.77	48.16
SelfPose-UNet [61]	45.83	47.24	47.35	45.15	48.72	47.00	46.15	46.45	**46.61**

Table 4. Comparison of different skeletal based 3D egocentric pose estimation methods on Mo^{2}Cap^2 dataset using MPJPE (mm).

Methods	Game	Gest.	Greeting	Lower Stretch	Pat	React	Talk	Upper Stretch	Walk	All
xR-EgoPose [60]	56.0	50.2	44.6	51.5	59.4	60.8	43.9	53.9	57.7	58.2
SelfPose [61]	60.4	54.6	44.7	56.5	57.7	52.7	56.4	53.6	55.4	54.7
Zhang et al. [76]	36.8	34.1	36.7	50.1	57.2	34.4	32.8	54.3	52.6	50.0
EgoFish3D [39]	48.0	48.2	42.5	47.3	48.8	53.6	47.2	36.2	48.9	46.1
Ego-STAN [48]	33.1	31.6	36.9	38.9	29.2	29.6	29.7	44.3	40.9	40.4
EgoGlass [77]	32.8	30.5	33.7	35.5	45.7	33.2	27.0	40.1	37.4	**37.7**

Table 5. Comparison of different skeletal based 3D egocentric pose estimation methods on xR-EgoPose dataset using MPJPE (mm).

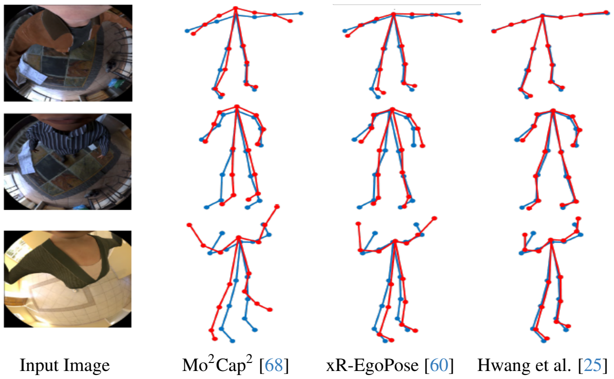


Figure 3. Qualitative comparison between three different state-of-the-art skeletal-based egocentric 3D pose estimation models on the xR-EgoPose dataset [60]. The predicted 3D poses (red) are superimposed onto the ground truth poses (blue).

5. Performance Analysis

In this section, we compare the performance of different state-of-the-art methods for the 3D egocentric human pose estimation on some of the popular egocentric datasets.

Performance of Skeletal-based Methods: Table 4 shows the performance of five different skeletal-based egocentric pose estimation methods across the different actions on the widely used Mo^{2}Cap^2 [68] dataset. The average MPJPE across all actions reduces from 66.8 mm in EgoFish3D [39] method to 46.61 mm in SelfPose-UNet [61] method. We see that 2D-3D lifting models [60, 61] achieved better results than direct 3D pose estimation methods, which may be due to the preserved uncertainty information of the joints. Figure 3 shows the qualitative evaluation of the 3D poses generated by three different skeletal-based methods on the xR-EgoPose [60] dataset. Table 5 compares six different skeletal based methods on the xR-EgoPose [60] dataset. Overall, the MPJPE of the methods is lower here than those tested in Mo^{2}Cap^2 [68] dataset, especially for actions with less visible joints. This could be because in this dataset most of the actions used for evaluation are relatively simpler.

Performance of body shape based Methods: Table 6 shows the performance of six different body shape based egocentric pose estimation methods on the AMASS [44] dataset. We can see that, AGRoL [15] outperforms other methods with its smooth motion generation, but it's limited to offline use. For real-time applications, EgoPoser [31] is more suitable as it provides more adaptability to diverse body sizes as well as robustness with hands out of view. Figure 4 qualitatively compares the 3D human pose and shape on three different body shape based methods using HPS [20] dataset.

Methods	MPJPE	MPJVE
CoolMoves [2]	7.83	100.54
Lee et al. [34]	5.87	19.11
AvatarPoser [30]	4.18	29.40
EgoPoser [31]	4.14	25.95
AGRoL [15]	3.86	50.94
AGRoL-Offline [15]	**3.71**	**18.59**

Table 6. Comparison of different body shape based 3D Egocentric Pose Estimation methods on AMASS [44] dataset using MPJPE (cm) and MPJVE (cm/s).

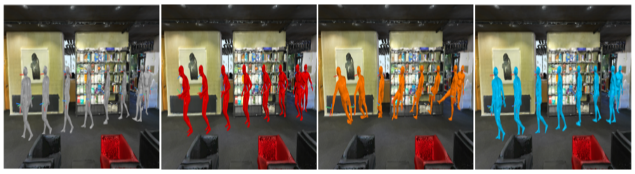


Figure 4. Qualitative comparison of three different state-of-the-art body shape based egocentric 3D pose estimation models on the HPS dataset. [20].

6. Conclusion and Future Directions

In this survey paper, we provide an overview of 3D egocentric human pose estimation using RGB images or video sequences, encompassing diverse datasets and estimation methodologies. Researchers have proposed diverse datasets with lightweight setups. However, the lack of standardized **benchmark datasets**, except for the recent Ego-Exo4D [18] dataset, poses a challenge for evaluating the robustness of different egocentric pose estimation models. While discussing the individual strengths and weaknesses of different skeletal and body shape based methods for egocentric pose estimation, we realize that most of the existing methods encounter difficulties with **in-the-wild scenarios** mainly due to insufficient training data. Notably, similar to traditional pose estimation, the biggest challenges of egocentric pose estimation models are strong occlusions and limited field of view, especially for the lower body joints. **Multi view consistency** may help to to solve this using additional 3D information. Moreover, **temporal and contextual information** can be utilized further to improve the robustness of the models considering these issues. Consequently, there exists ample scope for refining egocentric pose estimation approaches to better suit **real-time** technologies.

In conclusion, this survey paper serves as a comprehensive resource for researchers seeking to explore the existing egocentric pose estimation methods, understand prevalent challenges, and make further advancements.

7. Acknowledgements

This material is partially based upon work supported by the National Science Foundation under Grant No. 2153249 and 2316240. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] CMU Motion Capture Database. http://mocap.cs.cmu.edu. Accessed: March 2, 2024. 4, 6

[2] Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D Wilson. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5 (2):1–23, 2021. 6, 8

[3] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 2, 3, 4, 5

[4] Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. 3d human pose perception from egocentric stereo videos, 2023. 3

[5] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13253–13262, 2022. 6, 7

[6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[7] Andrea Bandini and José Zariffa. Analysis of the hands in egocentric vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

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

[9] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 3

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 3

[11] Hanz Cuevas-Velasquez, Charlie Hewitt, Sadegh Aliakbarian, and Tadas Baltrušaitis. Simpleego: Predicting probabilistic body pose from egocentric cameras. *arXiv preprint arXiv:2401.14785*, 2024. 6, 7

[12] Ameya Dhamanaskar, Mariella Dimiccoli, Enric Corona, Albert Pumarola, and Francesc Moreno-Noguer. Enhancing egocentric 3d pose estimation with third person views. *Pattern Recognition*, 138:109358, 2023. 5

[13] Ameya Dhamanaskar, Mariella Dimiccoli, Enric Corona, Albert Pumarola, and Francesc Moreno-Noguer. Enhancing egocentric 3d pose estimation with third person views. *Pattern Recognition*, 138:109358, 2023. 4

[14] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11687–11697, 2021. 6

[15] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 481–490, 2023. 7, 8

[16] Miniar Ben Gamra and Moulay A Akhloufi. A review of deep learning techniques for 2d and 3d human pose estimation. *Image and Vision Computing*, 114:104282, 2021. 1

[17] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014. 3

[18] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023. 3, 8

[19] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 1

[20] 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. 6, 8

[21] Tao Hu, Kripasindhu Sarkar, Lingjie Liu, Matthias Zwicker, and Christian Theobalt. Egorenderer: Rendering human avatars from egocentric camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14528–14538, 2021. 6

[22] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. In *2022 International Conference on 3D Vision (3DV)*, pages 197–208. IEEE, 2022. 3

[23] 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 *2017 international conference on 3D vision (3DV)*, pages 421–430. IEEE, 2017. 3

[24] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 3

[25] Juheon Hwang and Jiwoo Kang. Double discrete representation for 3d human pose estimation from head-mounted camera. In *2024 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–4. IEEE, 2024. 4, 8

[26] 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. 2

[27] Xiaopeng Ji, Qi Fang, Junting Dong, Qing Shuai, Wen Jiang, and Xiaowei Zhou. A survey on monocular 3d human pose estimation. *Virtual Reality & Intelligent Hardware*, 2:471–500, 2020. 2

[28] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509. IEEE, 2017. 4, 5

[29] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10986–10994. IEEE, 2021. 5

[30] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European Conference on Computer Vision*, pages 443–460. Springer, 2022. 6, 7, 8

[31] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time ego-body pose estimation in large scenes, 2023. 6, 7, 8

[32] Taeho Kang and Youngki Lee. Attention-propagation network for egocentric heatmap to 3d pose lifting. *arXiv preprint arXiv:2402.18330*, 2024. 5

[33] Taeho Kang, Kyungjin Lee, Jinrui Zhang, and Youngki Lee. Ego3dpose: Capturing 3d cues from binocular egocentric views. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 4, 5

[34] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. *arXiv preprint arXiv:2401.00847*, 2024. 7, 8

[35] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 6, 7

[36] Tianyi Li, Chi Zhang, Wei Su, and Yuehu Liu. Egoformer: Transformer-based motion context learning for egopose estimation. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4052–4057. IEEE, 2023. 4, 5

[37] 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. 5

[38] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yao Guo, and Guang-Zhong Yang. Ego+ x: An egocentric vision system for global 3d human pose estimation and social interaction characterization. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5271–5277. IEEE, 2022. 4, 5

[39] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yijun Chen, Yao Guo, and Guang-Zhong Yang. Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning. *IEEE Transactions on Multimedia*, 25:8880–8891, 2023. 3, 4, 5, 7, 8

[40] Yang Liu, Changzhen Qiu, and Zhiyong Zhang. Deep learning for 3d human pose estimation and mesh recovery: A survey. *arXiv preprint arXiv:2402.18844*, 2024. 1

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

[42] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, Shun Iwase, and Kris M Kitani. Kinematics-guided reinforcement learning for object-aware 3d ego-pose estimation. *arXiv preprint arXiv:2011.04837*, 2020. 6

[43] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34:25019–25032, 2021. 6

[44] 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. 6, 8

[45] Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, 8:133330–133348, 2020. 1

[46] 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. 4, 5

[47] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. Egocentric vision-based action recognition: A survey. *Neurocomputing*, 472:175–197, 2022. 1

[48] Jinman Park, Kimathi Kaai, Saad Hossain, Norikatsu Sumi, Sirisha Rambhatla, and Paul Fieguth. Domain-guided spatio-temporal self-attention for egocentric 3d pose estimation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1837–1849, 2023. 4, 5, 7

[49] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7307–7316, 2018. 3

[50] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3

[51] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 2, 3, 4

[52] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[53] Bilge Soran, Ali Farhadi, and Linda Shapiro. Action recognition in the presence of one egocentric and multiple static cameras. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V 12*, pages 178–193. Springer, 2015. 5

[54] Wei Su, Yuehu Liu, Shasha Li, and Zerun Cai. Proprioception-driven wearer pose estimation for egocentric video. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3728–3735. IEEE, 2022. 6, 7

[55] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 1

[56] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE international conference on computer vision*, pages 2602–2611, 2017. 3

[57] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*, 2016. 3

[58] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 6

[59] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2023. 1

[60] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7728–7738, 2019. 2, 3, 4, 5, 7, 8

[61] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando de la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6794–6806, 2023. 2, 5, 7, 8

[62] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[63] 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. 2

[64] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11500–11509, 2021. 2, 4, 5

[65] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021. 1

[66] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13157–13166, 2022. 2, 3, 4, 5

[67] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13031–13040, 2023. 2, 3, 4

[68] 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. 2, 3, 4, 7, 8

[69] 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. 5

[70] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296, 2018. 3

[71] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 735–750, 2018. 6

[72] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019. 6

[73] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2148–2157, 2018. 3

[74] 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*, pages 180–200. Springer, 2022. 3

[75] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. *arXiv preprint arXiv:2304.06024*, 2023. 6

[76] Yahui Zhang, Shaodi You, and Theo Gevers. Automatic calibration of the fisheye camera for egocentric 3d human pose estimation from a single image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1772–1781, 2021. 4, 5, 7

[77] Dongxu Zhao, Zhen Wei, Jisan Mahmud, and Jan-Michael Frahm. Egoglass: Egocentric-view human pose estimation from an eyeglass frame. In *2021 International Conference on 3D Vision (3DV)*, pages 32–41. IEEE, 2021. 4, 7

[78] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023. 1

[79] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184, 2021. 3

[80] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4): 1–19, 2023. 3