Scene-aware Egocentric 3D Human Pose Estimation

Jian Wang[1,2] Diogo Luvizon[1,2] Weipeng Xu[3] Lingjie Liu[1,2]
Kripasindhu Sarkar[4] Christian Theobalt[1,2]
[1]MPI Informatics [2]Saarland Informatics Campus [3]Meta Reality Labs [4]Google
{jianwang,dluvizon,lliu,theobalt}@mpi-inf.mpg.de xuweipeng@fb.com krsarkar@google.com

Abstract

Egocentric 3D human pose estimation with a single head-mounted fisheye camera has recently attracted attention due to its numerous applications in virtual and augmented reality. Existing methods still struggle in challenging poses where the human body is highly occluded or is closely interacting with the scene. To address this issue, we propose a scene-aware egocentric pose estimation method that guides the prediction of the egocentric pose with scene constraints. To this end, we propose an egocentric depth estimation network to predict the scene depth map from a wide-view egocentric fisheye camera while mitigating the occlusion of the human body with a depth-inpainting network. Next, we propose a scene-aware pose estimation network that projects the 2D image features and estimated depth map of the scene into a voxel space and regresses the 3D pose with a V2V network. The voxel-based feature representation provides the direct geometric connection between 2D image features and scene geometry, and further facilitates the V2V network to constrain the predicted pose based on the estimated scene geometry. To enable the training of the aforementioned networks, we also generated a synthetic dataset, called EgoGTA, and an in-the-wild dataset based on EgoPW, called EgoPW-Scene. The experimental results of our new evaluation sequences show that the predicted 3D egocentric poses are accurate and physically plausible in terms of human-scene interaction, demonstrating that our method outperforms the state-of-the-art methods both quantitatively and qualitatively.

1. Introduction

Egocentric 3D human pose estimation with head- or body-mounted cameras is extensively researched recently because it allows capturing the person moving around in a large space, while the traditional pose estimation methods can only record in a fixed volume. With this advantage, the egocentric pose estimation methods show great potential in various applications, including the xR technologies and mobile interaction applications.

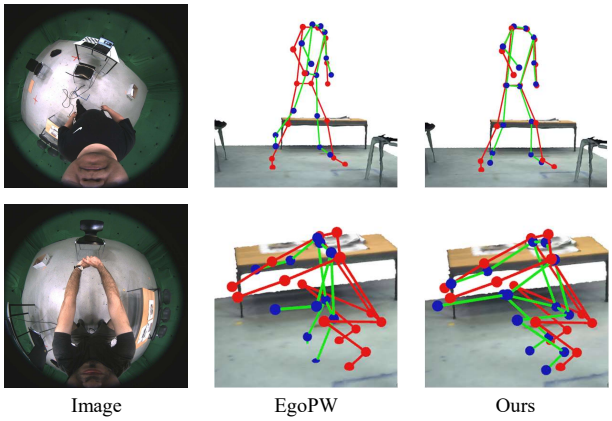

Figure 1. Previous egocentric pose estimation methods like EgoPW predict body poses that may suffer from body floating issue (the first row) or body-environment penetration issue (the second row). Our method predicts accurate and plausible poses complying with the scene constraints. The red skeletons are the ground truth poses and the green skeletons are the predicted poses.

In this work, we estimated the full 3D body pose from a single head-mounted fisheye camera. A number of works have been proposed, including Mo^{2}Cap^2 [39], xR-egopose [32], Global-EgoMocap [36], and EgoPW [35]. These methods have made significant progress in estimating egocentric poses. However, when taking account of the interaction between the human body and the surrounding environment, they still suffer from artifacts that contrast the physics plausibility, including body-environment penetrations or body floating (see the EgoPW results in Fig. 1), which is mostly ascribed to the ambiguity caused by the self-occluded and highly distorted human body in the egocentric view. This problem will render restrictions on subsequent applications including action recognition, human-object interaction recognition, and motion forecasting.

To address this issue, we propose a scene-aware pose estimation framework that leverages the scene context to constrain the prediction of an egocentric pose. This frame-

work produces accurate and physically plausible 3D human body poses from a single egocentric image, as illustrated in Fig. 1. Thanks to the wide-view fisheye camera mounted on the head, the scene context can be easily obtained even with only one egocentric image. To this end, we train an egocentric depth estimator to predict the depth map of the surrounding scene. In order to mitigate the occlusion caused by the human body, we predict the depth map including the visible human and leverage a depth-inpainting network to recover the depth behind the human body.

Next, we combine the projected 2D pose features and scene depth in a common voxel space and regress the 3D body pose heatmaps with a V2V network [22]. The 3D voxel representation projects the 2D poses and depth information from the distorted fisheye camera space to the canonical space, and further provides direct geometric connection between 2D image features and 3D scene geometry. This aggregation of 2D image features and 3D scene geometry facilitates the V2V network to learn the relative position and potential interactions between the human body joints and the surrounding environment and further enables the prediction of plausible poses under the scene constraints.

Since no available dataset can be used for train these networks, we proposed EgoGTA, a synthetic dataset based on the motion sequences of GTA-IM [3], and EgoPW-Scene, an in-the-wild dataset based on EgoPW [35]. Both of the datasets contain body pose labels and scene depth map labels for each egocentric frame.

To better evaluate the relationship between estimated egocentric pose and scene geometry, we collected a new test dataset containing ground truth joint positions in the egocentric view. The evaluation results on the new dataset, along with results on datasets in Wang *et al*. [36] and Mo^2Cap^2 [39] demonstrate that our method significantly outperforms existing methods both quantitatively and qualitatively. We also qualitatively evaluate our method on in-the-wild images. The predicted 3D poses are accurate and plausible even in challenging real-world scenes. To summarize, our contributions are listed as follows:

- The first scene-aware egocentric human pose estimation framework that predicts accurate and plausible egocentric pose with the awareness of scene context;
- Synthetic and in-the-wild egocentric datasets containing egocentric pose labels and scene geometry labels;[1]
- A new depth estimation and inpainting networks to predict the scene depth map behind the human body;
- By leveraging a voxel-based representation of body pose features and scene geometry jointly, our method outperforms the previous approaches and generates plausible poses considering the scene context.

[1]Datasets are released in our project page. Meta did not access or process the data and is not involved in the dataset release.

2. Related Work

2.1. Egocentric 3D Full Body Pose Estimation

Inspired by the new applications in augmented reality and by the limitations of traditional motion capture systems, Rhodin *et al*. [26] proposed the first egocentric motion capture system based on a pair of fisheye cameras. The following methods proposed new architectures [4,47] and new datasets [1,47] for stereo egocentric pose estimation. However, a stereo camera setup implies additional computation complexity and extra energy consumption, which is critical for low-power head-mounted devices, which are the main target applications.

The single head-mounted fisheye camera setup was first proposed by Xu *et al*. [39], who also introduced a two-stream CNN to cope with the low resolution of regions far from the fisheye camera, *i.e*., one branch for the full body and one branch for predicting the lower body joints from a zoom-in image. Tome *et al*. [32] proposed an encoder-decoder architecture to model the high uncertainty caused by severe occlusions present in this setup and Wang *et al*. [36] leverages motion capture data to learn a human motion prior, which is applied in an optimization method to obtain temporally stable poses for training and egocentric pose predictor. Another challenge is the strong image distortion caused by the fisheye lens, which can be mitigated with automatic camera calibration [46]. Other setups consider the camera facing forward and try to synthesize plausible human motion given only scene image evidences [18] or partially visible body parts [13, 15]. In our work, we do not consider this setup since in many poses only a few body extremities are visible in the image.

Obtaining real data in the egocentric setup is a time consuming process, therefore, many approaches rely on synthetic data for training [15,32]. A recent method has shown that the existing gap between synthetic and real data can be mitigated by domain adaptation techniques [35], but this approach still requires real data for the weak supervision part. Differently, our method has a simpler training strategy, and we reduce the gap between synthetic and real data by providing additional pseudo ground-truth scene labels for real sequences, including indoor and outdoor scenes.

2.2. Voxel Representation for Body Pose Estimation

Volumetric voxel representations have been extensively used with multiple view setups for the estimation of single [31, 33] and multiple human poses [34, 37, 40, 45], and for hand pose estimation [12, 19–21] from depth maps input. Considering a single image as input, Pavlakos *et al*. [25] proposed a coarse-to-fine approach to lift from 2D

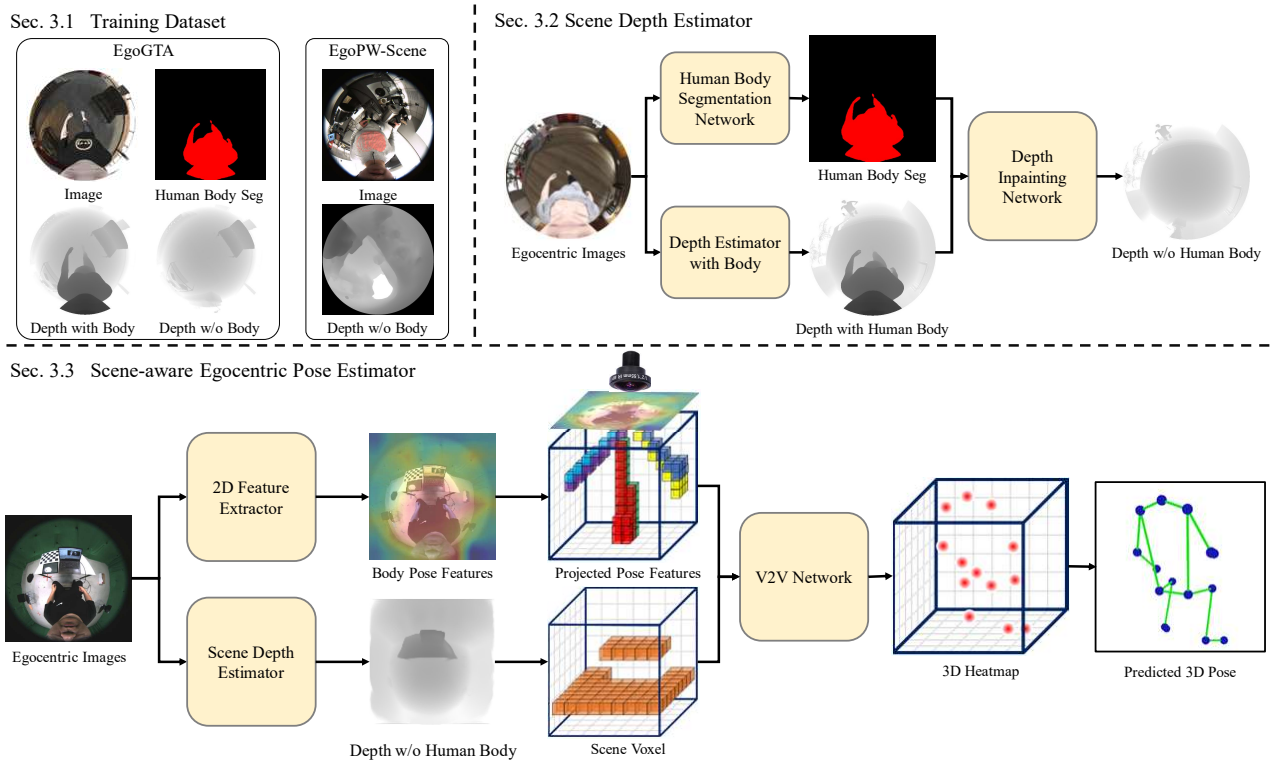

Figure 2. Overview of our method. We first render synthetic training dataset EgoGTA and in-the-wild training dataset EgoPW-Scene. Both datasets contain egocentric depth maps for subsequent training process (Sec. 3.1). Next, we train an egocentric scene depth estimator that predicts a depth map without the human body and a depth inpainting network (Sec. 3.2). Finally, we combine the 2D body pose features and scene depth map into a common voxel space. The 3D body pose heatmaps are regressed from the voxel space with a V2V network and the final pose prediction is obtained with soft-argmax (Sec. 3.3).

heatmaps to a voxel representation for 3D human pose estimation. Iskakov [14] later proposed a learnable triangulation method in voxel space that can generalize to single or multiple views, achieving less than 2 cm error in a controlled multi-view data. Despite the success of voxel representation for human and hand pose estimation from an external camera, this technique has not yet been explored for the interaction between human body and scene. In our work, we show the advantages of this representation for egocentric human pose estimation, especially when considering human-scene interaction.

2.3. Scene-aware Human Pose Estimation

In recent years, several approaches have been proposed to predict the pose of humans considering environmental and physical constraints from RGB [24, 30, 41] and inertial measurement units (IMU) [9, 42]. Some methods assume a simplified environment, such as a planar ground floor, to enforce a temporal sequence that is physically consistent with the universal law of gravity by assuming known camera poses [29, 30] or by tracking an object in the scene following a free flight trajectory [6]. Other approaches assume that the scene is given as input, either as a 3D reconstruction [9, 28] or as geometric primitives [43], whose positions can be refined in the optimization process. Bhatnagar *et al*. [2] proposed a method and dataset for human-object interactions. Taking into account the interaction between humans and furniture, holistic methods are able to estimate the position of humans and specific objects in the scene under the assumption of a planar floor [5, 38, 41], or even to estimate deformations in known objects based on human poses [17]. Contrary to the previous work, we make no strong assumptions about the objects and ground floor in the scene, but instead propose a method that learns to estimate the background scene geometry from a fisheye camera and explores the correlation between the human body and scene directly from egocentric data.

3. Method

We propose a new method for predicting accurate egocentric body pose by leveraging the estimated scene geometry. An overview of our method is shown in Fig. 2. In order to train the scene-aware network, we first generate a synthetic dataset based on the GTA-IM dataset [3], called

EgoGTA, and an in-the-wild dataset based on the EgoPW dataset [35], called EgoPW-Scene (Sec. 3.1). Next, we train a depth estimator to estimate the geometry of the surrounding scene and introduce the depth-inpainting network that estimates the depth behind the human body (Sec. 3.2). Finally, we combine 2D features and scene geometry in a common voxel space and predict the egocentric pose with a V2V network [22] (Sec. 3.3).

3.1. Training Dataset

Although many training datasets for egocentric pose estimation [32,35,39] have been proposed, they cannot yet train the scene-aware egocentric pose estimation network due to the lack of scene geometry information. To solve this, we introduce the EgoGTA dataset and EgoPW-Scene dataset (both will be made publicly available). Both datasets contain pose labels and depth maps of the scene for each egocentric frame, facilitating our training process. We show examples from both datasets, as illustrated in Fig. 2.

3.1.1 EgoGTA Dataset

In order to obtain precise ground truth human pose and scene geometry for training, we devise a new synthetic egocentric dataset based on GTA-IM [3], which contains various daily motions and ground truth scene geometry. For this, we first fit the SMPL-X model on the 3D joint trajectories from GTA-IM. Next, we attach a virtual fisheye camera to the forehead of the SMPL-X model and render the images, semantic labels, and depth map of the scene with and without the human body. In total, we obtained 320 K frames in 101 different sequences, each with a different human body texture. Here, we denote the EgoGTA dataset $\mathbb{S}_G = \{I_G, S_G, D_G^B, D_G^S, P_G\}$, including synthetic images I_G and their corresponding human body segmentation maps S_G, depth map with human body D_G^B, depth map of the scene without human body D_G^S, and egocentric pose labels P_G.

3.1.2 EgoPW-Scene Dataset

Since we want to generalize to data captured with a real head-mounted camera, we also extended the EgoPW [35] training dataset. For this, we first reconstruct the scene geometry from the egocentric image sequences of the EgoPW training dataset with a Structure-from-Motion (SfM) algorithm [10]. This step provides a dense reconstruction of the background scene. The global scale of the reconstruction is recovered from known objects present in the sequences, such as laptops and chairs. We further render the depth maps of the scene in the egocentric perspective based on the reconstructed geometry. Our EgoPW-Scene dataset contains 92 K frames in total, which are distributed in 30 sequences performed by 5 actors. The number of frames in the EgoPW-Scene dataset is less than EgoPW dataset since SfM fails on some sequences. Here, we denote the EgoPW-Scene dataset $\mathbb{S}_E = \{I_E, D_E^S, P_E\}$, including in-the-wild images I_E and their corresponding depth map of the scene without human body D_E^S, and egocentric pose labels P_E.

3.2. Scene Depth Estimator

In this section, we propose a depth estimation method to capture the scene geometry information in the egocentric perspective. Available depth estimation methods [7, 11, 16] can only generate depth maps with the human body, but are not able to infer the depth information behind the human, *i.e.*, the background scene depth. However, the area occluded by the human body, *e.g.* the areas of foot contact, are crucial for generating plausible poses, as demonstrated in Sec. 4.4. To predict the depth map of the scene behind the human body, we adopt a two-step approach. More specifically, we first estimate the depth map including the human body and the semantic segmentation of the human with two separated models. Then, we use a depth inpainting network to recover the depth behind the human body. This two-step strategy is necessary because the human visual evidences in the RGB images are too strong to be ignored by the depth estimator, therefore, it is easier to train the scene depth estimation as separated tasks.

We first train the depth estimator network $\mathcal{D}$, which takes as input a single egocentric image I and predicts the depth map with human body $\hat{D}^B$. The network architecture of $\mathcal{D}$ is the same as Hu *et al.* [11]'s work. To minimize the influence of the domain gap between synthetic and real data, the network is initially trained on the NYU-Depth V2 dataset [23] following [11], and further fine-tuned on the EgoGTA dataset.

Next, we train the segmentation network $\mathcal{S}$ for segmenting the human body. The network takes the egocentric image I as input and predicts the segmentation mask for the human body $\hat{S}$ as output. Following Yuan *et al.* [44], we use HRNet as our segmentation network. Similarly, to reduce the domain gap, we pretrain the network on the LIP dataset [8] and finetune the model on the EgoGTA dataset. We do not train network $\mathcal{D}$ and $\mathcal{S}$ on the EgoPW-Scene dataset since it lacks the ground truth segmentation maps and depth maps with the human body.

Finally, we propose a depth inpainting network $\mathcal{G}$ for generating the final depth map of the scene without human body. We first generate the masked depth map $\hat{D}^M = (1 - \hat{S}) \odot \hat{D}^B$, which is a Hadamard product between the background segmentation and the depth map with human body. Then, the masked depth map $\hat{D}^M$ and the segmentation mask $\hat{S}$ are fed into the inpainting network $\mathcal{G}$, which predicts the final depth map $\hat{D}^S$. We train the inpainting network $\mathcal{G}$ and finetune the depth estimation network $\mathcal{D}$ on both the EgoGTA and EgoPW-Scene datasets. During

training, we penalize the differences between the predicted depth maps and the ground truth depth of the background scene with L^S and also keep the depth map consistent in the non-human body regions with L^C. Specifically, the loss function is defined as follows:

$$\begin{aligned} L &= \lambda^S L^S + \lambda^C L^C, \quad \text{with} \\ L^S &= \left\| \hat{D}^S_G - D^S_G \right\|_2^2 + \left\| \hat{D}^S_E - D^S_E \right\|_2^2, \quad \text{and} \\ L^C &= \left\| (\hat{D}^S_G - \hat{D}^B_G)(1 - \hat{S}_G) \right\|_2^2 \\ &\quad + \left\| (\hat{D}^S_E - \hat{D}^B_E)(1 - \hat{S}_E) \right\|_2^2, \end{aligned} \tag{1}$$

where

$$\begin{aligned} \hat{D}^S_G &= \mathcal{G}(\hat{D}^M_G, \hat{S}_G); & \hat{D}^S_E &= \mathcal{G}(\hat{D}^M_E, \hat{S}_E); \\ \hat{D}^B_G &= \mathcal{D}(I_G); & \hat{D}^B_E &= \mathcal{D}(I_E); \\ \hat{S}_G &= \mathcal{S}(I_G); & \hat{S}_E &= \mathcal{S}(I_E), \end{aligned} \tag{2}$$

and λ^S and λ^C are the weights of the loss terms.

3.3. Scene-aware Egocentric Pose Estimator

In this section, we introduce our scene-aware egocentric pose estimator. We rely on the prior that human bodies are mostly in contact with the scene. However, estimating the contact explicitly from a single egocentric image is very challenging. Therefore, we rely on a data-driven approach by learning a model that predicts a plausible 3D pose given the estimated scene geometry and features extracted from the input image. To achieve this goal, we first leverage the EgoPW [35] body joints heatmap estimator to extract 2D body pose features F and use the scene depth estimator from Sec. 3.2 to estimate the depth map of the scene without human body $\hat{D}^S$. Afterwards, we project the body pose features and depth map into a 3D volumetric space considering the fisheye camera projection model. After obtaining the volumetric representation of human body features V_{body} and scene depth V_{scene}, the 3D body pose $\hat{P}$ is predicted from the volumetric representation with a V2V network [22].

Lifting the image features and depth maps to a 3D representation allows getting more plausible results, as inconsistent joint predictions can be behind the volumetric scene V_{scene} (pose-scene penetration) or spatially isolated from the voxelized scene geometry (pose floating), so they can be easily identified and adjusted by the volumetric convolutional network.

3.3.1 Scene and Body Encoding as a 3D Volume

In order to create the volumetric space, we first create a 3D bounding box around the person in the egocentric camera coordinate system of size $L \times L \times L$, where L denotes the length of the side of the bounding box in meters. The egocentric camera is placed at the center-top of the 3D bounding box so that the vertices of the bounding boxes are: $(\pm L/2, \pm L/2, 0)$ and $(\pm L/2, \pm L/2, L)$ under the egocentric camera coordinate system. Next, we discretize the bounding box by a volumetric cube $V \in R^{N,N,N,3}$. Each voxel $V_{xyz} \in R^3$ in position (x, y, z) is filled with the coordinates of its center under the egocentric camera coordinate system $(xL/N - L/2, yL/N - L/2, zL/N)$.

We project the 3D coordinates in V into the egocentric image space with the fisheye camera model [27]: $V_{\text{proj}} = \mathcal{P}(V)$, where $V_{\text{proj}} \in R^{N,N,N,2}$ and $\mathcal{P}$ is the fisheye camera projection function. The volumetric representation V_{body} of the human body is obtained by filling a cube $V_{\text{body}} \in R^{N,N,N,K}$ by bilinear sampling from the feature maps F with K channels using 2D coordinates in V_{proj}:

$$V_{\text{body}} = F\{V_{\text{proj}}\} \tag{3}$$

where $\{\cdot\}$ denotes bilinear sampling.

Then, we project the depth map to the 3D volumetric space. We first generate the point cloud of the scene geometry C from the depth map $\hat{D}^S$ with the fisheye camera projection function $C = \mathcal{P}^{-1}(\hat{D}^S)$. The volumetric representation of scene depth map V_{scene} is obtained by filling a binary cube $V_{\text{scene}} \in R^{N,N,N}$ by setting the voxel at (x, y, z) to 1 if there exists one point (x_p, y_p, z_p) in the point cloud C such that:

$$\left\| (\frac{xL}{N} - \frac{L}{2}, \frac{yL}{N} - \frac{L}{2}, \frac{zL}{N}) - (x_p, y_p, z_p) \right\| < \epsilon \tag{4}$$

where ϵ is the threshold distance. In our experiment, we set $L = 2.4$ m, $N = 64$, and $\epsilon = 0.04$ m. This setting can cover most types of the human motions and allows high accuracy of the predicted body pose.

3.3.2 Predicting 3D Body Pose with V2V Network

We feed the volumetric representation aggregated from V_{body} and V_{scene} into the volumetric convolutional network $\mathcal{N}$, which has a similar architecture as [22]. The V2V network produces the 3D heatmaps of the body joints:

$$V_{\text{heatmap}} = \mathcal{N}(V_{\text{body}}, V_{\text{scene}}) \tag{5}$$

Following [14], we compute the soft-argmax of V_{heatmap} across the spatial axes to infer the body pose $\hat{P}$. The predicted pose $\hat{P}$ is finally compared with the ground truth pose P_G from the EgoGTA dataset and P_E from the EgoPW-Scene dataset with the MSE loss.

4. Experiments

In this section, we evaluate our method considering existing and new datasets for egocentric monocular 3D human pose estimation. Please refer to the supplementary materials for the implementation details.

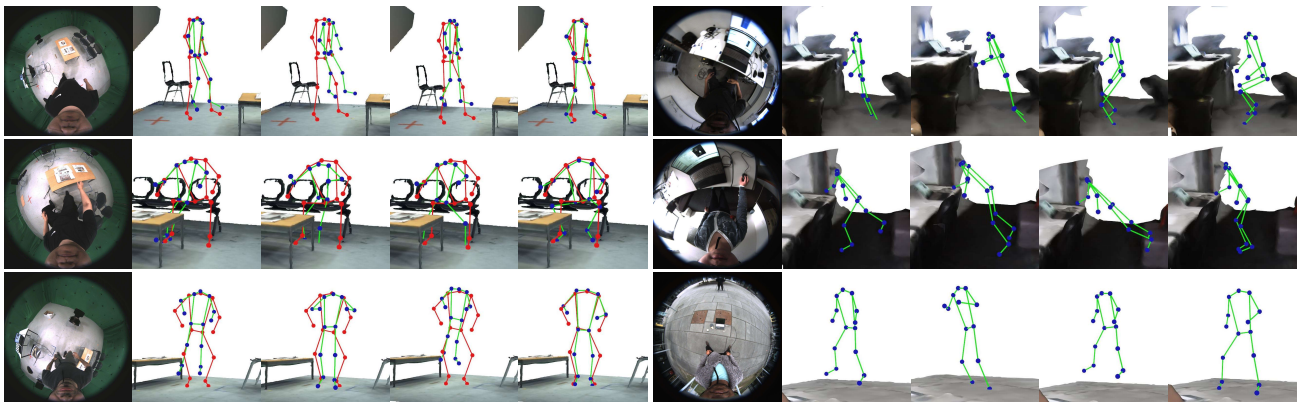

Figure 3. Qualitative comparison between our method and the state-of-the-art egocentric pose estimation methods. From left to right: input image, Mo^{2}Cap^2 result, xR-egopose result, EgoPW result and our result. The ground truth pose is shown in red. The input images from the left part are from our test dataset, while those in the right part come from the EgoPW [35] in-the-wild test sequences (without ground-truth poses). We also show the gt scene geometry of the in-the-studio data and scene geometry obtained by SFM method for the in-the-wild data. For better visualizing the interaction between human body and environment, please refer to our supplementary video.

4.1. Evaluation Datasets

Evaluating human-scene interaction requires precise annotations for camera pose and scene geometry. However, such information is not available in existing datasets for egocentric human pose estimation. To solve this issue, we collected a new real-world dataset using a head-mounted fisheye camera combined with a calibration board. The ground truth scene geometry is obtained with SfM method [10] from a multi-view capture system with 120 synced 4K resolution cameras and the ground truth egocentric camera pose is obtained by localizing a calibration board rigidly attached to the egocentric camera. This dataset contains around 28K frames of two actors, performing various human-scene interacting motions such as sitting, reading newspaper, and using a computer. This dataset is evenly split into training and testing splits. We finetuned the method on the training split before the evaluation. This dataset will be made publicly available and additional details of it are shown in the supplementary materials.

Besides our new test dataset, we also evaluate our methods in the test datasets from Wang *et al*. [36] and Mo^{2}Cap^2 [39]. The real-world dataset in Mo^{2}Cap^2 [39] contains 2.7K frames of two people captured in indoor and outdoor scenes, and the dataset in Wang *et al*. [36] contains 12K frames of two people captured in the studio.

4.2. Evaluation Metrics

We measure the accuracy of the estimated body pose with the MPJPE and PA-MPJPE. For the test dataset in Wang *et al*. [36] and Mo^{2}Cap^2 [39], we evaluate PA-MPJPE and BA-MPJPE [39] since the ground truth poses in the egocentric camera space are not provided. Further details of the metrics are shown in the supplementary materials.

4.3. Comparisons on 3D Pose Estimation

Method	MPJPE	PA-MPJPE
Our test dataset		
Mo^{2}Cap^2 [39]	200.3	121.2
xR-egopose [32]	241.3	133.9
EgoPW [35]	189.6	105.3
Ours	**118.5**	**92.75**
Method	PA-MPJPE	BA-MPJPE
Wang *et al*.'s dataset [36]		
Mo^{2}Cap^2 [39]	102.3	74.46
xR-egopose [32]	112.0	87.20
EgoPW [35]	81.71	64.87
Ours	**76.50**	**61.92**
Mo^{2}Cap^2 test dataset [39]		
Mo^{2}Cap^2 [39]	91.16	70.75
xR-egopose [32]	86.85	66.54
EgoPW [35]	83.17	64.33
Ours	**79.65**	**62.82**

Table 1. Performance of our method on our test dataset, Wang *et al*.'s test dataset [36] and Mo^{2}Cap^2 test dataset [39]. Our method outperforms the state-of-the-art methods EgoPW, Mo^{2}Cap^2 [39] and xR-egopose [32].

In this section, we compare our approach with previous single-frame-based methods, including EgoPW [35], xR-egopose [32] and Mo^{2}Cap^2 [39] on our test dataset under the "Our test data" in Table 1. Since the code for xR-egopose is not released, we use our implementation for the evaluation. In our dataset, the proposed method outperforms the previous state-of-the-art methods, EgoPW [35] by 37.5% on MPJPE and 11.9% on PA-MPJPE. We also compared with previous methods on the Wang *et al*.'s test dataset [36] and Mo^{2}Cap^2 test dataset [39] and show the re-

Method	Non pene.	Contact
Mo^{2}Cap^2 [39]	69.6%	23.1%
xR-egopose [32]	64.5%	38.3%
EgoPW [35]	71.7%	38.8%
Ours	**84.1%**	**89.4%**

Table 2. Comparisons of physical plausibility on our test dataset.

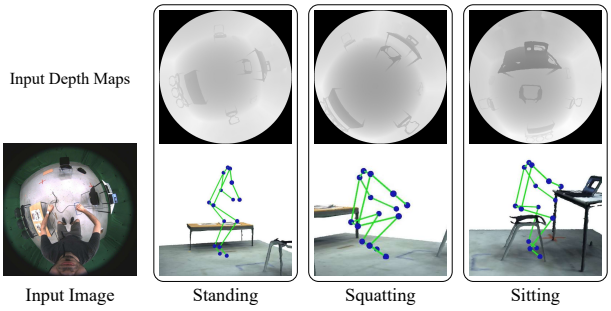

Figure 4. Predicted pose with different scene depth map input. Our network can generate different poses under different depth input and further disambiguate body poses under scene constraint.

sults in Table 1. On Wang *et al.*'s test dataset, our method performs better than EgoPW by 6.4%. On the Mo^{2}Cap^2 test dataset, our method performs better than EgoPW by 7.8%.

We also evaluate the physical plausibility of our predictions by calculating the percentage of predicted poses that are in contact with the scene and do not penetrate the scene in Table 2. We define a body pose as in contact with the scene if any body joint is less than 5cm from the scene mesh. A body pose suffers from the body floating issue if it is not in contact with the scene. Compared with previous approaches, our method generates body poses more physically plausible considering the constraints of the scene.

From the results in Table 1 and Table 2, we can see that our approach outperforms all previous methods on the single-frame egocentric pose estimation task. For the qualitative comparison, we show the results of our method on the studio dataset and in-the-wild sequences in Fig. 3. See our supplementary video for more qualitative evaluations. Our predicted poses are physically plausible under the scene constraint, whereas other methods generate poses suffering from body floating and penetration issues.

In order to further demonstrate our method performing predictions accordingly to the constraint of the scene geometry, we fix the input image and change the scene depth input to the depth map corresponding to the standing pose, squatting pose, and sitting pose. The results are presented in Fig. 4 and show that the predicted poses change to standing, squatting, and sitting to better adapt to the input changes of the scene geometry. This shows our method's ability to disambiguate poses under different scene constraints.

Method	MPJPE	PA-MPJPE
EgoPW+Optimizer	187.1	103.2
EgoPW+Depth	149.6	98.15
xR-egopose+Depth	180.5	103.7
Ours w/o Depth	188.1	105.1
Ours+Depth with Body	167.3	103.3
Ours+Depth w/o Body	135.7	95.84
Ours+Depth w/o Inpainting	124.2	95.00
Ours+GT Depth	**109.9**	**88.80**
Ours	*118.5*	*92.75*

Table 3. Results from our method compared to different baselines.

4.4. Ablation Study

Simple Combination of 2D Features and Depth Maps. In Sec. 3.3, we claim that the volumetric representation of egocentric 2D features and scene depth map is important for understanding the interaction between the human body and the surrounding scene. To provide evidence of this claim, in this experiment, we compare our method with baseline methods that simply combine the 2D image features and scene depth map. We set two baseline methods since there are two types of egocentric pose estimation methods, *i.e.* direct regression of 3D poses (xR-egopose [32]) and prediction of 2D position and depth for each joint (Mo^{2}Cap^2 [39] and EgoPW [35]). In the baseline method "xR-egopose + Depth", we concatenate 2D heatmaps and scene depth map as the input to the 3D pose regression network in xR-egopose. In the baseline "EgoPW + Depth", we concatenate the 2D features and the scene depth map and input them into the joint depth prediction network.

From the evaluation results shown in Table 3, both of the baseline methods perform worse than our proposed method. In "xR-egopose + Depth", simply combining the scene depth and 2D heatmaps cannot provide direct geometric supervision for the 3D pose regression network. In "EgoPW + Depth", though the joint depth estimation network performs better with the help of scene depth information, the 2D pose estimation network does not benefit from it. Both of the experiments demonstrate the effectiveness of our volumetric representation of 2D features and scene geometry, which provides direct geometry supervision for the physical plausibility of predicted egocentric poses.

Optimization. In this experiment, we compare our method with an optimization baseline that refines a 3D pose considering the scene constraint. Similar to Wang *et al.* [36] and EgoPW [35], we first train a VAE consisting of a CNN-based encoder f_{enc} and decoder f_{dec} to learn an egocentric motion prior. Then, We optimize the egocentric pose P by finding a latent vector z such that the corresponding pose $P = f_{\text{dec}}(z)$ minimizes the objective function

$E(P) = \lambda_R E_R + \lambda_J E_J + \lambda_C E_C$, where E_R is the egocentric reprojection term, E_J is the egocentric pose regularization term, and E_C is the contact term. The latent vector z is initialized with the estimated pose from the EgoPW method. The E_R and E_J are the same as those defined in Wang *et al*. [36]. Denote the nth joint in egocentric pose P as $P_n, n \in [0, N]$, where N is the number of joints, and the mth point in scene point cloud C as $C_m, m \in [0, M]$, where M is the number of points in the point cloud. The contact term E_C is defined as:

$$\begin{aligned} E_C &= \sum_{n \in [0,N]} d_n^2, \quad \text{if } d_n \leq \epsilon, \text{ otherwise } 0, \text{ and} \\ d_n &= \min_{m \in [0,M]} \|P_n - C_m\|_2. \end{aligned} \tag{6}$$

We first calculate the nearest distance d_n between each body joint and the projected point cloud C from the scene depth map. If the distance d_n of the nth joint is less than a margin ϵ, it is defined as in contact with the scene and minimized with the optimization framework.

The result of the optimization method is shown as "EgoPW+Optimizer" in Table 3, which demonstrates that the optimization framework is less effective than our method. This is because the accuracy of optimization method relies heavily on the initial pose. If the initial pose is not accurate, it will be difficult to determine the contact labels for each joint with the fixed distance margin. Without accurate contact labels, the optimization framework might force the joint that does not contact the scene to keep in contact, eventually resulting in wrong poses.

Scene Depth Estimator. In our work, we estimate the depth of the surrounding scene and infer the depth behind the human body with a depth inpainting network. In order to validate the effectiveness of the scene depth map, we remove the input depth map from the V2V network and show the results of "Ours w/o Depth" in Table 3. This baseline increases the MPJPE by about 70 mm, which is evidence of the relevant extra information provided by the scene depth.

To demonstrate the benefits of recovering the scene depth behind the human body, we evaluate our model using the estimated depth including the human body as the input to the V2V network. We also removed the human body area from the depth maps and use them as input to the V2V network. The results are shown in "Ours+Depth with Body" and "Ours+Depth w/o Body" in Table 3. Both of the baseline methods perform worse than our method because the area in the scene occluded by the human body can provide clues for generating plausible poses.

We also evaluate our method with ground truth depth maps in "Ours+GT Depth" in Table 3, which further improves over our estimated depth by 7.2% in terms of MPJPE and 4.2% in terms of PA-MPJPE. This demonstrates that the accuracy of our predicted pose benefits from better depth maps, but still our estimated scene depth already provides a significant improvement over the baselines.

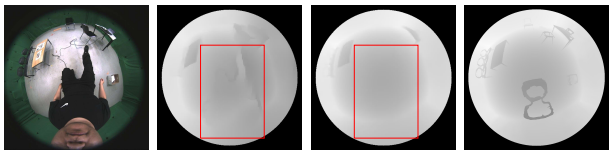

Figure 5. The qualitative depth estimation results with or without inpainting network. The depth map estimated without inpainting network show artifacts in the human body region (see the red box).

Finally, we compare with a baseline method without depth inpainting, *i.e.*, the human body is removed from the input image and the scene depth map is predicted directly with the network $\mathcal{D}$ in Sec. 3.2. The pose estimation accuracy is shown in "Ours+Depth w/o Inpainting" of Table 3 and the depth accuracy is shown in Table 4. Our method outperforms the baseline as it is more challenging to simultaneously estimate and inpaint the depth. Moreover, the network can be influenced by the segmented part in the input image as some extinct object, as shown in Fig 5.

Method	Abs-Rel	RMSE(m)
w/o Inpainting Network	0.3834	0.7856
Ours	0.1069	0.3365

Table 4. The quantitative results of depth estimator. We evaluate Abs-Rel and RMSE on our test dataset following [11].

5. Conclusion

In this paper, we have proposed a new approach to estimate egocentric human pose under the scene constraint. We firstly train a depth inpainting network for estimating the depth map of the scene without human body. Next, we combine the egocentric 2D features and scene depth map in a volumetric space and predict the egocentric pose with a V2V network. The experiments show that our method outperforms all of the baseline methods both qualitatively and quantitatively and our method can predict physically plausible poses in terms of human-scene interaction. In future, this method could be extended to estimate physically-plausible egocentric motion from a temporal sequence.

Limitations. The accuracy of voxel-based pose estimation network is constrained by the accuracy of estimated depth, especially where the scene is occluded by the human body. One solution is to leverage the temporal information to get a full view of the surrounding environment.

Acknowledgments Jian Wang, Diogo Luvizon, Lingjie Liu, and Christian Theobalt have been supported by the ERC Consolidator Grant 4DReply (770784).

References

[1] 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 *ECCV*, 2022. 2

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. *arXiv preprint arXiv:2204.06950*, 2022. 3

[3] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*. 2020. 2, 3, 4

[4] Y. Cha, T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, A. Ilie, A. State, Z. Xu, J. Frahm, and H. Fuchs. Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2993–3004, 2018. 2

[5] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8648–8657, 2019. 3

[6] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. In *The IEEE Conference on International Conference on Computer Vision (ICCV)*, pages 12365–12374, October 2021. 3

[7] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 4

[8] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 932–940, 2017. 4

[9] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 3

[10] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4, 6

[11] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019. 4, 8

[12] Fuyang Huang, Ailing Zeng, Minhao Liu, Jing Qin, and Qiang Xu. Structure-aware 3d hourglass network for hand pose estimation from single depth image. In *BMVC*, 2018. 2

[13] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera. In *ACM Symposium on User Interface Software & Technology*, page 98–111, 2020. 2

[14] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, 2019. 3, 5

[15] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. *ICCV*, 2021. 2

[16] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 4

[17] Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *International Conference on 3D Vision (3DV)*, 2022. 3

[18] 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. 2

[19] Jameel Malik, Ibrahim Abdelaziz, Ahmed Elhayek, Soshi Shimada, Sk Aziz Ali, Vladislav Golyanik, Christian Theobalt, and Didier Stricker. Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[20] Jameel Malik, Soshi Shimada, Ahmed Elhayek, Sk Aziz Ali, Christian Theobalt, Vladislav Golyanik, and Didier Stricker. Handvoxnet++: 3d hand shape and pose estimation using voxel-based neural networks. In *arXiv*, 2021. 2

[21] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[22] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *CVPR*, pages 5079–5088, 2018. 2, 4, 5

[23] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 4

[24] Georgios Pavlakos*, Ethan Weber*, , Matthew Tancik, and Angjoo Kanazawa. The one where they reconstructed 3d humans and environments in tv shows. In *ECCV*, 2022. 3

[25] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[26] 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 Trans. Graph.*, 35(6):162:1–162:11, 2016. 2

[27] Davide Scaramuzza and Katsushi Ikeuchi. Omnidirectional camera. 2014. 5

[28] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision*, pages 516–533. Springer, 2022. 3

[29] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics*, 40(4), 2021. 3

[30] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics*, 39(6), 2020. 3

[31] Masamichi Shimosaka, Yuichi Sagawa, Taketoshi Mori, and Tomomasa Sato. 3d voxel based online human pose estimation via robust and efficient hashing. In *2009 IEEE International Conference on Robotics and Automation*, pages 3577–3582, 2009. 2

[32] Denis Tomè, Patrick Peluse, Lourdes Agapito, and Hernán Badino. xr-egopose: Egocentric 3d human pose from an HMD camera. In *ICCV*, pages 7727–7737, 2019. 1, 2, 4, 6, 7

[33] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *European Conference on Computer Vision (ECCV'18)*, 2018. 2

[34] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[35] 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 *CVPR*, pages 13157–13166, June 2022. 1, 2, 4, 5, 6, 7

[36] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. *ICCV*, 2021. 1, 2, 6, 7, 8

[37] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct multi-view multi-person 3d human pose estimation. *Advances in Neural Information Processing Systems*, 2021. 2

[38] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343, 2021. 3

[39] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo^2cap^2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE Trans. Vis. Comput. Graph.*, 25(5):2093–2101, 2019. 1, 2, 4, 6, 7

[40] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[41] 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)*, pages 3959–3970, June 2022. 3

[42] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13167–13178, 2022. 3

[43] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *ACM Trans. Graph.*, 40(6), 2021. 3

[44] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. 2020. 4

[45] 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*, pages 1–1, 2022. 2

[46] 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. 2

[47] 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, 2021. 2