ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

VLG Computer Vision and Learning Group

3D Human Body Shape and Pose Estimation in Egocentric Data

Semester Project

Chuqiao Li
Department of Information Technology and Electrical Engineering

Advisors: Qianli Ma, Siwei Zhang, Dr.Yan Zhang
Supervisor: Prof. Dr. Siyu Tang, Prof. Dr. Marc Pollefeys

Aug 22, 2022

Abstract

Despite the thriving of 3D human pose and shape estimation(3DHPS) methods directly from single RGB images, it is shown that state-of-the-art methods suffer from severe performance decay when encountering egocentric data. In this work, based on EgoBody[64] dataset, we study the challenge of egocentric data from the perspectives of blurriness, truncated body, and weak camera calibration. We show that the main bottleneck is the bad inference of the truncated body parts. In order to tackle this problem, we propose a scene-based method that leverage the background segmentation mask corresponding to each input image. We developed a 2D scene branch that takes in the image corresponding segmentation and extracts scene features which are then absorbed into the final feature for the human body parameters prediction. We showed the comparative results and the detailed evaluation study analysis in this report.

Acknowledgements

I would like to express my special thanks of gratitude to Prof. Dr. Siyu Tang, Prof. Dr. Marc Pollefeys, Qianli Ma, Siwei Zhang, and Dr. Yan Zhang. I appreciate having the opportunity for doing an interesting project in VLG and the constant guidance, support, and motivation during the process. Thanks to them, I am able to acquire precious experience in doing research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Estimating 3D human pose and shape (3DHPS) directly from a single RGB image has attracted a lot of attention since it has many applications in computer graphics, sports motion analysis, Augmented and Virtual Reality (AR/VR), and beyond. Given an image or video sequence, 3DHPS methods[23, 24, 25, 18, 26, 5, 14, 21, 29] estimate the parameters of a human body models, such as SMPL[31] or SMPL-X[43] or directly the human meshes[29]. It remains a core challenging problem in the field of computer vision due to occlusion, truncation and complex articulated motion.

With the development of Deep learning architectures, recent works have achieved tremendous progress. The performance increases significantly on existing datasets such as H36M[16, 15], and MPI-INF-3DHP[35] which capture mostly indoor or in laboratory conditions. In terms of improving generalization, recent research[18, 26, 24, 22] focus on more realistic in-the-wild RGB images/videos where people always appear under occlusion or with various irregular motions. Facing these challenges, more works appear with the aim of solving body pose estimation with object occlusion, self-occlusion[24] or multi-person occlusion [22]. Other works attempt to introduce more information to the progress of estimation. SPEC[25] tried to estimate perspective camera parameters instead of applying a weak perspective or orthographic projection assumption during the estimation process. With the belief that monocular 3D scene estimation and monocular 3D human pose estimation should happen together, paper[14] proposed a dataset called PROX and showed introducing 3d scene constraints will lead to more accurate 3d reconstruction. With PROX dataset, more works[4, 33] include 3d scene information in the process of 3d pose estimation. However, all the above-mentioned 3DHPS methods were trained and evaluated on third-person-view datasets, and the study on egocentric view has received limited attention in the literature.

Being able to estimate the 3D body of the social interaction partner from egocentric views is crucial for understanding automated human interaction in the application scenario ranging from assistive robotics to AR/VR where the sensors mostly perceive the interaction partner in the first-person-view. Previous egocentric datasets suffer from a lack of annotation modalities, such as You2Me [39] which collects egocentric RGB images annotated 3d skeleton without body shape, or focused only on coarse interaction label[9, 36, 52] EgoBody[64] has filled in the gap and provides a large-scale egocentric dataset that captures 3d human motions under various social interaction scenarios. Interestingly, based on the benchmark provided in the paper[64], current state-of-the-art 3DHPS methods have severe performance decay when evaluated on the EgoBody dataset for unknown reasons. The possible factors causing the poor performance could be motion blur, severe body truncation, and people entering/exiting the field-of-view.

1.1 Focus of this Work

In this work, we are focusing on 3d human body shape and pose estimation in Egocentric data. In order to target the reason for 3DHPS methods' performance decline on egocentric view data, we have explored two state-of-the-art methods[24, 25] with the hope of tackling possible challenges which are linked to common characteristics of the EgoBody dataset. PARE[24] was first explored for solving the body occlusion due to scene content and truncation in the first-person-view under interaction scenarios. SPEC[25] was then explored with ground-truth camera information with the hope of achieving a better estimated 3d mesh scale. In order to locate the main error and distinguish the in-frame regression and out-of-frame inference, we proposed a new metric for 3d joints and reconstruction error evaluation.

After discovering the main reason causing the inaccurate which is the poor inference of truncated body parts, we target our goal to achieve more reasonable and accurate inference for the unseen body parts. In order to achieve this goal, we propose to introduce 2d scene information into the training process by segmenting the 3d scene mesh and rendering it to 2d masks corresponding to training images. Based on PARE network, we extend the network to encode the feature extracted from the scene masks. We compared the performance of the above-mentioned methods.

1.2 Thesis Organization

This thesis starts with the motivation of this work and then a brief related works description in Chapter 2. In Chapter3, we introduce the EgoBody dataset, its characteristics, and the potential challenges compared to other third-person-views datasets. we also present the theory of two 3DHPS methods that we exploited and our modification. We show the experiment results and analysis in Chapter 4. In the end, there is a short discussion about limitations and potential future work.

Chapter 2

Related Work

Deep Learning has powered 3D human mesh recovery[23, 24, 25, 18, 26, 5, 14, 21], facilitating the more challenges task of mesh recovery with severe body truncation and occlusion[24, 22, 18]. However, the previous methods were applied to third-person-view, 3d human pose and shape estimation in egocentric-view data is the main focus of our work.

2.1 Single-view 3D Human Shape and Pose Estimation

3D Human Shape and Pose Estimation is an active topic in recent years. Previous works have demonstrated impressive reconstruction with various sensors such as depth sensors[38, 48], monocular camera setting with the characteristic of convenience and efficiency is explored as well. But reconstructing 3d human mesh from a single image is challenging due to complex pose variations, occlusions, and limited 3d training data.

Pretrained parametric 3D human models[31, 41, 45, 43] are widely used by current works which estimate the pose and shape parameters and output the corresponding human mesh in 3D. Initial work explores 'bottom-up' regression methods and 'top-down' optimization methods[11, 50] using silhouettes or keypoints often with manual user intervention. These methods are fragile and do not generalize to in-the-wild images. SMPLify[3] was the first automated approach to fit the SMPL model to the output of a 2D off-the-shelf keypoint detector. Paper[28] uses silhouettes together with keypoints during fitting. on the other hand of the spectrum, Deep neural networks[12, 20, 40, 44, 53, 55] learn SMPL model coefficients directly from image pixels. Due to the lack of 3D ground truth for in-the-wild datasets, these approaches adopt weak supervision which is obtained from 2D keypoint re-projection loss[20, 53, 55], use intermediate 2D representations[40, 44] such as part segmentation, or leverage human in the loop[28]. Since it is challenging to regress human model coefficient directly from an input image, recent work further propose to leverage more information: temporal information[23, 61], various human body priors such as segmentation attention[24], accurate camera parameters[25, 63], and explore different optimization stragies[26, 20, 55].

Alternatively, instead of fitting a parametric human model, researchers also focused on non-parametric body reconstruction methods[27, 46, 56, 29] which directly regress 3D human body mesh from the input images. Various representation has been explored to represent human body, for instance, 3D mesh[27, 6], volumetric space[56], or occupancy field[46, 47]. Among these, Pose2Mesh[6] uses a cascaded model to reconstruct human mesh based on human pose representation. METRO[29] is a transformer-based method that models global interactions among joints and mesh vertices to reconstruct 3D human pose and mesh.

Pose estimation under egocentric view receives growing attention where usually camera viewer's 3D skeleton is estimated based on various inputs such as images, IMU data, scene cues, or body-object interactions[13,

17, 32, 49, 54, 62]. In this work, we are focusing on estimating the pose of the single person in egocentric view instead of the camera viewer's pose.

2.2 Datasets for 3D human pose, motion and interactions

Most benchmark datasets used for learning human pose and shape estimation focus on third-person-views [16, 51, 19, 34, 57, 59, 14, 58, 35, 42, 60, 65]. Among them, Human3.6M[16], HumanEva[51],TotalCapture[19], AMASS[34] use optical marker-based motion capture (mocap) systems to collect large-scale datasets with high quality. While providing accurate annotations, they are limited to constrained studio setup and also limited image complexity due to lack of background variation. PROX[14] captures people moving in 3D scenes from monocular RGB-D but without human interactions. Among egocentric datasets, lots of them focus on hand-object interactions and action recognition usually 3D ground-truth not provided[2, 7, 8, 10]. HPS[13] reconstructs the body pose and shape of camera wearer in 3D scenes. EgoMoCap[30] estimates the interact partner's shape and pose in outdoor scenarios. EgoBody[64] is the first motion capture dataset that collects calibrated egocentric and third-person-view images in various interaction scenarios with rich 3D ground-truth. Our work is based on this dataset.

Chapter 3

The Exploited Comparative Methods

3.1 EgoBody Dataset

Our work is based on the Egobody dataset[64], which collects 125 sequences from 36 subjects (18 male and 18 female) performing diverse social interactions in 15 indoor scenes. EgoBody provides both third-person-view data and egocentric data: “Multi-view (MV)Set” with in total of 219,731 synchronized frames captured from Azure Kinects, ”EgoSet” 199,111 RGB frames, and ”EgoSetinteractee” 175,611 frames with visible interactee in the egocentric view captured from HoloLens2. Image corresponding 3D human full-body pose and shape annotations and 3D scene mesh are provided. Examples are shown in Fig.3.1 Our work leverage ”EgoSet-interactee” dataset which is split into non-overlapping training, validation, and test sets with 90,124 EgoSet-interactee frames in training split, 23,332 EgoSet-interactee frames in test split, 62,155 Ego-interactee frames in test split.

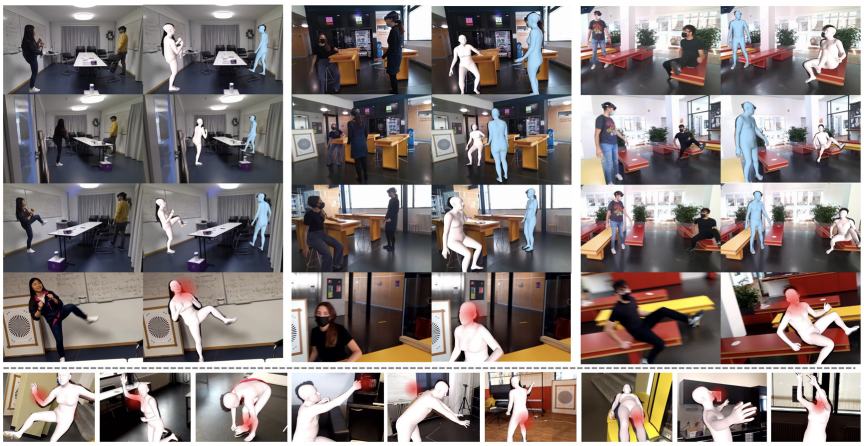
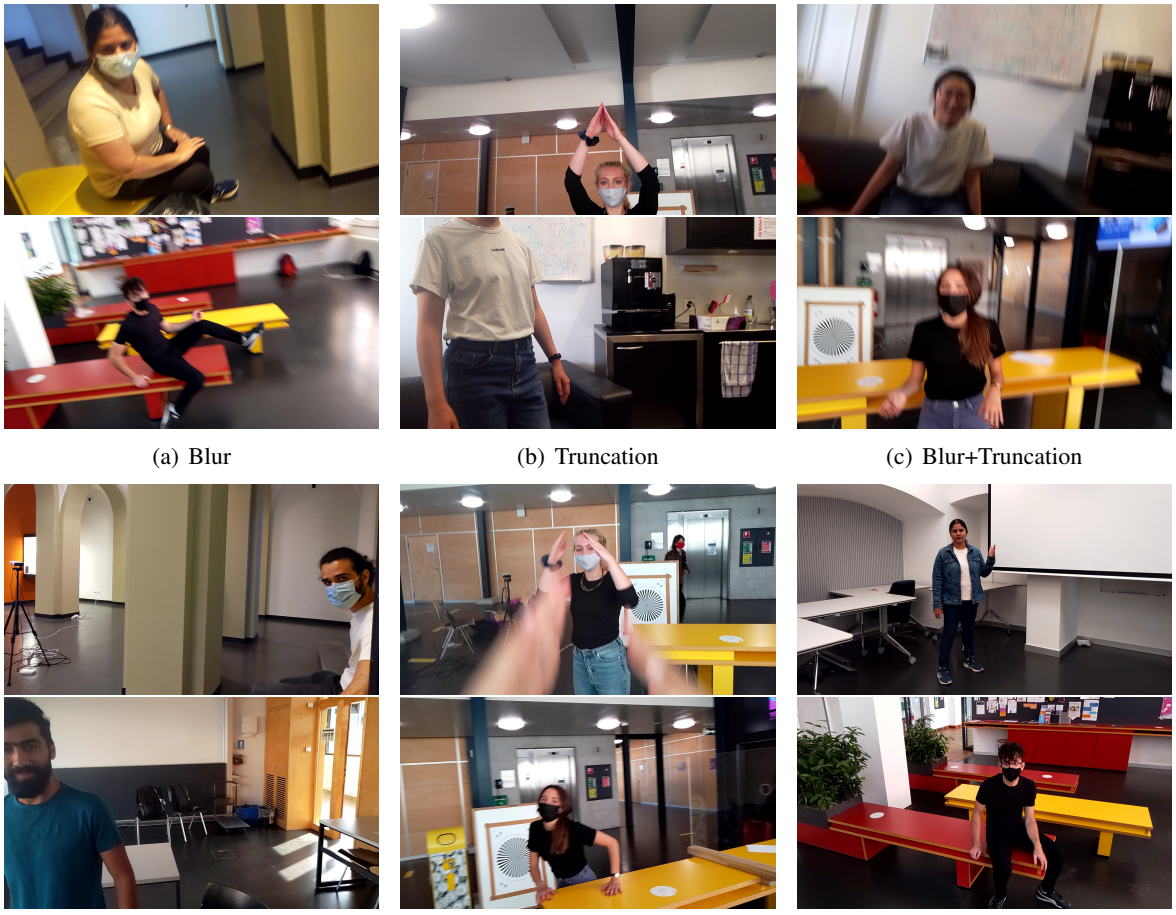


Figure 3.1: Reconstructed ground-truth bodies overlaid on third-person-view images from 3 Kinects (row 1-3), and the corresponding egocentric view image (row 4). Left/middle/right shows three different frames. Row 5 shows more examples from the egocentric view. Blue denotes the camera wearer, and pink denotes the interactee. Eye gaze of the camera wearer are in red circles.[64]

The Egobody dataset has multiple key challenges compared to third-person-view data, such as body truncation/occlusion, motion blur, people entering/exiting the field of view, and so on. Examples are shown

in Fig.3.2 Among them, blurriness and body truncation are the two most common and severe challenges. The blurriness is either due to the motion of the interactee or the head movement of the camera-viewer. While the truncation usually happens due to the close distance between people in various social interaction scenarios. And most of the time, these two phenomena appear together in the image frames which makes the data more challenging.



(a) Blur (b) Truncation (c) Blur+Truncation

(d) Enter/Exit FOV (e) Occlusion (f) Sharp Full Body

Figure 3.2: Example images of different challenge types from the EgoBody dataset

3.2 PARE: Part Attention Regressor for 3D Human Body Estimation

Since it has been shown that state-of-the-art 3DHPS methods which rely on global feature representations remain sensitive to partial occlusions, [24] propose a soft attention mechanism, called the Part Attention REgressor (PARE), that learns to predict body-part-guided attention masks. The attention mask was used for exploiting information about the visibility of individual body parts and, at the same time, leveraging information from neighboring body parts to predict occluded parts. It is shown in the paper that PARE could achieve better reconstruction results both on occlusion-specific and standard benchmarks. We are interested in PARE since the EgoBody dataset shares some common characteristics with occlusion-specific datasets and occlusion/truncation is one of the main challenges of the EgoBody dataset.

3.2.1 Methodology

In order to understand the visibility of body parts and if their location is occluded, PARE proposes a pixel-aligned structure, where each pixel corresponds to a region in the input image, and a feature volume is stored. Estimating attention mask and learning en-to-end features for 3D body estimation could be two separate tasks, thus PARE exploits the 2D part branch and 3D body branch to estimate attention weights and SMPL parameters individually. Finally, PARE leverages part segmentation as soft attention to balance the contribution of each feature differently in the 3D branch for different joints

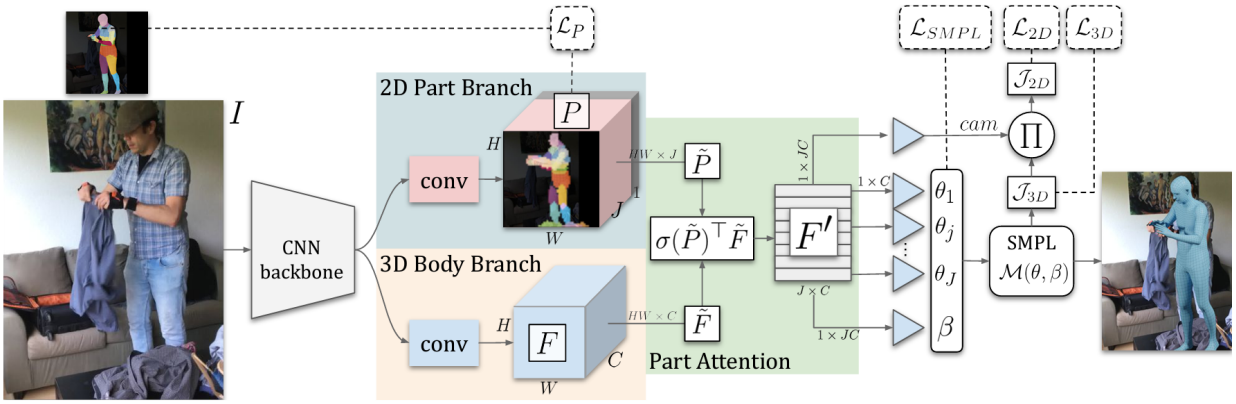


Figure 3.3: PARE model architecture. Given an input image, PARE extracts two pixel-level features P and F, which are fused by part attention (green box) leading to the final feature F for camera and SMPL body regression.[24]

The structure of PARE is shown in Fig.3.3: Given an input image I, the volumetric features was first extracted by a CNN backbone, and then there are two separated branches to extract two volumetric image features. The 2D part branch can be denoted as $P \in \mathbb{R}^{H \times W \times (J+1)}$, representing for J part attention masks, and 1 background mask, H and W denote the height and width of the feature. In jth mask, the pixel (h, w) stores the possibility of belonging to body part j. Meanwhile, the 3D branch for body parameter regression could be described as $F \in \mathbb{R}^{H \times W \times C}$. Let $F' \in \mathbb{R}^{J \times C}$ denotes the final feature, P_j and F_c denote the jth layer of P and cth layer of F, the element at (j, c) of F' is computed as:

$$F'_{j,c} = \sum_{h,w} \sigma(P_j) \odot F_c \tag{3.1}$$

where $\odot$ is the Hadamard product. This product can be implemented in a efficient way as: $F' = \sigma(\tilde{P})^\top \tilde{F}$, where $\tilde{P} \in \mathbb{R}^{HW \times J}$ is the reshaped P without the background layer, and $\tilde{F} \in \mathbb{R}^{HW \times C}$ is the reshaped F.The 2D branch is supervised with ground-truth segmentation labels, in order to help the mask of the visible part converge to the ground-truth segmentation. PARE exploits a hybrid training strategy that uses ground-truth segmentation labels to supervise the 2D part branch only in the initial stage, and then the branch is trained without supervision. F' is used for regressing SMPL body pose and shape parameters and also a weak-perspective camera model with scale and translation parameters.

The total loss is computed as:

$$L = \lambda_{3D} L_{3D} + \lambda_{2D} L_{2D} + \lambda_{SMPL} L_{SMPL} + \lambda_P L_P \tag{3.2}$$

with:

$$
\begin{aligned}
L_{3D} &= \|J_{3D} - \hat{J}_{3D}\|_F^2, \\
L_{2D} &= \|J_{2D} - \hat{J}_{2D}\|_F^2, \\
L_{SMPL} &= \|\theta - \hat{\theta}\|_2^2 + \|\beta - \hat{\beta}\|_2^2, \\
L_P &= \frac{1}{HW} \sum_{h,w} CrossEntropy(\sigma(P_{h,w}), \hat{P}_{h,w}),
\end{aligned}
\tag{3.3}
$$

where parameter with hat denotes the ground-truth. Notably, the prediction of 2D keypoints is calculated with SMPL 3D joint locations J_{3D} by projection: $J_{2D} \in \mathbb{R}^{J\times 2} = s\Pi(RJ_{3D}) + t$, where $S \in SO(3)$ is the camera rotation matrix and Π represents the orthogonal projection.

3.3 SPEC: Seeing People in the Wild with an Estimated Camera

As shown in Fig.3.4, current state-of-the-art 3DHPS methods always make simplifying assumptions on camera, referred to as IWP-cam: weak-perspective projection, large constant focal length, and zero camera rotation. The underlying assumption is that the camera is placed far from the person such that the depth variation of the person in the z coordinate is negligible compared to the distance to the camera. But this assumption tends to cause an error in the 3D reconstruction result under various scenarios when the camera has significant pitch and smaller focal lengths with foreshortening distortion. In order to tackle this problem, SPEC[25] was proposed as the first in-the-wild 3DHPS method that estimates the perspective camera parameters and integrates this information into the estimation process of the 3D human body and pose. And the paper[25] has shown quantitatively and qualitatively that their method is more accurate than prior methods on the standard benchmark (3DPW) as well as two new datasets with more challenging camera views and varying focal lengths. We are interested in this method since close camera distance is one characteristic under many social interaction scenarios the ground-truth camera information is available for the EgoBody dataset. We want to figure out if it would be helpful if we include accurate camera parameters in the estimation process.

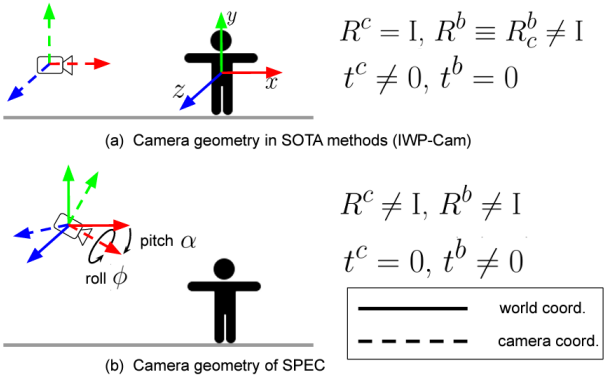


Figure 3.4: Illustration of IWP-cam and SPEC. R^c and t^c are camera rotation and translation. R^b and t^b are body orientation and translation. All are defined in world coordinate[25].

3.3.1 Methodology

In order to introduce more accurate camera parameters in the 3DHPS estimation, SPEC has two separate branches: CameraClib and SPEC backbone. The overall structure is shown in Fig.3.5. First, the CameraClib network is trained to take an input of a single RGB image and output the estimated camera parameters: the field of view, camera pitch, and roll. Then the rest of the SPEC network is trained to concatenate the camera calibration to the image features and take them as input for the 3D body shape and pose estimation.

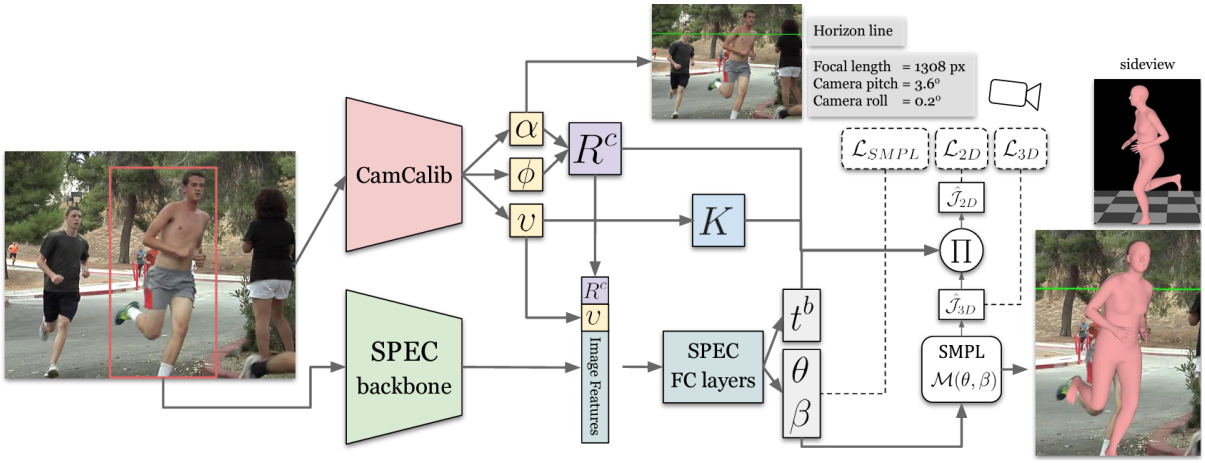


Figure 3.5: SPEC overview.[25]

By releasing the zero camera rotation constraints and estimating camera rotation R^c, SPEC bypass the camera-relative body orientationR_b^c and thus disentangles the camera rotations from body orientations. This enables handling of the above-mentioned foreshortening distortion and at the same time maintaining the xz-plane aligned ground plane at $[0, y, 0]$. The CameraClin network takes a single RGB image as input and outputs the estimated camera rotation R^c and focal length f. Since the camera matrix can be parameterized by pitchα, rollϕ, and yaw, while focal length f_y can be calculated with vertical field of view(vfov)ν. Following [66], they assume $f_x = f_y = f$ and zero camera yaw. They placed their camera at $t_c = [0, 0, 0]$ as depicted in Fig.3.4 and leave body translation t_b for the downstream body estimator. Thus the overall parameters for CameraCalib to estimate are pitchα, rollϕ, and vfovν. Notably, CameraCalib takes in the full-frame image instead of cropped image patch around the person since the author believes that full-frame images contain more details that can be helpful to the camera calibration.

After the estimation of camera pitchα, rollϕ and vfovν, rotation R^c and K are calculated and then incorporated into the estimation process as shown in Fig.3.5. For human body estimation, a cropped image is taken as input by a CNN backbone and outputs the image features. These features are then concatenated with R^c and ν and then fed into HMR to regress SMPL parameters, poseθ, shapeβ, and body translation t_b. The overall loss can be calculated as:

$$L = \lambda_{3D} L_{3D} + \lambda_{2D} L_{2D} + \lambda_{SMPL} L_{SMPL} \tag{3.4}$$

with:

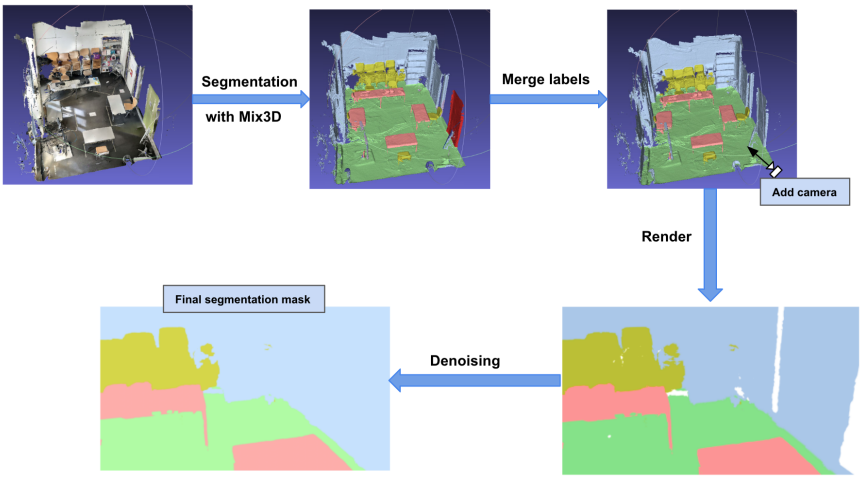
$$\begin{aligned} L_{3D} &= \|J_{3D} - \hat{J}_{3D}\|_F^2, \\ L_{2D} &= \|J_{2D} - \hat{J}_{2D}\|_F^2, \\ L_{SMPL} &= \|\theta - \hat{\theta}\|_2^2 + \|\beta - \hat{\beta}\|_2^2, \\ \hat{J}_{2D} &= K[R^c| - t^b]\hat{J}_{3D} \end{aligned} \tag{3.5}$$

3.4 Our proposed scene-based part attention method

After evaluation and further analysis of the implementation of the above-mentioned methods on Egocentric data, we have driven the conclusion that accurate camera calibration is not very helpful under the current setup and partial attention can facilitate the estimation of the in-frame body parts and can tackle the occlusion challenge of these visible parts. However, the unseen body parts are still poorly and unreasonably predicted, such as a straight standing body but with bending knees which make the person end up floating above the ground. The detailed results can be seen in Chapter 4. Therefore, we propose to encode the scene information into the estimation process as an additional guide for the unseen body parts inference. For example, when a person is sitting behind a table with the lower body parts truncated, if the model is aware of the background content, in this case, table and chairs, the model would tend to infer the person to be sitting. While if the person is standing behind a table, and the model is aware that there is no chair in the background, it would tend to infer the person is standing instead of sitting. The relative location information can be also very helpful in this case, for example, a standing person is much taller than a sitting person, etc.

3.4.1 Acquisition of image-corresponding scene information

Since often it is time-consuming and unfeasible to get corresponding image pairs with and without the person, ie, the corresponding background RGB images. But the 3D scan of the background can be acquired relatively easily. Since EgoBody makes the background scan available, we leverage the 3D meshes to generate the segmentation mask of the background corresponding to each RGB image. We first use Mix3D[37] to get the segmented 3D background meshes. The output labels use the format of ScanNet, whereas in our case, fine-grained labeling is not necessary. We merged the unnecessary labels, and keep 6 labels in the end, ie, wall, floor, table, chair, sofa, and other furniture. And then with the per-frame camera intrinsic and extrinsic information, we can render the segmented 3D meshes into images with lightning disabled. The rendered images serve as an initialization for the final segmentation mask. The flows of the segmented 3d meshes, such as the incomplete or additional noisy 3d parts cause the output segmentation to be noisy as well. After the rendering, the borderline will have a color value in between the label colors which might be due to different labeled parts corresponding to one image pixel. We first convert the output label images to segmentation masks leaving the empty part and the non-label color to be the label of the background, i.e.wall. Then we conduct some denoising techniques to get a cleaner mask. The overall process is described in Fig.3.6.



(a) Acquisition process

(b) Overlay of the original image and the final mask

Figure 3.6: The segmentation mask acquisition process.

3.4.2 Methodology

As described in Fig.3.8, based on PARE, we add an additional 2D scene branch into the original architecture. The 2D scene branch adopted SegNet[1] with the structure of encoder and decoder as its backbone. As shown in Fig.3.7, it is pretrained with the input of segmentation masks/RGB label images and output of segmentation masks, with the hope of extracting features with scene info in the middle layers.

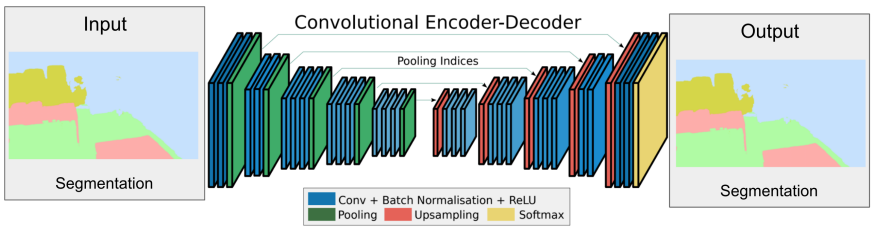


Figure 3.7: Structure of scene branch

The network takes the input of cropped RGB images and the corresponding cropped segmentation mask/RGB label image as input, and outputs the body shape and pose parameters, which is the same as PARE. The segmentation input is taken by the pretrained SegNet and the feature extracted by the middle layer is then concatenated with the output feature of the 3D body branch to form a new feature. Then the

new feature is incorporated with the part attention mask with the same procedure as PARE and then fed into the following networks to estimate body parameters.

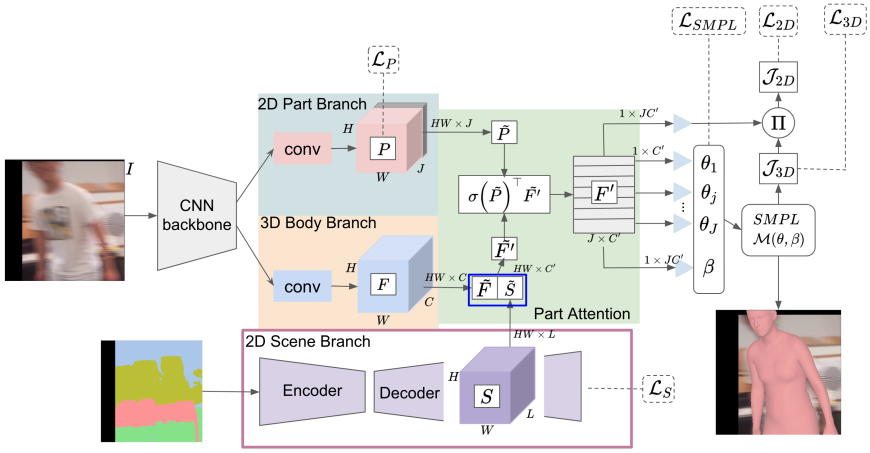


Figure 3.8: Structure of our proposed method.

The total loss is computed as:

$$L = \lambda_{3D}L_{3D} + \lambda_{2D}L_{2D} + \lambda_{SMPL}L_{SMPL} + \lambda_P L_P + \lambda_S L_S \tag{3.6}$$

with:

$$\begin{aligned} L_{3D} &= \|J_{3D} - \hat{J}_{3D}\|_F^2, \\ L_{2D} &= \|J_{2D} - \hat{J}_{2D}\|_F^2, \\ L_{SMPL} &= \|\theta - \hat{\theta}\|_2^2 + \|\beta - \hat{\beta}\|_2^2, \\ L_P &= \frac{1}{HW}\sum_{h,w} CrossEntropy(\sigma(P_{h,w}), \hat{P}_{h,w}), \\ L_S &= \frac{1}{H'W'}\sum_{h',w'} CrossEntropy(S'_{h',w'}, \hat{S}'_{h',w'}), \end{aligned} \tag{3.7}$$

With h', w' denote the height and width of the input cropped image, $S'_{h',w'}$is the prediction of the 2d scene branch and $\hat{S}'_{h',w'}$ is the ground-truth segmentation mask. L_S constraints the scene branch output to maintain the same information as the input. The other loss terms are the same as PARE. But since we are fine-tuning the pretrained PARE network, we omit L_P during the training process. However, if training from scratch, L_P will be needed in the early stage as described in PARE.

Chapter 4

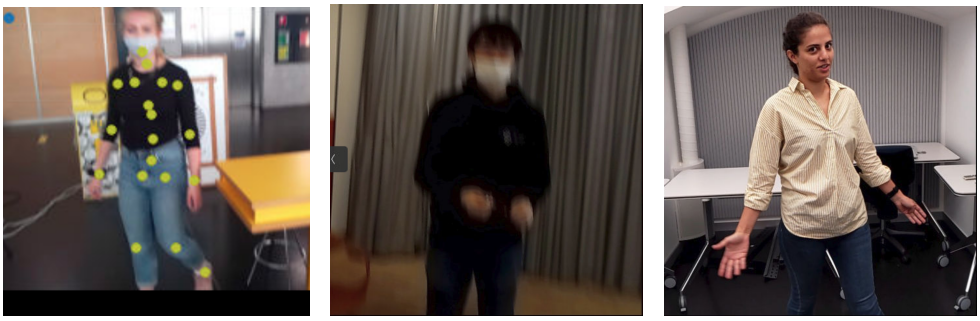
Experiments and Results

4.1 Egobody DataSet Analysis

According to the paper[64], existing state-of-the-art(SoTA) 3DHPS methods has a severe performance drop when tested on the test split of the dataset. Although there is no conclusion on what exactly caused the variation in performance, the author gives insights on the key challenging facts that might impact performance: motion blur and joint visibility. In order to find out what is the key bottle-neck of SoTA methods on the Egobody dataset, we conduct a detailed evaluation study on two SoTA methods, PARE[24] and SPEC[25].

4.1.1 Study on the effect of visible joints and blurriness

In order to better understand if the bottleneck lies in the joint visibility, we proposed a new metric which measures reconstruction error of only the in-frame joints, i.e. the joints inside bounding box, which is corresponding to 'select' in all tables. We acquire the selected joints/vertices by projecting the reconstructed 3D mesh/joints onto the image with the predicted camera translation and select those inside the bounding box. An example is depicted in Fig.4.1(a),only the green joints in frame are being calculated in the error. We also further study on the reconstruction of the lower body joints(knees,ankles,feet) and upper body joints(knees,ankles,feet), which correspond to 'lower' and 'upper' in Table4.1.



(a) Joints selected (b) Blur subset example (c) Sharp subset example

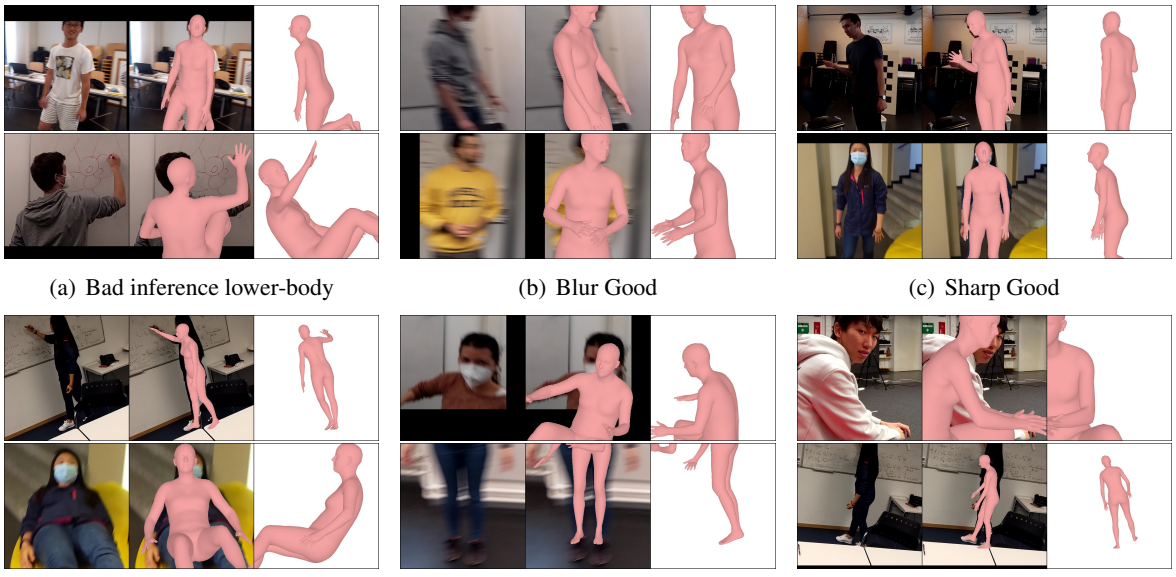
Figure 4.1: (a) shows the selected joints(green) which has to be inside the bounding box. (b) and (c) show the examples from blur subset and sharp subset selected with the sharpness score(Blur: score$<$20, Sharp: score$>$100).

In order to study the effect of blurriness on these two methods, we select a blur subset and a sharp subset from the EgoBody test set using the sharpness score introduced in [64]. The thresholds are set as score<20 for blur images, score>100 for sharp images, example images are shown in Fig.4.1(c) and Fig.4.1(b). There are 4300 images in the blur subset, meanwhile, 27889 images in the sharp subset. We have tested PARE on these different data sets and with different joint sets whose results are shown in the following table.

Dataset	MPJPE	MPJPE$_{select}$	MPJPE$_{lower}$	MPJPE$_{upper}$	V2V	V2V$_{select}$
EgoBody Testset	117.35	73.18	237.72	111.52	131.16	130.36
Blur Subset	128.72	86.32	273.02	114.05	140.31	139.79
Sharp Subset	115.96	72.05	234.79	111.79	130.16	129.32

Table 4.1: Evaluation of PARE on EgoBody full test set/blur subset/sharp subset. All metrics are in mm.

As we can see, the selected joint loss decrease significantly in all the cases. The errors (MPJPE, V2V) of the full test set are very close to those of the sharp subset. When it comes to the extreme blur case, all errors increase but the largest error among the listed errors is always $MPJPE_{lower}$, the lower body joints error, which most likely are the truncated parts. Therefore, we came to the conclusion that the main error is due to the bad inference of the truncated parts which are not inside the bounding box. Some visualization is shown in Fig.4.2, the bad inference of truncated body parts is mainly focused on the lower body joints. Even though the example is very blurry, the reconstruction can still be very good, and the bad results are mainly due to the very limited number of joints shown in the image. Notably, the bad scale is also another major problem.



(a) Bad inference lower-body (b) Blur Good (c) Sharp Good

(d) Bad body scale (e) Blur Bad (f) Sharp Bad

Figure 4.2: Qualitative results on PARE model. From left to right: input image, mesh overlaid on the image, sideview of the mesh.

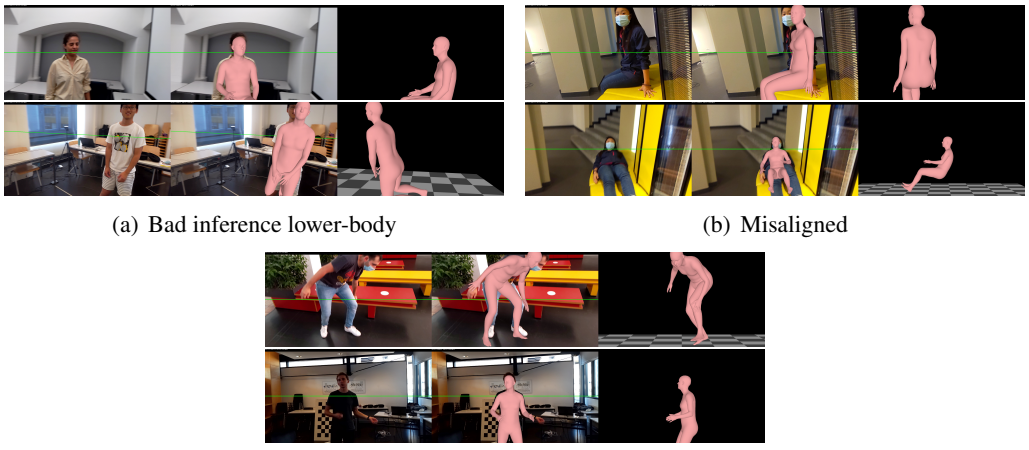
4.1.2 Study on the effect of accurate camera calibration

As we mentioned above, we want to tackle the bad scale of the reconstructed mesh. As introduced in Chapter 3, by leveraging accurate camera calibration on the body and pose estimation, SPEC[25] showed that this can facilitate more accurate reconstruction. We want to find out whether using the accurate camera parameters can help with EgoBody test data. We test SPEC on the EgoBody test split with different camera calibrations, the results are shown in Table 4.2. Among those, $f = 5000, R^c = I$ represents the weak-perspective camera calibration; $f = f_{gt}, R^c = I$ use the ground-truth focal length provided by Egobody with the camera rotation $R^c = I$; $f = f_{pred}, R^c = I$ use the focal length predicted by the pretrained CamCalib model of SPEC[25], and still camera rotation $R^c = I$; $f = f_{pred}, R^c = R^c_{pred}$ we use all the information provided by the CameraCalib model.

Camera Parameters	MPJPE	MPJPE$_{select}$	PA-MPJPE	PA-MPJPE$_{select}$	V2V	V2V$_{select}$	PA-V2V	PA-V2V$_{select}$
$f = 5000, R^c = I$	121.56	71.09	68.37	39.76	137.18	136.64	85.06	84.73
$f = f_{gt}, R^c = I$	121.07	71.08	67.77	38.79	136.43	135.86	84.81	84.44
$f = f_{pred}, R^c = I$	121.10	71.17	67.71	38.76	136.44	135.87	84.77	84.41
$f = f_{pred}, R^c = R^c_{pred}$	122.95	73.92	67.52	39.43	138.69	138.16	84.63	84.29

Table 4.2: Evaluation of SPEC[25] on EgoBody full test set with different camera parameters. All metrics are in mm.

As shown in the table above, there is no significant decrease in the reconstruction error when using a more accurate camera calibration when evaluated on the Egobody test set. And the reconstruction errors are bigger than those of PARE. We tend to draw the conclusion that for now, adding more accurate camera calibration information does not have a big influence on the current setup. We show some visualization of SPEC reconstruction in Fig.4.3.



(a) Bad inference lower-body

(b) Misaligned

(c) Good Reconstruction

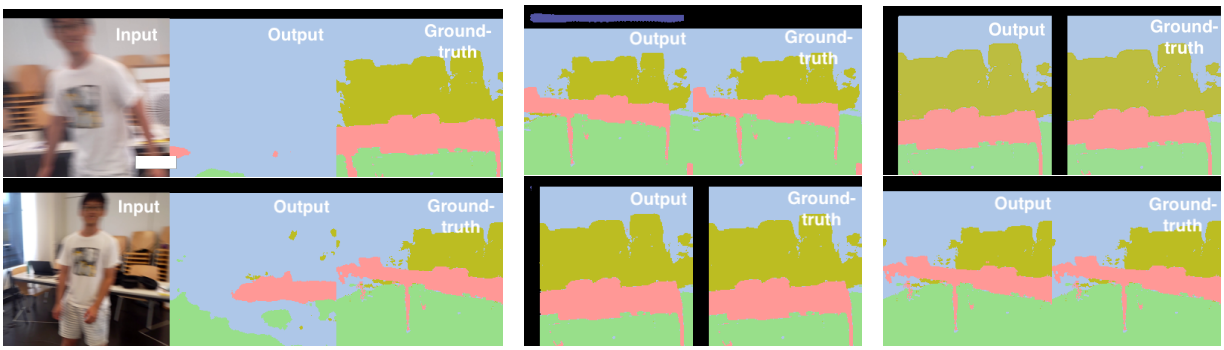
Figure 4.3: Qualitative results on SPEC model. From left to right: input image, mesh overlaid on the image, sideview of the mesh.

4.2 Train Proposed Scene-based Method on EgoBody Dataset

As described in Chapter 3, we propose a scene-based part attention method based on PARE structure. In order to make adjustments to PARE network and fit-tune the model on the Egobody dataset, we need to acquire a background scene segmentation mask corresponding to each image in the EgoBody dataset. The scene mask is generated by rendering the segmented 3D scene which is provided by Egobody Dataset. For a fair comparison, we also need to fine-tune the original PARE model on the EgoBody training split.

4.2.1 Pretraining the 2D scene branch

Before the training step, we propose to pretrain the 2D scene branch which uses the SegNet backbone with various inputs corresponding to the output of segmentation mask. We propose to train on three types of input: original cropped image, RGB cropped label image and cropped segmentation mask(same as output). Notably, the RGB cropped label image is acquired by converting the segmentation mask to an RGB image with each label corresponding to one specific color. The performance of the pretrianed models is shown in Fig.4.4. With the hope of enforcing the scene branch to infer the background even with human in the cropped image, we use the original image as the input. However, it seems too difficult for the current model to learn that since the output segmentation is quite different from the ground-truth segmentation mask. The other two cases work fine. Therefore, we decided to only use label image and mask as the input for the later training stage.



(a) original image input (b) label image input (c) mask input

Figure 4.4: (a)is the example input image, output mask, ground-truth mask triplets. (b) and (s) are the paired output mask and ground-truth mask, the visualization of input would be the same as ground-truth. (b) use RGB label image as input, (c) directly use ground-truth mask as input.

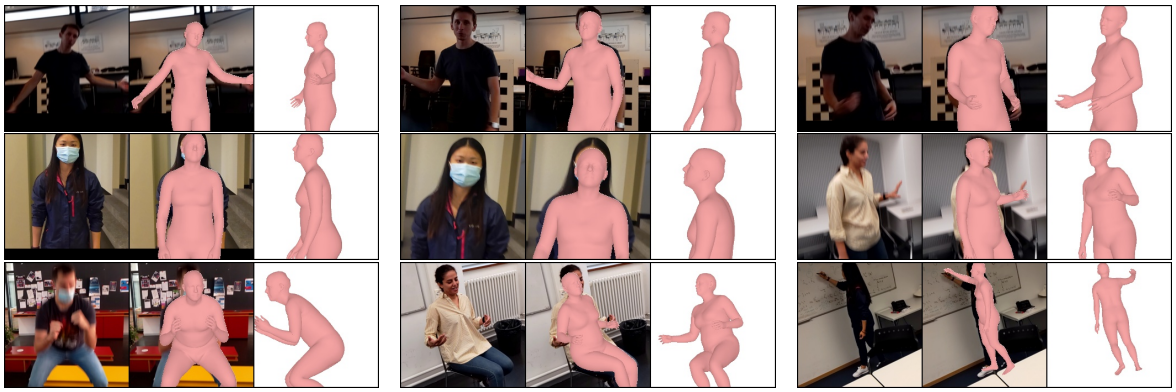
4.2.2 Training the proposed scene-based method

Based on PARE[24], we add an additional scene branch with the hope of encoding scene information into the estimation process. In the training phase, we use the pretrained PARE model weights as the initial weights for the 2D part branch and 3D body branch, meanwhile, we use the above-mentioned pretrained SegNet weights as the initialization for the 2D scene branch. We train the model on EgoBody training split with the same hyper-parameters as PARE. For a fair comparison, we also fine-tune PARE model on the same training split. The evaluation results on the test split are shown in Table4.3.

Model	MPJPE	PA-MPJPE	V2V
EFT-ft[64]	123.9	78.4	135.0
SPIN-ft[64]	106.5	67.1	120.9
METRO-ft[64]	98.5	66.9	110.5
PARE	117.4	76.0	131.16
PARE-ft	88.7	59.6	99.8
PARE-scene(mask)	90.8	61.4	101.5
PARE-scene(image)	86.4	59.8	95.1

Table 4.3: Evaluation of fine-tuned EFT[18], SPIN[26], METRO[29], PARE[24], our proposed scene method with mask input, our proposed method with label image input on EgoBody full test set. All metrics are in mm.

As it is shown in the above table, the fine-tuned PARE[24] model has better results than EFT[18], SPIN[26], METRO[29]. Since the result of our proposed method does not show an obvious improvement, we can not draw the conclusion that the scene branch is beneficial for the 3D human pose and shape estimation. However, the potential could be further explored by considering various ways of combining the scene feature and the original PARE feature or different backbones for extracting scene features.



(a) PARE-ft (b) PARE-scene(mask) (c) PARE-scene(image)

Figure 4.5: Qualitative results on fine-tuned PARE model, our proposed scene model with mask input and our proposed scene model with label image input. From left to right: input image, mesh overlaid on the image, sideview of the mesh.

The visualization of the output reconstruction is shown in Fig.4.5 Notably, different from PARE's training data which has 49 ground-truth 2D keypoints(25 openpose keypoints, 24 ground-truth keypoints), EgoBody only provides 25 openpose keypoints. Therefore, our 2D keypoint loss is only conditioned on those 25 keypoints, which might be weaker than the original PARE 2D keypoint loss.

Chapter 5

Discussion

In this work, in terms to understand the main challenge of egocentric data, we conduct various evaluation studies on the EgoBody dataset. We have conducted state-of-the-art 3DHPS method PARE[24] on the full test split, blur images subset, and sharp images subset in order to study the effect of blurriness. We also developed the metric to evaluate the error of in-frame joints which are inside the bounding box of the input image. We derived the conclusion that the main error is caused by the bad inference of the truncated joints which are mostly the lower-body joints. We evaluated SPEC[25] on the test split with different camera calibrations. We found that adding more accurate camera parameters to the estimation process does not have a big influence on the reconstruction. In order to have a more reasonable unseen body part prediction, we propose a scene-based method that adds an additional 2D scene branch to PARE[24] structure. In the future, research on the scene branch backbone and the scene constraint loss term could be conducted for a more comprehensive study.

Bibliography

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

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

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

[4] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020.

[5] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1964–1973, 2021.

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

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

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

[9] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012.

[10] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *2011 international conference on computer vision*, pages 407–414. IEEE, 2011.

[11] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009.

[12] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019.

[13] 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.

[14] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[15] Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. Latent structured models for human pose estimation. In *2011 International Conference on Computer Vision*, pages 2220–2227. IEEE, 2011.

[16] 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.

[17] 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.

[18] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021.

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

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

[21] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019.

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

[23] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

[24] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021.

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

[26] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019.

[27] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[28] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017.

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

[30] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding. *arXiv preprint arXiv:2011.13341*, 2020.

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

[32] 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.

[33] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. *arXiv preprint arXiv:2206.09106*, 2022.

[34] 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.

[35] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

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

[37] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3d: Out-of-context data augmentation for 3d scenes. In *2021 International Conference on 3D Vision (3DV)*, pages 116–125. IEEE, 2021.

[38] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011.

[39] 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.

[40] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018.

[41] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *European Conference on Computer Vision*, pages 598–613. Springer, 2020.

[42] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021.

[43] 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.

[44] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018.

[45] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[46] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019.

[47] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020.

[48] Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3d scene reconstruction with multi-layer depth and epipolar transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2172–2182, 2019.

[49] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K Hodgins. Motion capture from body-mounted cameras. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011.

[50] Leonid Sigal, Alexandru Balan, and Michael Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. *Advances in neural information processing systems*, 20, 2007.

[51] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010.

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

[53] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. 2017.

[54] 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. *arXiv preprint arXiv:2011.01519*, 2020.

[55] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *Advances in Neural Information Processing Systems*, 30, 2017.

[56] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36, 2018.

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

[58] Timo Von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1533–1547, 2016.

[59] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3959–3970, 2022.

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

[61] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11038–11049, 2022.

[62] 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.

[63] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12971–12980, 2021.

[64] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape, motion and social interactions from head-mounted devices. *arXiv preprint arXiv:2112.07642*, 2021.

[65] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for real-time multi-person motion capture using multiple video cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1324–1333, 2020.

[66] Rui Zhu, Xingyi Yang, Yannick Hold-Geoffroy, Federico Perazzi, Jonathan Eisenmann, Kalyan Sunkavalli, and Manmohan Chandraker. Single view metrology in the wild. In *European Conference on Computer Vision*, pages 316–333. Springer, 2020.