Università degli Studi di Padova

Institut de Robòtica i Informàtica Industrial

University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Computer Science

3D EgoPose Estimation: Leveraging Multi-View perspectives in the Egobody Dataset

Supervisor
Prof. Lamberto Ballan
University of Padova

Supervisor
Prof. Mariella Dimiccoli
Institut de Robòtica i Informàtica Industrial

Master Candidate
Syed Riaz Raza

Student ID
2041583

Academic Year
2023-2024

Abstract

This thesis addresses the problem of multi-person 3D pose estimation from videos captured by a wearable camera. We focused on the recently introduced EgoBody dataset, a synchronized multi-view dataset, including three third-views and one first-person view of two interacting individuals. Notably, in the first-view the wearer is hidden behind the camera, hence presenting a challenge for his/her accurate 3D pose estimation. The proposed architecture leverages information from both Kinect (third-view) and Holo (first-view) cameras at training time to effectively estimate 3D pose of the camera wearer only from the egocentric view at test time. This thesis proposes an original architecture's design and evaluates its performance on the EgoBody dataset.

Contents

Listing of figures

Listing of tables

1 Introduction

Human pose estimation (HPE) has become a fundamental tool in computer vision, enabling applications in action recognition [11], human-computer interaction [12], and augmented reality [13]. Traditionally, HPE methods relied on images or videos captured from a third-person perspective. However, egocentric videos, captured from a head-mounted camera (HMC) provide a more natural and immersive perspective for tasks like understanding human interaction and manipulation in real-world environments.

Consequently, egocentric video analysis has emerged as a prominent field within computer vision, focusing on understanding human activities and interactions from a first-person (egocentric) perspective. This perspective offers unique challenges and opportunities, particularly in scenarios capturing interactions between individuals.

In particular, accurate egocentric 3D pose estimation has the potential to revolutionize various fields:

- **Augmented Reality (AR):** Imagine an AR system that overlays virtual objects onto the real world based on a user's hand and body pose in an egocentric video. This could be used for tasks like furniture placement in a room, virtual prototyping of products, or even enhancing physical therapy exercises.
- **Virtual Reality (VR):** Egocentric pose estimation can be used to create more immersive VR experiences by accurately reflecting the user's body movements within the virtual world. This can enhance the feeling of presence and improve the user's interaction with virtual objects.

- **Human-Computer Interaction (HCI):** By understanding a user's posture and gestures through egocentric pose estimation, HCI systems can become more intuitive and natural. Imagine controlling a robotic arm or navigating a virtual environment using just your body movements captured by an HMC.
- **Action Recognition and Analysis:** Egocentric pose data can be used to analyze and understand human actions in various contexts. This could be used for applications like sports analysis, gait analysis for medical purposes, or even sign language recognition.

However, egocentric HPE introduces novel technical challenges compared to traditional third-person approaches. The primary challenge arises from the fact that the camera wearer's body is almost fully occluded behind the camera. Additionally, the egocentric perspective introduces major complexity due to motion blur and variations in illumination caused by continuous head movements.

Several research efforts have addressed these challenges and explored techniques for egocentric HPE. Some works focused on single-person pose estimation in egocentric videos, often employing deep learning architectures like Convolutional Neural Networks (CNNs) to directly regress pose parameters from the egocentric frames [5]. A recent work build a shared feature space for egocentric and third-view poses and leveraged it at inference time for enhancing 3D pose estimation [9]. However, for scenarios involving multiple people interacting in the scene, additional considerations are needed.

Recent studies have explored **multi-person pose estimation** in egocentric videos. Some approaches leverage prior knowledge about the scene layout or interaction context to improve pose estimation accuracy [3]. For instance, the work by [14] utilizes information about object affordances and hand-object interactions to enhance pose estimation in egocentric settings.The work in [8] utilizes information about the visible (from the first -person perspective) interacting person to improve the 3D camera wearer pose estimation. However, relying on such contextual cues can limit the generalizability of the model to scenarios where this information may not be readily available.

The work in has [5] contributed valuable insights the domain of Model-fitting in the Loop by utilizing SMPL [15]. These studies have demonstrated the importance of leveraging contextual cues and multi-view information for robust pose estimation.

The paper addressed the challenge of **multi-person pose estimation** in egocentric videos, particularly focusing on addressing missing data due to self-occlusions. We propose a novel approach that leverages information from complementary viewpoints within an egocentric dataset to achieve accurate pose estimation for both individuals in the scene, even when one

person is occluded in the HMC view. Our work builds upon the success of existing deep learning architectures for HPE while introducing a novel training strategy that allows the model to learn from complete data (both people visible) and generalize to scenarios with missing data (one person occluded).

The paper also addressed another challenge of pose estimation in synchronized **multi-view egocentric videos** using data from both Hololens and Kinect cameras. The dataset utilized for this study, EgoBody, provides multiple subsets with synchronized views capturing interactions between two individuals. This dataset presents a unique scenario where one individual, the wearer of the Hololens camera, is not visible in the captured frames, posing a significant challenge for traditional pose estimation methods.

The overarching goal of this research is to develop an architecture that can effectively leverage information from both Kinect and Hololens cameras at training to accurately estimate the poses of both individuals at test time, where only the Hololens view is available. This architecture is crucial for applications such as augmented reality scenarios, where understanding the interactions between individuals is essential for immersive experiences.

1.0.1 Contributions

The contributions of this thesis include:

- Proposing a novel architecture for pose estimation trained in synchronized multi-view egocentric videos captured using Hololens and Kinect cameras.
- Leveraging information from third-view and first-view during training to learn parameters for accurate 3D pose estimation of two interacting individuals from the only egocentric perspective provided by the Hololens.
- Demonstrating the effectiveness of the proposed architecture through experimental evaluation on the EgoBody dataset.

1.0.2 Composition of Thesis

The composition of this thesis is structured into several key chapters, each contributing to a comprehensive exploration of the proposed approach for **Multi-View Multi-Person Human**

Pose Estimation (MV MP HPE) using the **EgoBody dataset** and modified **Skinned Multi-Person Linear Model (SMPL)**.

1. **Introduction:** This opening chapter sets the stage for the research by outlining the main objectives, the motivation behind the study, and the specific challenges addressed by the proposed methodologies.

2. **Background:**

   (a) This section introduces the **Skinned Multi-Person Linear Model (SMPL)**, a key foundation for much of the work in 3D human pose and shape estimation. The nuances of SMPL and its relevance to accurate and realistic human modeling are discussed.

   (b) Joint location estimation is explored in depth, highlighting the critical role of accurately identifying joint positions in the effective application of SMPL models for human pose estimation tasks.

   (c) The adaptation and application of SMPL across various datasets are examined, demonstrating the flexibility and challenges of utilizing SMPL in diverse research contexts.

3. **Related Work:**

   (a) A comprehensive survey of existing methodologies in single-person and multi-person pose estimation sets the context for the proposed research, showcasing the evolution of techniques and the state-of-the-art.

   (b) The role of **Human Mesh Recovery (HMR)** within the framework of SMPL modeling is detailed, emphasizing its importance in achieving realistic and accurate human pose reconstructions.

   (c) Various strategies for addressing missing data in pose estimation tasks are discussed, an area of significant challenge in the field.

   (d) The section concludes with a detailed examination of approaches to human pose estimation using egocentric data, a relatively less explored area that the current research contributes to significantly.

4. **The EgoBody Dataset:**

   (a) The characteristics that distinguish the EgoBody dataset from existing datasets in the domain of human pose estimation are highlighted, underscoring its value and potential for research.

(b) A comparison with other datasets illustrates the unique advantages and challenges presented by the EgoBody dataset, providing justification for its selection in this study.

5. **Proposed Approach:**

   (a) The chapter begins with an overview of the network architecture modifications designed to leverage the EgoBody dataset for multi-view, multi-person human pose estimation, highlighting the innovative aspects of the approach.

   (b) A significant portion is dedicated to discussing the integration of Kinect and HoloLens views, elaborating on the methods and challenges associated with merging these perspectives for enhanced pose estimation accuracy.

   (c) Detailed descriptions of the **preprocessing steps** and **data augmentation techniques** employed enhance the dataset's utility for the research objectives, showcasing the meticulous approach to dataset preparation.

   (d) The implementation details, including the choice of loss functions and training methodologies, are provided, offering insights into the practical aspects of bringing the proposed approach to fruition.

6. **Experimental Results:**

   (a) This chapter presents a thorough evaluation of the proposed methodologies, starting with the definition of evaluation metrics and proceeding to a detailed analysis of quantitative and qualitative results.

   (b) Baselines are established for comparison, and the performance of the proposed approach is critically examined against these benchmarks.

   (c) A discussion section delves into the implications of the findings, evaluating the effectiveness of the loss functions and the overall impact of the research on the field.

7. **Conclusion:** The concluding chapter synthesizes the research findings, highlighting the contributions to the field of multi-view multi-person pose estimation . Reflections on the advancements achieved, alongside considerations for future research, draw the study to a close.

2 Background

2.1 Introduction to the Skinned Multi-Person Linear Model (SMPL)

SMPL is a widely used representation of human body shape and pose aimed at facilitating 3D pose estimation and human mesh recovery [15]. SMPL represents the human body as a deformable mesh parameterized by three sets of parameters: pose parameters, shape parameters, camera parameters, and global translation respectively θ, β, γ, ϕ are optimized SMPL-X parameters.

1. **Pose Parameters:** These parameters encode the joint rotations of the human body in a 3D space. By adjusting the pose parameters, one can control the orientation and configuration of the body joints, enabling the representation of various body poses and movements.

   A pose vector of 24x3 scalar values that keeps the relative rotations of joints with respective to their parameters. Each rotation is encoded as a arbitrary 3D vector in axis-angle rotation representation. In the SMPL model, the human skeleton is described by a hierarchy of 24 joints as shown by the white dots in the below figure. This hierarchy is defined by a kinematic tree that keeps the parent relation for each joint

2. **Shape Parameters:** Shape parameters (shape vector of 10 scalar values) determine the geometric variations in the human body shape. These parameters capture individual

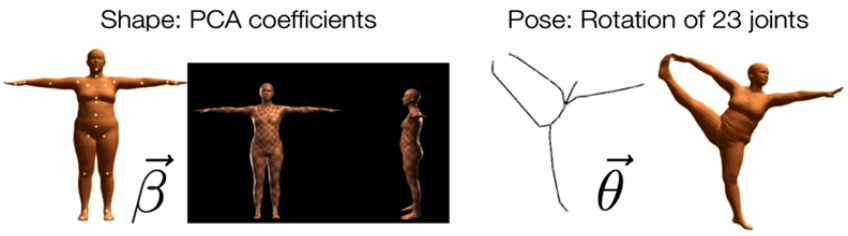

Figure 2.1: SMPL Parameters Representation

differences such as height, weight, and body proportions, allowing for the generation of diverse human body shapes within the SMPL framework.

3. **Global Translation:** The global translation parameter specifies the overall position and orientation of the human body in the world coordinate system. It accounts for translations and rotations of the entire body relative to the camera viewpoint.

By estimating these SMPL parameters from input images, one can reconstruct the 3D human pose and shape, facilitating tasks such as motion capture, animation, and virtual try-on.

2.2 Joint Location estimation

Thanks to the fixed mesh topology of the SMPL model, each joint location could be estimated as an average of surrounding vertices.

This average is represented by a joint regression matrix learned from the data-set that defines a sparse set of vertex weight for each joint. As shown in the below figure, the knee joint will be calculated as a linear combination of red vertices, each with a different weight.
The below code shows how to regress joint locations from the rest-pose mesh:

```
v_shape: 6890x3 #the mesh in neutral T-pose calculated from a shape parameter of 10
    scalar values.
self.J_regressor: 24x6890 #the regression matrix that maps 6890 vertex to 24 joint
    locations
self.J: 24x3 #24 joint (x,y,z) locations
self.J = self.J_regressor.dot(v_shaped)
```

2.3 SMPL IN DIFFERENT DATASETS

SMPL-Body offers a versatile framework for distributing created meshes and poses akin to a standardized format for human bodies, comparable to a "pdf" for bodies. However, the specific definition of joints and the arrangement of the kinematic tree can vary significantly across different 3D datasets used for human pose estimation and modeling2.2. In practice, the actual joints of the human body are never directly observed but are inferred, often through motion capture (mocap) systems.

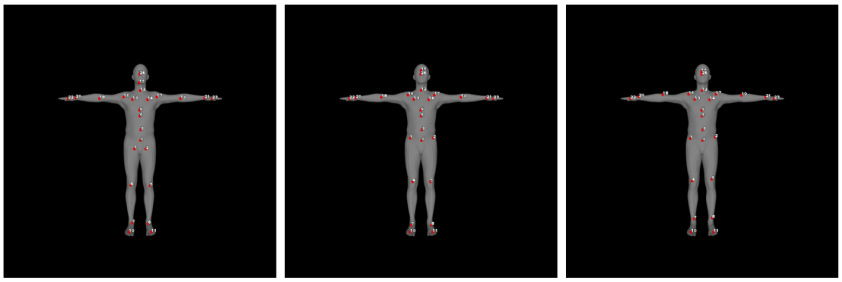

Figure 2.2: Joint Regressor in different datasets

For instance, when utilizing datasets like Human3.6M (H3.6M), it becomes necessary to transform the SMPL model's joints to align with the specific 3D joint definitions used in the dataset. This transformation accounts for differences in joint naming and hierarchy, where higher numbers of joints defined in the dataset may need to be mapped or reduced according to the dataset's kinematic tree structure. Fig 2.3.

Additionally, when training models with 2D joint annotations from sources like OpenPose, there is a need to map these 2D joints to corresponding 3D joints that project into the image plane and align with the observed 2D annotations. Tools like the JointMapper provided in [16] facilitate this mapping process, allowing for seamless integration of SMPL or SMPL-X joints with OpenPose's Coco_25 or coco_19 joints. Some datasets introduce additional joints beyond the standard SMPL definition, such as 45 or 49 joints, which serve as supplementary joints for joint regression tasks.
Mapping these extra joints requires knowledge of joint names and their associated kinematics, as outlined in the joint mapping information provided in [17]. Understanding and adapting

Figure 2.3: The **left image** represent SMPL joint ids to joint names, the **middle one** represent unique joints connectivity each method use (SPIN for us) and with that we can visualize a Kinematic tree shown in the **right image**

to the nuances of joint definitions and kinematic trees across datasets are essential for ensuring accurate and compatible pose estimation across diverse datasets and applications within the field of human body modeling and analysis.

3
Related work

3.1 Single-Person Pose Estimation:

Early works on egocentric HPE focused on single-person pose estimation. One approach utilizes Convolutional Neural Networks (CNNs) to directly regress pose parameters from egocentric frames [5]. However, these methods struggle with complex scenarios involving multiple people interacting in the scene. Single-person 3D HPE methods can be categorized into two main approaches based on their output representation: skeleton-only and human mesh recovery (HMR).

1. Skeleton-Only Approaches:

- Focus on estimating the 3D locations of key body joints (e.g., elbows, wrists, ankles) as the final result.
- Do not use a detailed human body model to reconstruct a complete 3D human mesh.
- Can be further divided into:
  - Direct Estimation: 3.1(a) These methods directly predict the 3D pose from a 2D image, without first estimating a 2D pose representation.
  - 2D to 3D Lifting: 3.1(b)Inspired by advancements in 2D HPE, these methods estimate 3D pose by leveraging an intermediate step of predicting 2D pose first.

2. Human Mesh Recovery (HMR) Approaches:

- Utilize a parametric human body model to recover a more detailed 3D human mesh representation.
- This mesh goes beyond just joint locations and captures the overall shape of the body.
- Typically involve a 3D pose and shape network that estimates parameters controlling the body model, ultimately generating the 3D mesh. 3.1(c)

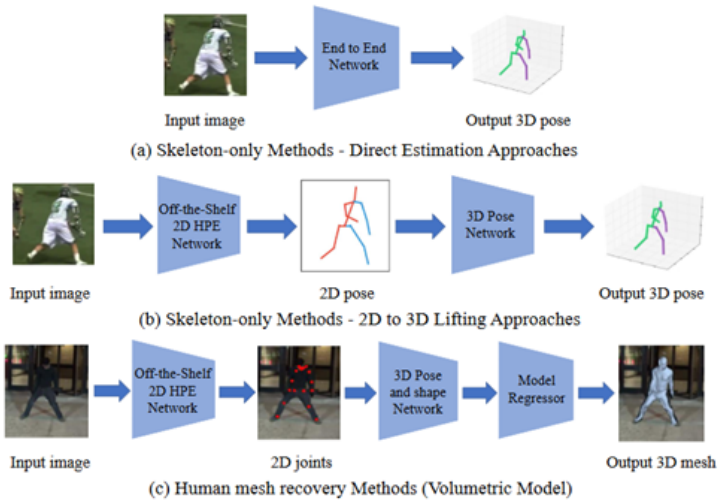

Figure 3.1: Single-person 3D HPE frameworks. (a) Direct estimation approaches directly estimate the 3D human pose from 2D images. (b) 2D to 3D lifting approaches leverage the predicted 2D human pose (intermediate representation) for 3D pose estimation. (c) Human mesh recovery methods incorporate parametric body models to recover a high-quality 3D human mesh. The 3D pose and shape parameters inferred by the 3D pose and shape network are fed into the model regressor to reconstruct 3D human mesh.[1]

3.2 Multi-Person Pose Estimation:

Several recent studies have explored multi-person pose estimation in egocentric videos. Some approaches leverage contextual information about the scene or interaction context to improve pose estimation accuracy. For instance, [3] utilizes knowledge about object affordances and hand-object interactions to enhance pose estimation. However, relying on such contextual cues limits generalizability to scenarios where this information might not be readily available.

Similar to 2D multi-person HPE, 3D multi-person HPE can be categorized into two main approaches: top-down and bottom-up (illustrated in Figure 3.2)

Top-Down Approach 3.2(a)

1. Human Detection: This methd first utilizes a human detection network to identify individual people in the image or video frame. This results in bounding boxes or keypoint detections for each person.

2. 3D Pose Estimation: For each detected person, a separate 3D pose network estimates their individual 3D pose representation. This representation could be in the form of joint locations in 3D space or parameters of a 3D body model.

3. World Coordinate Alignment: Finally, all estimated 3D poses are aligned to a common world coordinate system, allowing for analysis of relative positions and interactions between people.

Bottom-Up Approach 3.2(b)

1. Joint and Depth Estimation: This approach directly estimates the locations of all body joints in the image or video frame, along with a depth map for each pixel.

2. Person Association: Body parts are then associated with specific individuals based on estimated root joint depth and relative depth of different body parts. This step groups the estimated joints into individual 3D pose representations for each person in the scene.

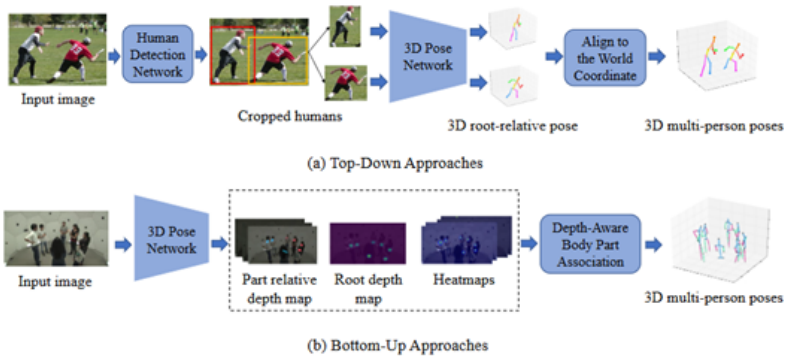

Figure 3.2: Illustration of the multi-person 3D HPE frameworks. (a) Top-Down methods first detect single-person regions by human detection network. For each single-person region, individual 3D poses can be estimated by 3D pose network. Then all 3D poses are aligned to the world coordinate. (b) Bottom-Up methods first estimate all body joints and depth maps, then associate body parts to each person according to the root depth and part relative depth.

3.3 Multi-View Pose Estimation:

Multi-view images can reduce the ambiguity significantly. However, it is challenging to fuse information from multiple views. Typical methods include fusing multi-view 2D heatmaps [18], enforcing multiple view consistency [19], triangulation [2], and utilizing the SMPL model [20].

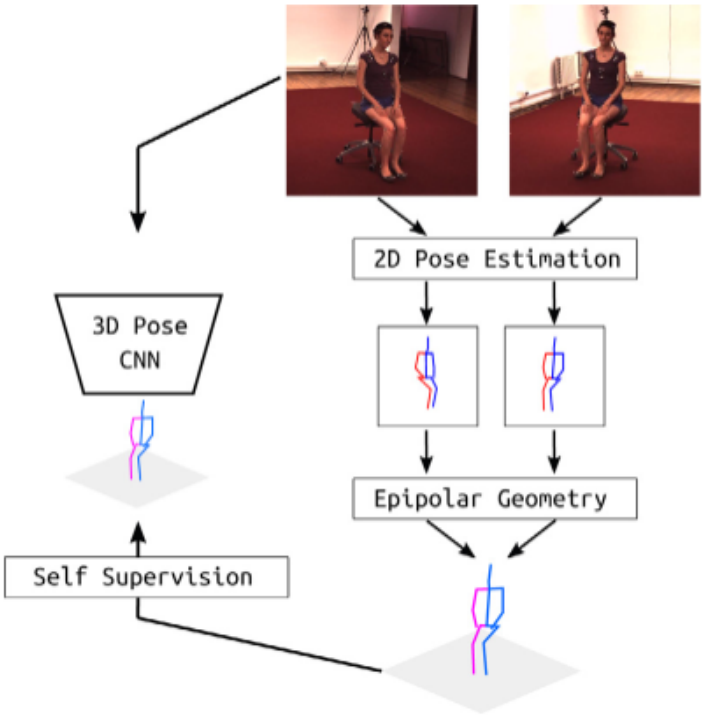

Figure 3.3: Main framework of EpipolarPose [2].

Triangulation is another fundamental method for reconstruction in computer vision. EpipolarPose [2] uses the epipolar geometry method to recover the 3D pose from the 2D poses and uses it as a supervision signal to train the 3D pose estimation model, as shown in Fig. 3.3. first propose a baseline method that feeds the 2D joint confidences and 2D positions of all views produced by the 2D pose detector to the algebraic triangulation module to obtain the 3D pose.

Markerless motion capture has been a subject of research in computer vision for over a decade. Initial efforts focused on tracking the 3D skeleton or geometric model of the human

body across multi-view sequences [21]. However, these tracking-based methods require initialization in the first frame and are susceptible to local optima and tracking failures. Consequently, more recent approaches have shifted towards a bottom-up strategy, reconstructing 3D pose from 2D features detected in images [22]. Notably, recent advancements [23] have demonstrated impressive results by integrating statistical body models with deep learning-based 2D detectors.

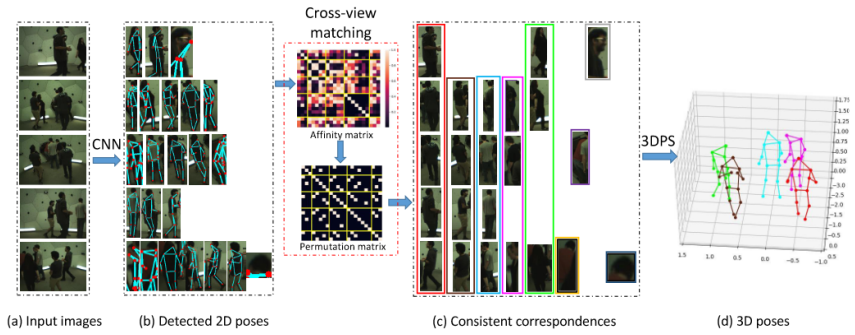

Figure 3.4: Overview of the proposed approach of [3]. Given images from a few calibrated camera

In this thesis, our emphasis is on multi-person multi-view 3D pose estimation. Previous research predominantly employs 3DPS models, where nodes represent the 3D locations of body joints and edges encode pairwise relations between them [24],[25],[26],[27]. [3] FastMvPose worked on identifying and matching correspondences using for cross-view matching.The main challenge of this problem is to find the cross-view correspondences among noisy and incomplete 2D pose predictions.Fig 3.4

A great work work done by [4], MvP represents skeleton joints as learnable query embeddings and let them progressively attend to and reason over the multi-view information from the input images to directly regress the actual 3D joint locations, greatly decreasing computational cost. MvP also have also compared their methods with the previous one. Typically, the state space for each joint is represented as a 3D grid, discretizing the 3D space. The likelihood of a joint's location is determined by a joint detector applied across all 2D views, and pairwise potentials between joints are defined by skeletal constraints [24],[25] or body parts detected in 2D views [27], [28]. Subsequently, the 3D poses of multiple individuals are jointly inferred through maximum a posteriori estimation.

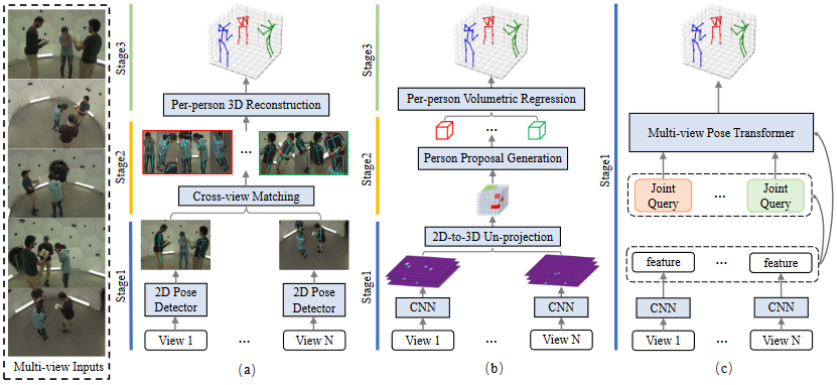

Figure 3.5: (a) reconstruction-based or (b) volumetric representation based, which incur heavy computation burden. (c) [4] MvP method solves this task as a direct regression problem without relying on any intermediate task by a novel Multi-view Pose Transformer, and largely simplifies the pipeline and boosts the efficiency.

3.4 Human Mesh Recovery (HMR) with SMPL

The Human Mesh Recovery (HMR) [6] model builds upon the SMPL representation to recover detailed 3D human meshes from 2D images. HMR leverages deep learning techniques to predict the SMPL parameters directly from input images, enabling the reconstruction of accurate 3D human poses and shapes. Their are different method and approaches proposed after [6] to reduce the reconstruction error, some of these approaches are shown in 3.6.

The integration of SMPL and HMR into our proposed architecture enables us to leverage the rich representational power of these models for pose estimation in synchronized multi-view egocentric videos. By predicting SMPL parameters for each individual in the scene, we can reconstruct their poses and shapes, facilitating a wide range of applications in augmented reality, virtual reality, healthcare, and security.

Single-image and video-based HPE methods have achieved significant progress [1, 4]. However, applying these approaches directly to egocentric videos is challenging due to self-occlusions and limited field of view. Several works have addressed egocentric HPE, with some focusing on single-person pose estimation [5]. Others explore multi-person pose estimation, but often require additional information or assumptions about the scene or interaction context [3].

The key challenge in our work is estimating the pose of a person occluded in the HoloLens view. Recent approaches like [6] utilize model-fitting techniques for 3D pose reconstruction, but these methods may struggle with occlusions. Our work addresses this challenge by leveraging the complete data from the Kinect views to learn a model that can generalize to the single-view HoloLens data.

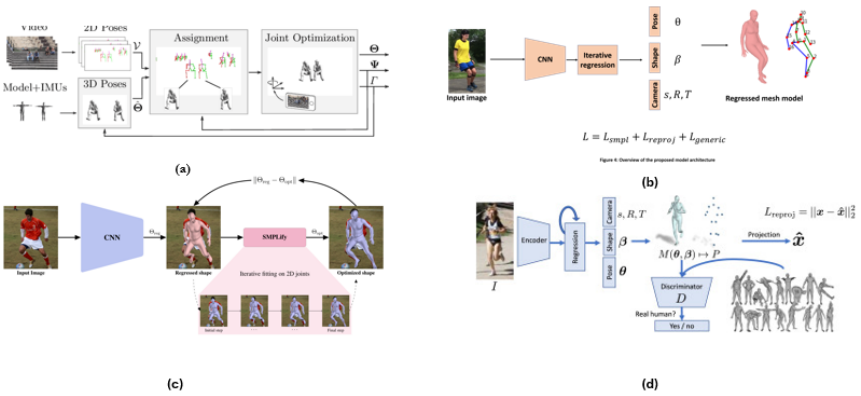

Figure 3.6: (a) Method works by fitting the SMPL body model to the measured IMU orientations they obtain initial 3D poses Θ. (b) The proposed model takes as input a human body image and output 3D body deformable model and camera parameters. (c) SPIN[5] trains a deep network for 3D human pose and shape estimation through a tight collaboration between a regression-based and an iterative optimization-based approach (d) [6] They take a picture image and use a special program (convolutional encoder) to understand the main shapes and features and guess a 3D model, trying different versions until they find one that matches the person's joints and finally, they check their guess against a "checker" (discriminator) to make sure it looks like a real person.[7]

Human pose estimation (HPE) in egocentric videos presents unique challenges compared to traditional third-person approaches. This section reviews existing research addressing these challenges, particularly focusing on multi-person pose estimation with missing data due to occlusions.

3.5 Addressing Missing Data:

A significant challenge in egocentric HPE is handling missing data due to self-occlusions. Some works address this by employing model-fitting techniques for 3D pose reconstruction [6]. However, these methods can struggle with heavily occluded body parts.

3.6 HPE with Egocentric Data:

Research in egocentric or first-person view video understanding has seen significant advancements in recent years, fueled by studies such as "You2Me" [8], "First2Third" [9], "Mo2Cap2" [29], and "HPS" [30]. This domain poses unique challenges compared to traditional third-person approaches, such as limited field of view due to self-occlusions and variations in illumination caused by head movements. These works collectively aim to understand and interpret

visual content from the wearer's perspective, yet they each contribute unique perspectives and methodologies to the field.

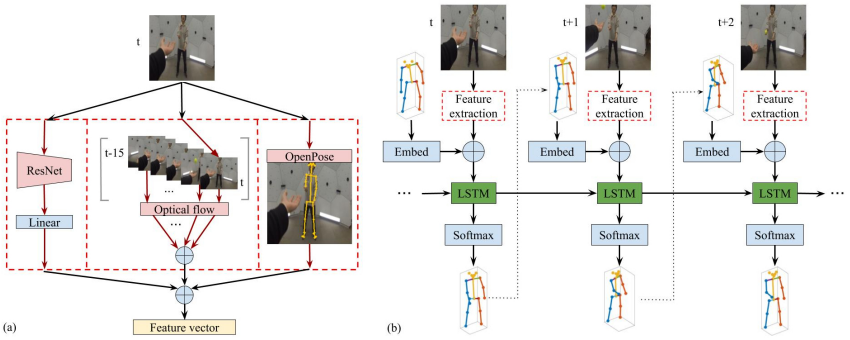

Figure 3.7: An overview of You2Me network.[8]

Learning from interactee (You2Me): This approach leverage information from complementary views of interactee(second-person) within an egocentric dataset. You2Me[8] propose a learning- based approach to estimate the camera wearer's 3D body pose from egocentric video sequences. Their key insight is to leverage interactions with another person—whose body pose we can directly observe—as a signal inherently linked to the body pose of the first-person subject.

Learning From Multiple View (First2Third):[9] exploit synchronized recordings from head-mounted cameras and third view camera to capture complete multi-view data. During training, the model learns the relationship between features in one view (e.g., hand visible in first-person) and the corresponding pose information from another view (e.g., full body from third-person). This allows them to estimate pose in scenarios where parts of the body are occluded in the egocentric view. "First2Third" dataset is the closest to our work for enhancing egocentric 3D pose estimation by incorporating third-person views, bridging the gap between egocentric and third-person perspectives.By leveraging both egocentric and third-person viewpoints, the work aims to improve the accuracy and robustness of pose estimation algorithms. The incorporation of third-person views offers additional context and complementary information that can help resolve ambiguities and challenges inherent in egocentric

Model-Based Reconstruction (Mo2Cap2, HPS): Mo2Cap2 [29] and HPS [30] focus on reconstructing 3D pose by fitting a parametric body model (e.g., SMPL) to image features in

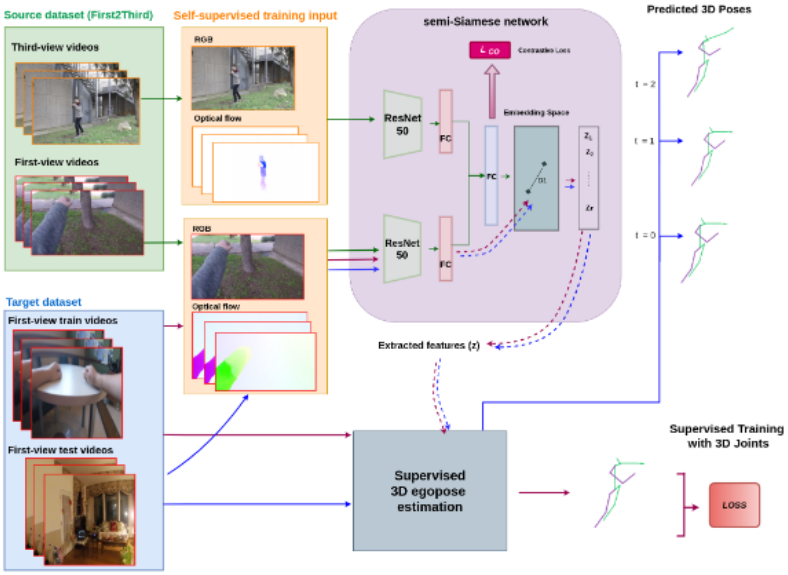

Figure 3.8: An overview model of First2Third which uses a semi-Siamese architecture.[9]

the egocentric frame. While these methods can handle occlusions to some extent, their accuracy can be limited by the quality of the initial pose estimation and the complexity of the body model fitting process.

You2Me focuses on personalized video summarization, leveraging user interaction for content selection. In contrast, "First2Third" and "Mo2Cap2" tackle challenges in human pose estimation and activity recognition from egocentric videos, emphasizing the translation of egocentric observations into conventional third-person representations. HPS introduces a hierarchical approach to address scale variations and occlusions in egocentric videos, enhancing pose estimation accuracy.

3.7 Our Approach in Context:

Our proposed method differentiates itself from existing work in several ways:

1. **Leveraging Multiple Views:** We exploit information from complementary viewpoints within an egocentric dataset. This allows the model to learn from complete data (both people visible) in some views and generalize to scenarios with missing data (one person occluded) in the HMC view.

2. **Leveraging of EgoView for Multi-Person Identification:** We exploit information from egocentric dataset,with the parameters from Kinect View, it is impossible to identify the camera wearer and second person. We compared the parameters based on their points distance in space with EgoView. This allows the model to fuse the complete data coming from EgoView and Kinect View(both people visible in kinect and only one in holoview), we this we completed optimization for final parameters .

3. **Learning from Complete Data:** Our approach utilizes a training strategy that allows the model to learn a robust mapping between image features and pose parameters during training, even when complete data is available. This knowledge is then transferred to estimate poses in scenarios with missing data during testing.

4. **Focus on SMPL Parameters:** Our model predicts SMPL parameters, enabling not only 2D joint localization but also 3D pose reconstruction, crucial for various applications like AR overlays.

4

The EgoBody Dataset

This work utilizes the EgoBody dataset [10] for training and evaluating our proposed approach for multi-person pose estimation in egocentric videos with missing data. Visualized dataset can be seen in the image below. for each frame and each subject, the corresponding SMPL-X body parameters **global translation** γ in $\mathbb{R}^3$, **body shape** β in $\mathbb{R}^{10}$, **pose** θ in $\mathbb{R}^{96}$ (body and hand) and **facial expression** φ in $\mathbb{R}^{10}$, including the Here's why EgoBody is particularly well-suited for our research:

4.1 Characteristics of EgoBody:

- **Multiple Views:** The EgoBody dataset provides synchronized views of a scene captured from four different perspectives: Main Camera, sub1 camera, sub2 camera (all Kinect), and Holo Camera. This multi-view structure allows us to exploit complementary viewpoints for pose estimation, addressing the challenge of missing data due to occlusions in the HMC view.

- **Ground Truth Pose:** Each frame in the dataset is annotated with ground truth pose information for all people present in the scene. This information is provided in the form of SMPL parameters, enabling not only 2D joint localization but also 3D pose reconstruction, crucial for various applications.

- **Diverse Interactions:** The EgoBody dataset captures a wide range of social interactions between two people in various indoor environments. This diversity allows the model to

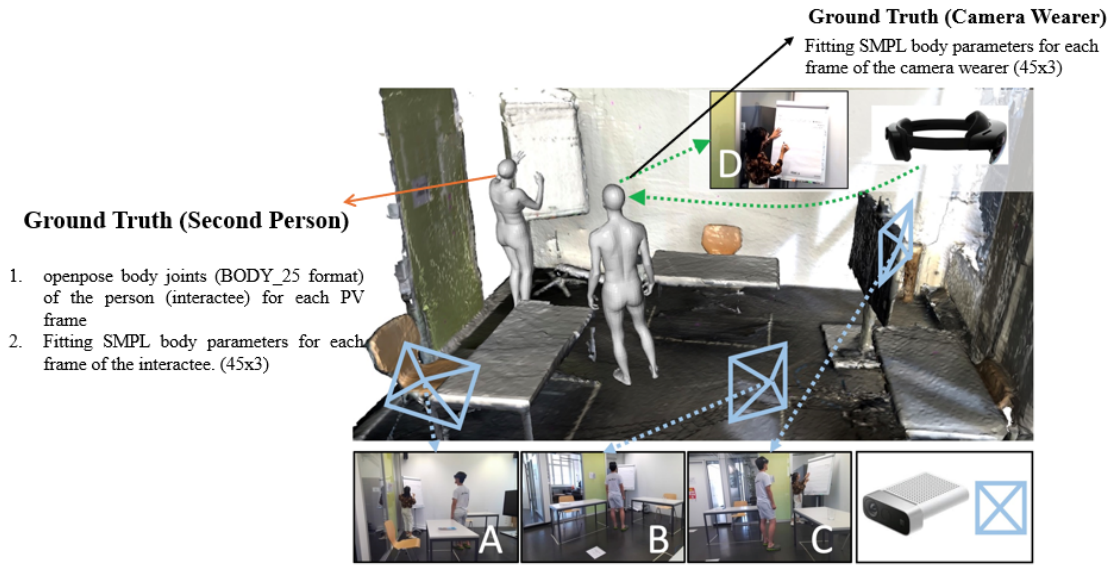

Figure 4.1: Overview of EgoBody Dataset: Capture setup. Multiple Azure Kinects capture the interactions from different views (A, B, C), and a synchronized HoloLens2 worn by one subject captures the egocentric view image (D), as well as the eye gaze (red circle) of the camera wearer.[10]

learn robust pose estimation capabilities that generalize well to unseen scenarios.

EgoBody dataset contains 125 sequences, 36 subjects and 15 indoor scenes. The dataset has 3 subsets:

1. EgoSet (egocentric RGB subset of EgoBody) Egocentric RGB frames captured from the HoloLens, calibrated and synchronized with the Kinect frames

2. MVSet: (third-person view RGBD, 3D scene, eye gaze, etc.) Synchronized frames captured from the Azure Kinects, from multiple third-person views.

3. EgoSet interactee: Frames where the interactee is visible in the egocentric view.

Ground Truth is given in the fitting parameters of SMPL /SMPL-X, both of these models work on three parameters the only difference is in SMPL-X has more joints with extra info about Hands and Face. The GT is always in the coordinate system of the master kinect RGB camera. The dataset has extensive detail about camera parameters

4.2 Comparison to Existing Datasets:

Several egocentric pose estimation datasets exist. However, EgoBody offers distinct advantages for our specific research goals, we did a comparison of Datasets for 3D HPE, with and without data attributes like Egocentric, Interaction, Gaze, and Parametric Model Table 4.1, Table 4.2.

Table 4.1: Comparison of Datasets for 3D HPE, without data attributes like Egocentric, Interaction, Gaze, and Parametric Model

Dataset	Year	Capture system	Environment	Single person	Multi-person	Single view	Multi-view
HumanEva	2010	Marker-based MoCap	Indoor	Yes	No	Yes	Yes
Human3.6M	2014	Marker-based MoCap	Indoor	Yes	No	Yes	Yes
CMU Panoptic	2016	Marker-less MoCap	Indoor	Yes	Yes	Yes	Yes
MPI-INF-3DHP	2017	Marker-less MoCap	Both	Yes	No	Yes	Yes
TotalCapture	2017	Marker-based MoCap with IMUs	Indoor	Yes	No	Yes	Yes
3DPW	2018	Hand-held cameras with IMUs	Both	Yes	Yes	Yes	No
MuPoTS-3D	2018	Marker-less MoCap	Both	Yes	Yes	Yes	Yes
AMASS	2019	Marker-based MoCap	Both	Yes	No	Yes	Yes
NBA2K	2020	NBA2K19 game engine	Indoor	Yes	No	Yes	No
GTA-IM	2020	GTA game engine	Indoor	Yes	No	Yes	No
Occlusion-Person	2020	Unreal Engine 4 game engine	Indoor	Yes	No	Yes	Yes

Table 4.2: Comparison of Datasets for 3D HPE, with additional data attributes like Egocentric, Interaction, Gaze, and Parametric Model

Name	3rd Person	Egocentric	Multi-Person	Multi-View	Interact	Gaze	Parametric Model	3D Scene
You2Me	X	✓	✓	X	✓	X	SMPL-X	X
Mo2Cap2	X	✓	X	X	X	X		X
PROX	✓	X	X	X	X	✓	SMPL-X	✓
First2Third	✓	✓	X	✓	X	X		X
CharadesEgo	✓	✓	X	X	X	X		X
GIMO	✓	✓	X	X	X	✓	SMPL-X	✓
HPS	X	✓	X	X	X	X	SMPL	✓
Assembly101	✓	✓	X	X	X	X		X
HUMBI	✓	X	X	X	X	✓		X
Ego Body	✓	✓	✓	✓	✓	✓	**SMPL-X**	✓

5

Proposed Approach

This section details our proposed deep learning architecture for human pose estimation (HPE) in egocentric videos with missing data. Our approach leverages the information from both the Kinect and HoloLens views during training and generalizes this knowledge to estimate poses from single HoloLens frames during testing, where one person might be occluded.
In this chapter, we will explain the overview of our Model MVMP HMR 5.1, SMPL and JRegressor, SPIN Optimatization Method and our Regression model which we used in submodels SP HMR, EgoHMR, MP HMR.

5.1 Architecture of MVMP-HMR:

Our architecture Figure 5.1 builds upon the well-established Human Mesh Recovery (HMR) [1] for human pose estimation. However, we modify the HMR model and added it our architecture to handle two people present in two scene and address the challenge of missing data in the HoloLens view.

The network takes images from both the Kinect and HoloLens views as input during training stage and only the HoloLens View input during testing stage. Whole architecture can be divided into 4 parts: (a) Kinect Views (MP-HMR),(b) HoloLens View (SP-HMR), (c) HoloLens View (Ego-HMR), (d) Addressing Person Identification, (e) SMPL Optimization. Here's a breakdown of the processing for each view:

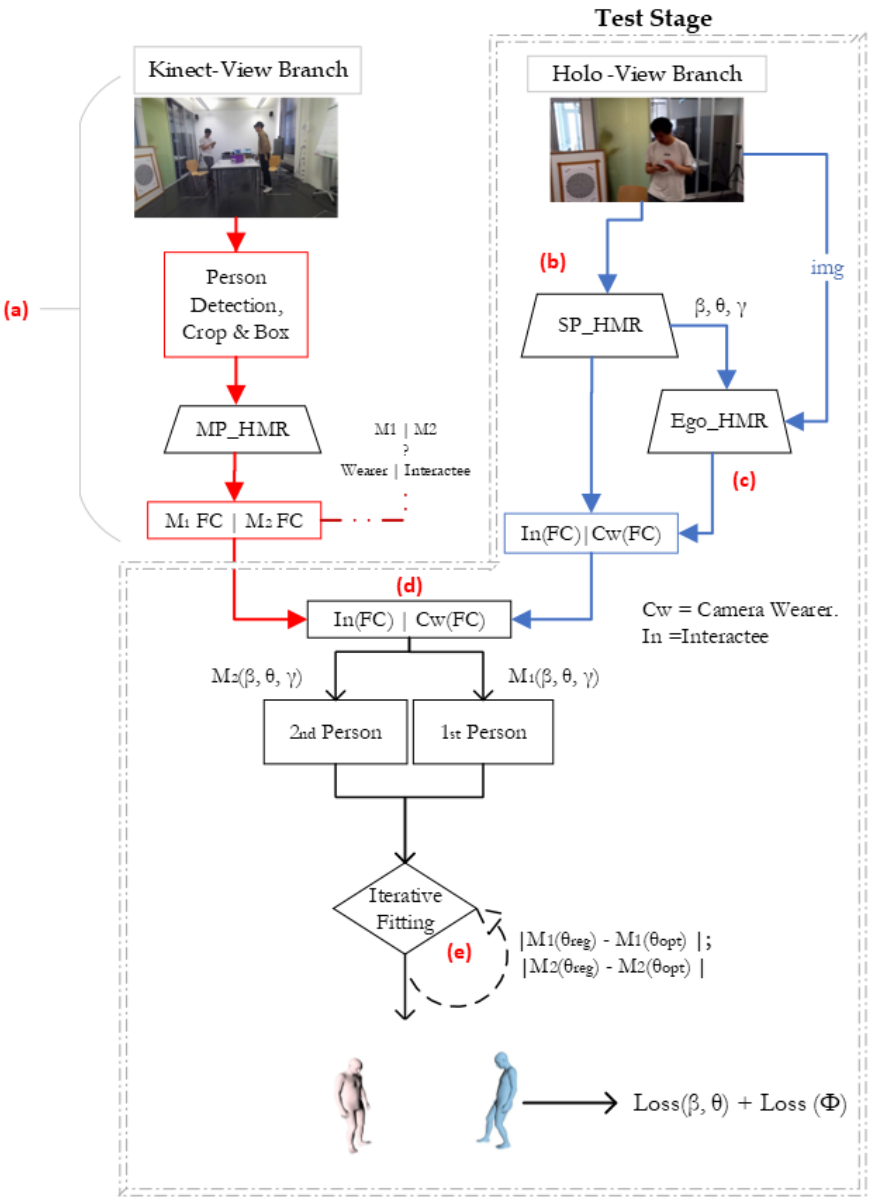

Figure 5.1: Overview of our architecture: MVMP HMR. Whole architecture can be divided into 4 parts: (a) Kinect Views (MP-HMR), (b) HoloLens View (SP-HMR), (c) HoloLens View (Ego-HMR), (d) Addressing Person Identification, (e) SMPL Optimization

5.1.1 Kinect Views (MP-HMR - Multi-Person HMR):

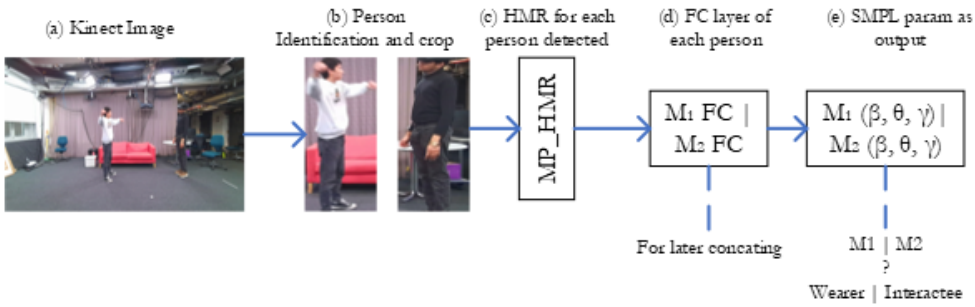

Figure 5.2: Overview of Kinect branch, this is the detail part of 5.1 (a)

- **Person Identification and Pose Estimation:** The network first identifies individual people within the Kinect image. This was achieved using Yolo3, whose accuracy is much better than CV2 HOG Descriptor.
- **Cropping and Individual Pose Estimation:** For each identified person, the image is cropped around the bounding box and fed into a separate HMR branch. This branch estimates the individual's SMPL parameters (M1(β, θ, γ) and M2(β, θ, γ)) using the standard HMR architecture.
- **Temporary SMPL Parameters:** The predicted SMPL parameters which represents human pose and shape can be used to project both 2D and 3D joint locations. These initial estimates serve as temporary SMPL parameters, to identify correspondences of Camera Wearer and Interactee by the results from EgoView Branch. After identification corresponding FC will be merged with M1, M2 FC.

5.1.2 HoloLens View (SP-HMR):

- **Single Person Identification (SP-HMR):** The HoloLens image is processed through a dedicated Single-Person HMR (SP-HMR) branch. This branch identifies the person visible in the HoloLens view (the "interactee") and estimates their pose parameters using the HMR architecture.
- **Contextual Information Concatenation:** Inside the SP-HMR before passing the Fully-Connected layer, the features extracted from the HoloLens image by SP-HMR are then concatenated with additional information, which are the **mean-SMPL parameters β, θ, γ**, model is initialized with **idle SMPL body pose**.

- **Output** The concatenated features from SP-HMR are then processed to obtain temporary SMPL parameters for the interactee (visible person) in the HoloLens view.

5.1.3 HoloLens View (Ego-HMR):

- **EgoHMR:** The HoloLens image is processed through a dedicated (EgoHMR) branch. This branch identifies the person not visible in the HoloLens view (the "camera-wearer") and estimates their pose parameters using the HMR architecture.
- **Contextual Information Concatenation:** Inside the Ego-HMR before passing the Fully-Connected layer, the features extracted from the HoloLens image by SP-HMR are then concatenated with additional information, which are the output SMPL parameters from SP-HMR. The Ego-HMR branch leverages the information from both the image and the contextual cues (SMPL Parameters) to refine the pose estimation for the camera wearer, even though they are occluded in the HoloLens view.
- **Output:** The concatenated features from SP-Ego-HMR are then processed to obtain temporary SMPL parameters for the camera-wearer (invisible person) in the HoloLens view
- **Temporary SMPL Parameters from Ego-HMR:** The predicted SMPL parameters which represents human pose and shape can be used to project both 2D and 3D joint locations. These initial estimates serve as **temporary SMPL parameters**, to identify correspondences of Camera Wearer and Interactee by the results from KinectView Branch. After identification corresponding FC will be merged with M1, M2 FC

5.1.4 Identifying Correspondences from two branches:

- A key challenge arises in distinguishing between the camera wearer and the second person (interactee) in the output from the MP-HMR branch processing the Kinect views. To address this, we propose a similarity measure, we compare the distance between the pose parameters of the two people predicted by MP-HMR with the pose parameters estimated by SP-HMR for the visible person in the HoloLens view.

- Based on the similarity measure, we can identify which person in the MP-HMR output corresponds to the camera wearer (occluded in the HoloLens view) and which one corresponds to the interactee (visible in the HoloLens view) foroverview check Fig 5.3 .

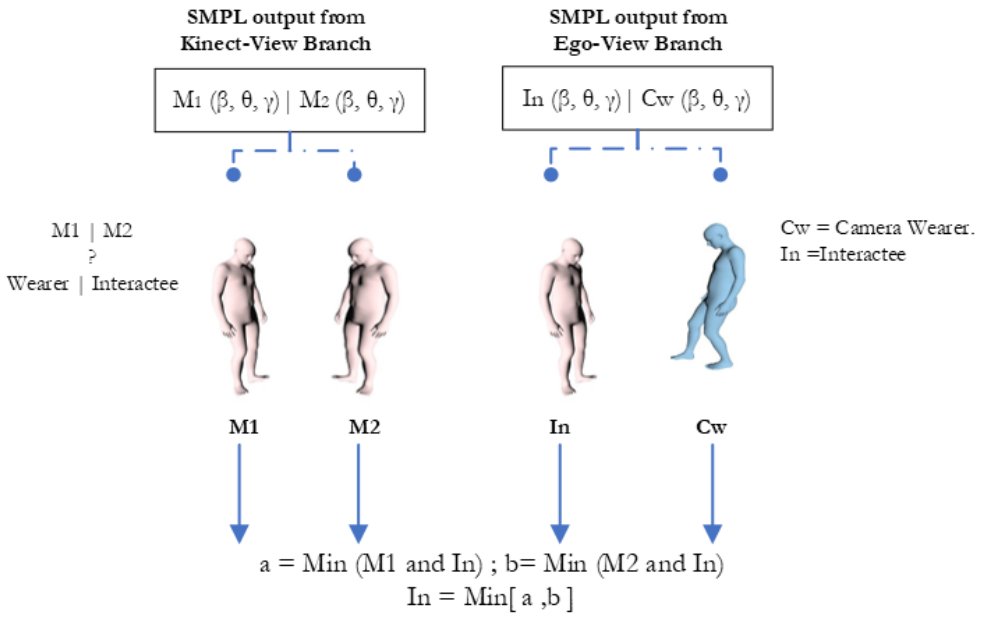

Figure 5.3: Identifying Correspondences of Camera Wearer and Second Person from two branches, this is the detail part of 5.1 (d)

5.1.5 Estimating final parameter for our MVMP-HMR:

- Once the Camera Wearer and Interactee is identified from the MP-HMR output, both the corresponding the temporary SMPL parameters from SP-HMR (representing the interactee) and Ego-HMR (representing the camera-wearer)are concatenated(FC) with temporary SMPL parameters from respective bodies in MP-HMR.
- The predicted parameters are than fitted using techniques mentioned in SPIN.

5.2 SMPL

The Skinned Multi-Person Linear (SMPL) model [15] is a widely used representation for human body shape and pose estimation. It provides a function $M(\theta, \beta)$ that takes pose parameters θ and shape parameters β as input and returns a body mesh $M \in \mathbb{R}^{N\times 3}$, where $N = 6890$ vertices. This function allows for the generation of detailed 3D human body meshes based on specific pose and shape configurations.
One of the key advantages of the SMPL model is its ability to define body joints X directly from

the generated mesh vertices. Specifically, the body joints X can be represented as a linear combination of the mesh vertices, enabling the efficient extraction of joint locations from the mesh geometry. This linear relationship can be encapsulated using a pre-trained linear regressor W, which maps the mesh vertices to the desired joint locations.
Mathematically, for k joints of interest, the major body joints $X \in \mathbb{R}^{k\times 3}$ can be defined as:

$$\mathbf{X} = \mathbf{W}\,\mathbf{M}$$

where X represents the joint locations. M is the body mesh generated by the SMPL model. W is a pre-trained regressor matrix that maps mesh vertices to joint locations.
We stimate camera translation given the model joints and 2D keypoints by minimizing a weighted least squares loss. We find camera translation that brings 3D joints S closest to 2D the corresponding joints_2d:

```
Input:
    S: (B, 49, 3) #3D joint locations
    joints: (B, 49, 3) #2D joint locations and confidence
Returns:
    (B, 3) #camera translation vectors
```

By utilizing the SMPL model and the associated linear regressor W, researchers can effectively extract joint information from generated body meshes, facilitating tasks such as pose estimation, motion capture, and animation. The SMPL model's parameterization of pose (θ) and shape (β) allows for versatile and customizable representations of human bodies, making it a valuable tool in computer graphics and vision research.

5.3 SPIN for Optimization

Our method leverages the SPIN (SMPL oPtimization IN the loop) [5] framework, which was developed with a specific strategy in mind. SPIN harnesses the synergy between two paradigms to effectively train a deep regressor for human pose and shape estimation (see Figure 5.1 e). In our approach, during the training process, input images are passed through the network to obtain regressed parameters denoted as Θ_{reg}.
Instead of directly applying standard 2D reprojection losses at this stage, the regressed parameters are utilized to initialize an optimization routine. Typically, starting this optimization from a mean pose value can be slow. However, with a reasonable initial estimate, the optimization

process can be significantly accelerated. This allows us to integrate the fitting routine directly within the training loop.

Let $\Theta_{\text{opt}} = \{\theta_{\text{opt}}, \beta_{\text{opt}}\}$ represent the optimized set of model parameters obtained through iterative fitting. These optimized values are explicitly tuned so that the resulting shape $M_{\text{opt}} = M(\theta_{\text{opt}}, \beta_{\text{opt}})$ and the reprojected joints J_{opt} align closely with the 2D keypoints. Using these optimized parameters, we can directly supervise the network function f on the parameter level using the loss:

$$L_{3D} = ||\Theta_{\text{reg}} - \Theta_{\text{opt}}|| \quad (3)$$

and/or on the mesh level using the loss:

$$L_M = ||M_{\text{reg}} - M_{\text{opt}}|| \quad (4)$$

This approach differs significantly from applying a reprojection loss solely on the 2D joints. Rather than requiring the network to identify parameters that satisfy the joints' reprojection, we provide it with a parametric solution corresponding to a feasible 3D shape directly. Essentially, we circumvent the network's search in the parameter space by supplying privileged parameters Θ_{opt} that are very close to the optimal solution.

An important characteristic of SPIN is its inherent self-improvement. A good initial estimate Θ_{reg} from the network leads to improved optimization results Θ_{opt}, which in turn provides enhanced supervision to the network. This iterative process within the training loop is crucial as it fosters a close collaboration between the two components, leading to continuous improvement and refinement. This improvement and optimization can be compared with other methods, check Table 5.1

Method	Reconstruction Error (%)
NBF [31]	59.9
HMR [6]	56.8
SPIN [5]	41.1

Table 5.1: Reconstruction Error Comparison: The numbers are mean reconstruction errors in mm. We compare with approaches that output a mesh of the human body. The approaches make use of 3D ground truth too. Using SPIN outperforms the state-of-the-art by significant margins

5.4 Dataset Preprocessing

For preprocessing the EgoBody dataset, we undertook several steps to adapt the ground truth annotations and camera data for our model's requirements. Initially, the dataset was annotated based on **Kinect camera data**, with pose parameters provided for **45** joints. To enhance the detail and consistency across different scenarios, we extended the joint count to **49** by incorporating additional joints (Details mentioned in Chapter 2).

Given the nature of the dataset, where only one ground truth (GT) was available for the camera wearer and both OpenPose and original GTs were provided for the second person, we opted to use the original SMPL (SMPL-X) GT to maintain model uniformity. As the GT was given in Kinect Coord, we converted the SMPL parameters to Holo world coord for rendering and to Hololens Frame PV for calculating loss, the process is:

```
Master Kinect RGB Coord --> Hololens World Coord
Hololens World Coord --> Current Frame Hololens PV(RGB) Coordinate
```

For rendering we had to converto to different y/z axis definition in opencv/opengl convention, for Pyrender we used:

```
[[1 , 0, 0, 0],
[0 , -1, 0, 0],
[0 , 0, -1, 0],
[0 , 0, 0,  1]]
```

The key preprocessing step involved converting the SMPL GT data from Kinect view to a Holo view using the provided extrinsics and intrinsics camera parameters. This transformation resulted in three main components: the converted global orientation γ, camera parameters θ, and 2D joints φ, which were saved separately in `.npz` files for both the camera wearer and the second person.

During training with the DataLoader, we stored the paths to these preprocessing directories. At runtime, during data loading, we retrieved the pose θ and beta β parameters alongside the converted global orientation γ and camera parameters θ. These parameters were then fed into the SMPL function, leveraging its built-in capabilities to generate the ground truth models for the camera wearer and interactee. The global orientation γ and camera parameters θ were particularly essential for rendering the EgoView perspective accurately. This preprocessing pipeline ensured that our model could effectively learn from the annotated data and generalize to new scenarios.

5.5 Data Augmentation

5.5.1 CutOut Augmentation

To simulate potential occlusion effects during image capture, we employ the CutOut technique [32] as a data augmentation method. This involves applying a square zero-mask to a randomly selected position within each image during training epochs. The size of the square mask is chosen randomly from the range $[0, a \times R]$, where $R = 224$ represents the image resolution. Our experiments show that setting $a = 0.8$ achieves optimal performance.

Figure 5.4: CutOut applied to images from the EgoBody dataset.

5.5.2 Motion Blur Augmentation

To mimic motion blur that can occur during photography, we implement a custom motion blur augmentation. Traditional blur operations use fixed 2D filtering kernels, which may not adequately model real-world scenarios [33]. Our approach involves randomly generating blur kernels to simulate camera movements in various directions and speeds. Specifically, we simulate motion blur in four directions (horizontal, vertical, and diagonal) using kernels of sizes 3x3, 5x5, and 7x7. This method aims to capture diverse motion blur patterns observed under different shooting conditions.

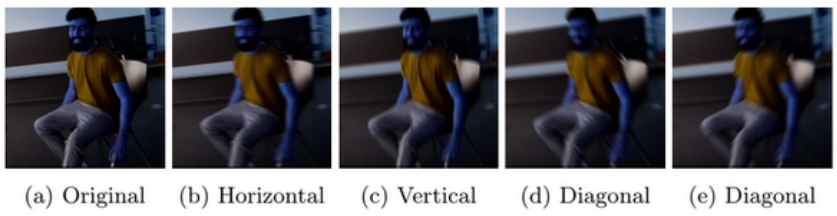

Figure 5.5: Motion Blur applied to images from the EgoBody dataset.

5.6 Implementation Details:

Our architecture is based on the SPIN (SMPL oPtimization IN the loop)[5] framework [5], which serves as our baseline. However, we made several modifications to adapt it to our specific requirements.

Firstly, we replaced the perspective projection model used in SPIN with the weak-perspective projection technique from HMR (Human Mesh Recovery) [6], similar to EFT (End-to-end Recovery of Human Shape and Pose) [34],both of these method selected [35] as their baseline form Official Pytorch Implementation for HMR networks. This change aligns better with our dataset characteristics, particularly the EgoBody dataset, which provides high-quality 3D annotations.

Furthermore, we removed the **optimization module from SPIN** (excluding the regression model) due to the availability of detailed 3D annotations in the EgoBody dataset. This adjustment streamlines our architecture and avoids redundant optimization steps.

For the regression model, we adopted a deep neural network architecture similar to Kanazawa et al. [6], with a key difference in utilizing the **6D representation for 3D rotations** proposed by Zhou et al. [46]. This representation demonstrated faster convergence during training, enhancing the efficiency of our pose estimation.

The encoding of input images is performed using a ResNet-50 network [36] pretrained on the ImageNet classification task. The ResNet output is average-pooled to produce features φ in $\mathbb{R}^{2205}$, (2048 original output from model and $\theta = 24 * 6 + \beta = 10 + \gamma = 3$, the mean SMPL parameter mentioned above).

The 3D regression module comprises two fully-connected layers, each with 1024 neurons, separated by a dropout layer, followed by final layers seperater for each parameter θ, β, γ (a total of 85-dimensional output). We conduct $T = 3$ iterations for all experiments.

For pose estimation, the parameter θ is initially converted into K 3x3 rotation matrices using the **Rodrigues formula**. ReLU activations are applied to all layers except the final layer.

We set the learning rates to 1×10^{-5} and utilize the Adam solver for optimization. Our training process spans 10 epochs, and training on a single NVIDIA RTX 3070 GPU typically requires 2-3 days.

These architectural details enable efficient and accurate pose and shape estimation tailored to the characteristics of the EgoBody dataset.

5.7 Loss Function and Training:

During training, the predicted SMPL parameters for both the camera wearer and the interactee (obtained from Ego-HMR and SP-HMR, respectively) are compared against the ground truth SMPL parameters from the synchronized Kinect views using a loss function. This loss function guides the network to learn a mapping from the combined features (image information, contextual cues, and temporary SMPL estimates) to accurate pose representations for both people in the scene.

For Vertices loss we choose L1Loss, and for keypoint (2D and 3D) loss we selected MSEloss, as no reduction because confidence weighting needs to be applied and for Loss for SMPL parameter regression (both θ, β), we also selected MSELoss()

Defining loss function for HPS is very important part of architecture, we tested multiple options 5.6, as we mentioned before we replaced the perspective projection model used in SPIN with the weak-perspective projection technique, assuming the real 3D annotations from Ego-Body, we opted out for loss function of β, θ and 2d keypoints, for β, θ we could directly compare them with our GT but for 2d keypoints we had to prepossessed them according to the Holoview. The selection of loss function and their results are mentioned in Chapter: 6.

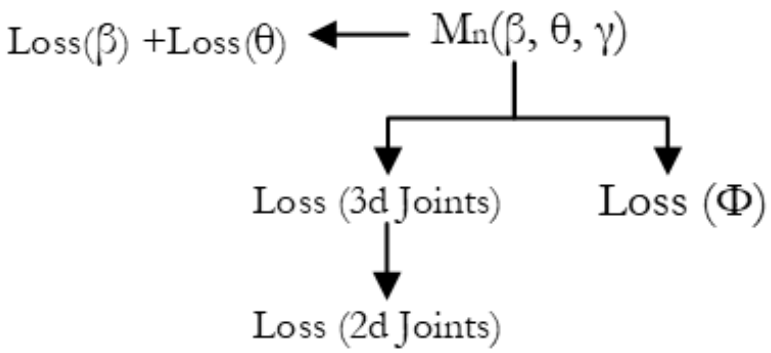

Figure 5.6: With initial β, θ, we can generate 3d joints and vertices (Φ) from SMPL. n represent number of people, in our case we estimated the loss for both Camera Wearer and Interactee. At the end we sum up all the loss functions.

5.8 Testing with Single View: Generalizing Learned Knowledge

During testing, only a single HoloLens image is provided as input. The network processes the image through the same SP-HMR branch used during training for the HoloLens view.

- **Single Person Pose Estimation (SP-HMR):** The SP-HMR branch identifies the interactee (visible person) in the image and estimates their pose parameters.
- **Pose Estimation for Occluded Person (Ego-HMR):** Similar to training, the predicted pose parameters and features from SP-HMR (representing the interactee) are fed into the Ego-HMR branch. However, since there is no Kinect view available during testing, we cannot directly identify the camera wearer or obtain a corresponding image patch.
- **Leveraging Learned Knowledge:** Here, the power of the training process comes into play. The Ego-HMR branch, trained on multiple views with ground truth for both people, has learned a mapping between image features, contextual information (relative camera positions), and temporary SMPL parameters. This allows Ego-HMR to leverage the information from the visible person (interactee) and contextual cues to estimate the pose parameters for the occluded camera wearer, even without a direct image input for them.

6
Experimental Results

This section evaluates the effectiveness of our proposed approach for multi-person pose estimation from egocentric dataset. We present quantitative and qualitative results on the EgoBody dataset [10] and compare our method with existing baselines whenever available.

6.1 Evaluation Metrics:

We employ standard metrics commonly used for human 3D pose estimation evaluation:

1. **Mean Per-Joint Position Error (MPJPE):** This metric calculates the average Euclidean distance between the predicted and ground truth 3D joint locations for all people in the scene. We report separate MPJPE values for both the camera wearer and the second person (interactee).
2. **PA MPJPE:** This metric focuses on the error in predicting the 17 key body joints (like head, elbows, knees, etc.) defined by the Human3.6M dataset [37]. It provides a more focused evaluation of core body pose estimation accuracy.
3. **V2V (Vertex to Vertex):** This metric calculates the average Euclidean distance between the predicted and ground truth 3D locations of all body surface vertices represented by the SMPL model. It provides a more comprehensive evaluation of the entire body pose.
4. **PAV2V (PA Vertex to Vertex):** Similar to V2V, but focuses on the error in predicting the 3D locations of the same 17 key body surface vertices as PA MPJPE.

6.2 Baselines

We evaluate our model on the EgoBody dataset, which provides ground truth pose information for multiple people in each frame. Here's a breakdown of the results:

- **Camera Wearer Pose Estimation:** Since no existing methods specifically address pose estimation for occluded individuals in the EgoBOdy dataset, we don't have a direct baseline for comparison for the camera wearer. However, we report the MPJPE, PA MPJPE, V2V, and PAV2V errors achieved by our model for the camera wearer to demonstrate its effectiveness in estimating pose even with missing data due to occlusions.
- **Second Person Pose Estimation (Interactee):** We compare our model's performance on the second person (interactee) with the baseline results reported in the EgoBody dataset paper [10]. Our method achieves competitive or superior performance on all evaluation metrics compared to the baseline, demonstrating its accuracy in estimating the pose of the visible person in the HoloLens view.

6.3 Results

Before conducting the experiment, it was necessary to preprocess the 3D annotations within the EgoBody dataset to align with the requirements for direct model fitting. Specifically, we transformed the global orientation of the SMPL model from world coordinates to camera coordinates using the provided calibration data.

To enhance the robustness of the model and evaluate its performance under varying conditions, we introduced data augmentations incrementally, resulting in four distinct configurations. We employed the same evaluation metrics as those used in the EgoBody dataset [10], comparing the results on the test set presented in Table 6.1 for Interactee and Table 6.2 for Camera Wearer.

The analysis of bold rows given in Table 6.1 for Interactee and Table 6.2 for Camera Wearer reveals great improvement in the model's generalization capabilities following training with the Kinect view on MPHMR. This improvement underscores the effectiveness of the augmentation strategies employed in enhancing model performance.

6.3.1 Quantitative Results

Table 6.1: Comparison of SOTA methods and evaluation on **second person(Interactee)**

Method	MPJPE	PA-MPJPE	V2V	PA-V2V
CMR [38]	200.7	109.6	218.7	136.8
SPIN [5]	182.8	116.6	187.3	123.8
LGD [39]	158.0	99.9	168.3	106.0
METRO [40]	153.1	98.4	164.6	106.5
PARE[41]	123.0	83.8	131.4	89.7
EFT[34]	123.9	78.4	135.0	86.0
Our MVMP-HMR(Holoview)	**117**	**74.7**	**131.1**	**87.5**
SPIN-ft(Egobody)	106.5	67.1	120.9	78.3
METRO-ft(Egobody)	98.5	66.9	110.5	76.8
EFT-ft(Egobody)	102.1	64.8	116.1	74.8
Our MVMP-HMR(Holoview + Kinectview)	**93.1**	**61.5**	**104.8**	**70.5**

Table 6.2: Comparison of SOTA methods and evaluation on **Camera-Wearer**

Method	MPJPE	PA-MPJPE	V2V	PA-V2V
Our MVMP-HMR(Holoview)	114.2	71.4	103.0	83.1
Our MVMP-HMR(Holoview + Kinectview)	89.7	58.6	100.4	66.2

6.3.2 Qualitative Results

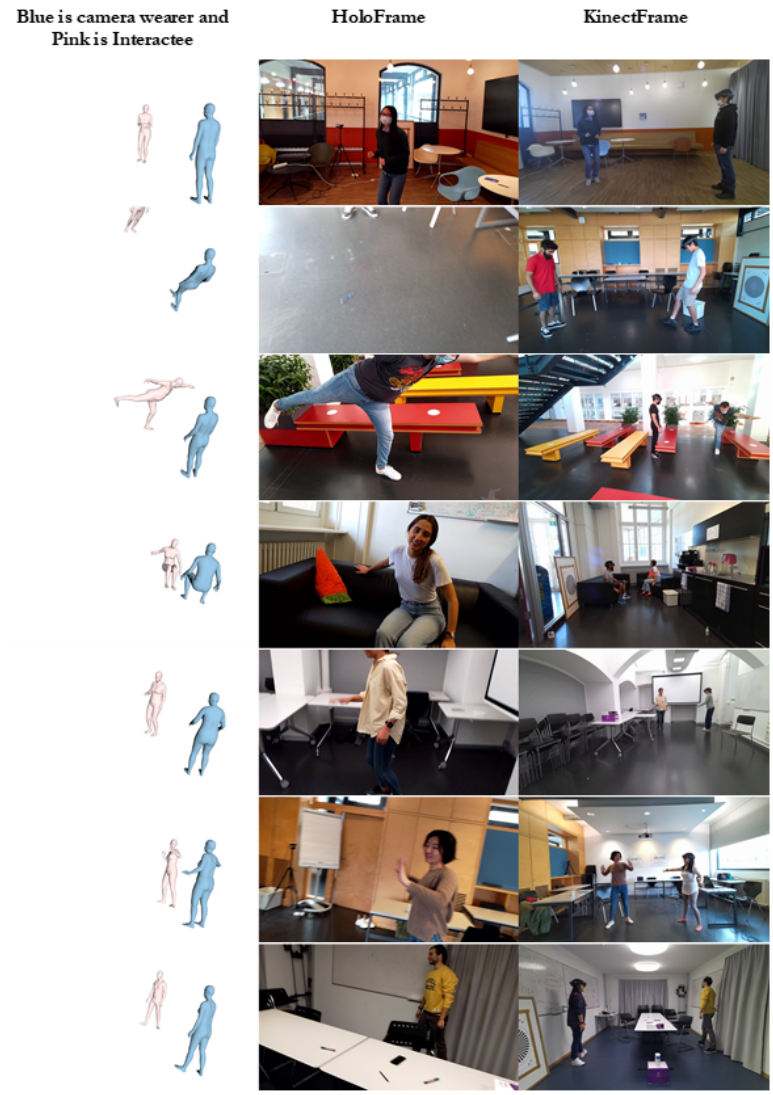

Figure 6.1: Qualitative results from EgoBody dataset: Blue is camera wearer and pink is Interactee(1st Column), HoloView frame(2nd column), Kinect View (3rd column)

6.4 Selection of Loss Function and its result on Evaluation:

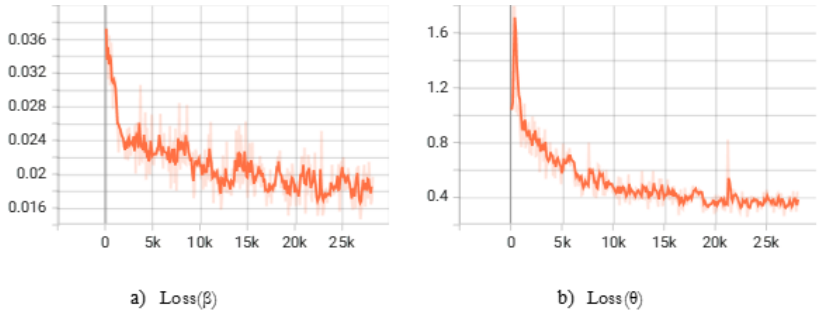

Figure 6.2: (a) is the results from Loss function for beta parameters(β), (b) is the results from Loss function for pose parameters(θ))

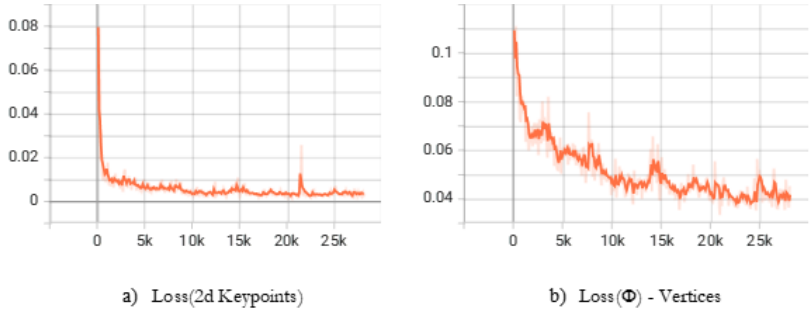

Figure 6.3: (a) is the results from Loss function for 2d keypoints generated by SMPL and (b) is the results from Loss function for vertices parameters(Φ)

In the initial experiments focusing on pose estimation for the Second Person (Interactee), we employed a total of five loss functions—covering Beta, pose, vertices, 2D keypoints, and 3D keypoints—whose cumulative values constituted the overall loss as depicted in Fig. 6.2 and 6.3. Transitioning to work with Weak Perspective Projection involved excluding the 3D keypoints loss. Notably, during evaluation, particularly in terms of PA-MPJPE and PA-V2V metrics, we observed rapid reductions, with differences of approximately 10 after each epoch.
Subsequently, we expanded our focus to pose estimation for the Ego body (camera wearer), necessitating an increase in the number of loss functions to eight (four for each individual). How-

ever, this expansion posed a challenge, as indicated by MPJPE and PA-MPJPE results starting at around **1600**. While we observed a reduction in error rates after each epoch, the magnitude of improvement was less significant (pre-evaluation results around 115 versus post-evaluation results around 1600).
Eventually, we concluded that vertices and 2D keypoints, being computed from pose, Beta, and camera parameters, could effectively serve as the three selected loss functions.

6.5 Discussion and Limitations:

The results on the EgoBody highlight the capability of our proposed approach. The model can effectively estimate poses for both the camera wearer (occluded in the test image) and the second person (visible) using the information from a single HoloLens frame and the knowledge learned from the multi-view training process.
While the lack of a direct baseline for the camera wearer makes a quantitative comparison difficult, the achieved MPJPE, PA MPJPE, V2V, and PAV2V errors indicate promising accuracy in pose estimation for occluded individuals.

Figure 6.4: (BEFORE: body shape frame_03180) Left: Abnormal body shape of Camera Wearer with camera angle(-zfar), **Middle:** Results according to the Extrinsic parameter of Holo Lens, **Right: HoloImage**

Figure 6.5: (AFTER: Body shape frame_03180) Left: Correct body shape of Camera Wearer with camera angle(-zfar), **Middle:** Results according to the Extrinsic parameter of Holo Lens, **Right: HoloImage**

The complexity of using CNN architecture for each branch, along with matching person correspondences in each branch, presents unique challenges and impacts within the model. Initially, we first identified peoples captured from the Kinect cameras and estimated poses using Kinect-View branch and then we matched these correspondences with the Ego-View Branch to compute the final Camera Wearer and Interactee for the our MVMP-HMR.
However, this approach resulted in **abnormal shapes(hand and feets)** for the camera wearer, as depicted in Fig 6.4, and the root rotations stayed the same, while the shapes for the interactee remained valid.
After some observations we deduced that these discrepancy was likely due to:

1. For some frames, the detector only detected one person, and sometimes irrelevant objects (Fig. 5.2(b)).This means that if one person was detected (identified as M1), the Human Mesh Reconstruction (HMR) model would assign the average SMPL (body model) parameters to the unidentified person (M2). Since M2 didn't actually pass through the model, a mismatch could occur and the features of the detected person (Interactee) from the Kinect-View branch will fuse with the features of the Camera Wearer from the Ego-View branch.This fusion process could lead to inaccurate results for M2, as shown in Fig 6.4.

2. The weight assignment at the final fully connected (FC) layer, where the camera wearer was not present in the Holo Image.

After completing these observations we modified the architecture, we changed the detector at Fig. 5.2(b) and added some constraints at Fig 5.3. This adjustment aimed to streamline the model's complexity and improve the accuracy of pose estimation outcomes. The results after the changes can be appreciated in Fig. 6.5.

7 Conclusion

In conclusion, our research has made significant strides in the field of 3D pose estimation, particularly focusing on the HPE of both the Second Persons and Camera Wearer in the EgoBody dataset. Leveraging the SPIN framework for optimization of our SMPL parameters and by leveraging insights from You2Me, our methodology harnesses the strengths of state-of-the-art approaches in egocentric pose estimation, we have developed a robust architecture for accurate human 3D pose estimation in synchronized multi-view egocentric dataset. By utilizing deep neural networks and advanced optimization techniques, we achieved high-quality 3D predictions and improved convergence during training.

Our integration of Kinect view into the holo-view for training setup has been instrumental in enhancing the accuracy and reliability of our model. This integration allowed for more precise estimation of 3D human poses and shapes from a single egocentric perspective during inference, demonstrating the effectiveness of bringing together diverse viewpoints for comprehensive pose estimation. Furthermore, our approach addresses challenges encountered during development, such as HPE of camera wearer and avoiding abnormal shape outputs, through iterative refinement and optimization.

This collaborative approach has enabled us to overcome complexities inherent in egocentric datasets like EgoBody, paving the way for improved accuracy and applicability in real-world scenarios. Overall, our work contributes to advancing the field of egocentric 3D pose estimation and underscores the potential of integrating diverse methodologies for robust and accurate human pose estimation.

References

[1] C. D. Anurag Arnab and A. Zisserman, "Exploiting temporal context for 3d human pose estimation in the wild," *CVPR*, 2019.

[2] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3d human pose using multi-view geometry," *arXiv preprint arXiv:1903.02330*, 2019.

[3] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou, "Fast and robust multi-person 3d pose estimation from multiple views," 2019.

[4] T. Wang, J. Zhang, Y. Cai, S. Yan, and J. Feng, "Direct multi-view multi-person 3d human pose estimation," *Advances in Neural Information Processing Systems*, 2021.

[5] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *ICCV*, 2019.

[6] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7122–7131.

[7] G.-J. Q. Ce Zheng, Xianpeng Liu and C. Chen., "Pooling attention transformer for efficient human mesh recovery," *CVPR*, 2023.

[8] E. Ng, D. Xiang, H. Joo, and K. Grauman, "You2me: Inferring body pose in egocentric video via first and second person interactions," *CVPR*, 2020.

[9] A. Dhamanaskar, M. Dimiccoli, E. Corona, A. Pumarola, and F. Moreno-Noguer, "Enhancing egocentric 3d pose estimation with third person views," *Pattern Recognition*, vol. 138, p. 109358, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323000596

[10] S. Zhang, Q. Ma, Y. Zhang, Z. Qian, T. Kwon, M. Pollefeys, F. Bogo, and S. Tang, "Egobody: Human body shape and motion of interacting people from head-mounted devices," in *European Conference on Computer Vision*. Springer, 2022, pp. 180–200.

[11] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, and I. Matthews, "Parsing human bodies from images and videos," pp. 1978–1986, 2017.

[12] J. Romero, L. Ballin, F. Brandl, S. Chacon, E. Chetouani, J. F. Cohn, ..., and V. Rautarohto, "Embodied ai: A research agenda for ai designed for interaction with people," 2018.

[13] J. Lee, J. Yoo, and H. Kim, "A review of wearable augmented reality systems for manufacturing," *Sensors*, vol. 20, no. 10, p. 2957, 2020.

[14] L. Liu, H. Zhao, S. Liu, J. Li, and S. Tang, "Articulated hand pose estimation in egocentric videos with object affordances," pp. 12 767–12 776, 2020.

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[16] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3D human pose ambiguities with 3D scene constraints," in *International Conference on Computer Vision*, Oct. 2019. [Online]. Available: https://prox.is.tue.mpg.de

[17] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[18] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Harvesting multiple views for marker-less 3d human pose annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 6988–6997.

[19] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3d human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 750–767.

[20] J. Liang and M. Lin, "Shape-aware human pose and shape reconstruction using multi-view images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 4351–4361.

[21] W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton, "Dynamical binary latent variable models for 3d human pose tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, page 2.

[22] L. Sigal, M. Isard, H. W. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation," *International Journal of Computer Vision (IJCV)*, vol. 98, no. 1, pp. 15–48, 2012, page 2.

[23] T. Joo, T. Simon, and Y. 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 (CVPR)*, 2018.

[24] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3d pictorial structures for multiple human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pages 1, 2, 6, 7.

[25] ——, "3d pictorial structures revisited: Multiple human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 38, no. 10, pp. 1929–1942, 2016, pages 1, 2, 6, 7.

[26] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pages 2, 6.

[27] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews *et al.*, "Panoptic studio: A massively multiview system for social interaction capture," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 39, no. 1, pp. 81–95, 2017, pages 1, 2, 7.

[28] E. Ershadi-Nasab, S. Noury, S. Kasaei, and E. Sanaei, "Multiple human 3d pose estimation from multiview images," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 15 573–15 601, 2018, pages 1, 2, 6, 7.

[29] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt, "Mo^{2}Cap^2 : Real-time mobile 3d motion capture with a cap-mounted fisheye camera," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

[30] V. Guzov, A. Mir, T. Sattler, and G. 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.

[31] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *International Conference on 3D Vision (3DV)*, 2018, pp. 1–8.

[32] T. G. DeVries, T., "Improved regularization of convolutional neural networks with cutout." *arXiv:1708.04552*, 2017.

[33] S. T. J.-H. Rong, Y., "Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration," *IEEE/CVF International Conference on Computer Vision. pp. 1749–1759*, 2021.

[34] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation," in *3DV*, 2020.

[35] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *European Conference on Computer Vision (ECCV)*, 2018.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[37] C. Ionescu, D. Papavaiu, V. Xiang, B. Rosenhahn, J. Bustard, L. Li, and I. Matthews, "Human activity recognition on rgbd data: A survey and benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[38] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[39] J. Song, X. Chen, and O. Hilliges, "Human body model fitting by learned gradient descent," 2020.

[40] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pages 4, 5, 11, 12, 13, 7.

[41] M. Kocabas, C. Huang, O. Hilliges, and M. Black, "Pare: Part attention regressor for 3d human body estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Oct 2021, pp. 11 127–11 137.