EGOCENTRIC HUMAN MOTION CAPTURE

JIAN WANG

Dissertation zur Erlangung des Grades des
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

Saarbrücken, 2024

Date of Colloquium	November 15, 2024
Dean of the Faculty	Prof. Dr. Roland Speicher
Chair of the Committee	Prof. Dr. Philipp Slusallek
Reviewers	Prof. Dr. Christian Theobalt
	Prof. Dr. Hans-Peter Seidel
	Prof. Dr. Gerard Pons-Moll
Academic Assistant	Dr. Lin Geng Foo

ABSTRACT

The human motion capture (mocap) technology has wide applications, especially in entertainment, sports analysis, and human-computer interactions. Among the motion capture techniques, egocentric motion capture provides a unique perspective from the individual's point of view. Being able to capture human motion in an unconstrained environment, egocentric motion capture is crucial for AR/VR applications.

This thesis focuses on the task of egocentric motion capture with a single, head-mounted, downward-facing fisheye camera. This setup can provide a broad field of view, which enables the capture of both body movements and interactions within the environment.

Despite the advantages of egocentric cameras, this setup suffers from several challenges, which are discussed in this thesis. These challenges include global motion estimation, self-occlusion, fisheye lens distortion, and the lack of large-scale training datasets. This thesis addresses these challenges by introducing new datasets and technical contributions:

To address the lack of large-scale training datasets, the thesis presents new datasets, including EgoPW, EgoGTA, and EgoWholeBody. These datasets cover a wide range of motions and environments, containing detailed annotations for human motion and scene geometry. By proposing new datasets, this thesis also reduces the gap between synthetic and real-world data. To capture global human motion, the thesis employs the SLAM method to obtain the global camera poses. The camera poses and the initial local human motion estimations are simultaneously optimized with the motion prior. The thesis also presents methods to overcome the issue of self-occlusion. These include leveraging temporal information, applying human motion priors, and incorporating scene geometry information. To mitigate the fisheye distortion issue, this thesis introduces FisheyeViT. It rectifies fisheye distortion with image patches and employs a Vision Transformer (ViT) network for feature extraction.

All of the methods in this thesis provide new solutions to some of the main challenges of egocentric motion capture with different technical and dataset contributions. These contributions enhance the capability to capture human motion under unconstrained scenarios, which offers new possibilities for applications in VR, AR, interactive gaming, and more.

ZUSAMMENFASSUNG

Technologien zur rechnergestützten Erfassung menschlicher Bewegungen (Mocap) finden Anwendung in unterschiedlichen Bereichen, beispielsweise in der Unterhaltungsbranche, der Sportanalyse oder der Mensch-Computer-Interaktion. Die egozentrische Bewegungserfassung sticht hierbei heraus und liefert dadurch einen entscheidenden Beitrag für AR und VR-Anwendungen, dass sie die menschliche Bewegung und Wahrnehmung aus Sicht des Trägers erfasst und somit die Nutzung in uneingeschränkten Umgebungen ermöglicht.

Diese Arbeit befasst sich mit der Aufgabe der egozentrischen Bewegungserfassung auf Basis einer einzelnen, am Kopf montierten, nach unten gerichteten Fischaugenkamera. Das breite Sichtfeld dieses Systems ermöglicht nicht nur die Erfassung von Körperbewegungen, sondern auch von Interaktionen in der Umgebung.

Trotz der Vorteile egozentrischer Kameras geht die Verwendung dieses Systems mit einigen Problemen einher, die in dieser Arbeit thematisiert werden. Die vorgestellten Datensätze und technischen Methoden bearbeiten unter anderem die Herausforderung der Schätzung der globalen Bewegung, die erschwerte Schätzung der Bewegung durch starke gegenseitige Überdeckungen verschiedener Körperteile, die Verzerrung der erfassten Bilder durch des Fischaugen-Objektivs und der Mangel an großen Trainingsdatensätzen.

Um den Mangel an großen Trainingsdatensätzen zu beheben, stellt die Arbeit die Datensätze EgoPW, EgoGTA und EgoWholeBody vor. Diese Datensätze decken ein breites Spektrum an Bewegungen und Umgebungen ab und enthalten detaillierte Annotationen für menschliche Bewegungen sowie die Geometrie der Szene. Durch die Einführung dieser Datensätze versucht diese Arbeit auch, die Unterschiede zwischen synthetischen und realen Daten zu reduzieren.

Zur Erfassung der globalen menschlichen Bewegung wird in dieser Arbeit die SLAM-Methode eingesetzt, um die globalen Kamerapositionen zu ermitteln. Die Kameraposen und initialen lokalen Bewegungsschätzungen des Menschen werden gemeinsam unter Betrachtung der Einhaltung wahrscheinlicher menschlicher Bewegungen optimiert.

Die erschwerten Bedingungen durch gegenseitige Abdeckung verschiedener Körperteile wird durch die Nutzung zeitlicher Informationen, die Integration der a-priori Verteilung menschlicher Bewegungen, sowie die

Einbeziehung von Informationen über die Geometrie der Szene adressiert.

Um das Problem der Fischaugenverzerrung zu reduzieren, wird in dieser Arbeit FisheyeViT vorgestellt. Es korrigiert die Verzerrungen der Fischaugenkamera und verwendet ein Vision Transformer (ViT)-Netzwerk zur Merkmalsextraktion.

Alle in dieser Arbeit vorgestellten Methoden bieten neue Lösungen für einige der größten Herausforderungen der egozentrischen Bewegungserfassung. Diese Beiträge erweiterten den Raum möglicher Szenarien und Umgebungen zur Erfassung menschlicher Bewegungen, was neue Möglichkeiten für Anwendungen in VR, AR, interaktiven Spielen und mehr bietet.

ACKNOWLEDGMENTS

I am deeply grateful for having spent many years as a Ph.D. candidate in the Graphics, Vision, and Video group, which evolved into the Visual Computing and Artificial Intelligence department during this journey. My Ph.D. career and this thesis would not have been possible without their guidance, encouragement, and assistance.

First and foremost, I would like to thank my supervisor, Prof. Christian Theobalt, for his unwavering support, insightful advice, and constructive feedback. His expertise and mentorship have been invaluable throughout my research.

I am immensely thankful to my colleagues and friends who have been part of this journey. The discussions and collaborations have made this experience enjoyable and rewarding. I would especially like to thank those who directly helped me with my Ph.D. projects: Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, Zhe Cao, and Thabo Beeler. Additionally, I extend my gratitude to those whose excellent projects I had the privilege to participate in: Hiroyasu Akada, Christen Millerdurai, Jianchun Chen, and Andrea Boscolo.

I also want to acknowledge the administrative staff who ensured everything ran smoothly: Sabine Budde, Ellen Fries, everyone at IST, and our administrators, Gereon Fox, and Pramod Ramesh Rao. I extend my appreciation to my colleagues, friends, and office mates at the VCAI group for creating a stimulating and supportive environment.

Special thanks to the members of my thesis committee, Professor Gerard Pons-Moll, Professor Hans-Peter Seidel, and Professor Christian Theobalt, for their valuable suggestions and input, which have greatly enhanced the quality of my thesis. I also want to thank Andrea Boscolo and Wanyue Zhang, who helped to proofread the thesis and gave valuable suggestions. I want to especially thank Gereon Fox and Olaf Dünkel for their help with the German abstract.

A heartfelt thanks to my family for their unconditional love and support. To my parents, Shancai Wang and Xiaoyu Jiang, for believing in me and encouraging me to pursue my dreams. To my best friends, Yunshu Jiang and Zhao Ma, for their constant encouragement, understanding, and timely help.

Thank you all for your incredible support and for being a part of this significant milestone in my life.

CONTENTS

1 Introduction 1
1.1 Motivation 1
1.2 Overview 4
1.3 Structure 6
1.4 Contributions 7
1.5 Publications and Preprints 8

2 Related Work 11
2.1 Egocentric Human Motion Capture 11
2.1.1 Inside-In Egocentric Motion Capture 11
2.1.2 Egocentric Motion Capture with Other Setups . . . 13
2.2 Dataset for Egocentric Human Motion Capture 16
2.2.1 Dataset with Down-facing Cameras 16
2.2.2 Dataset with other setups 18
2.3 Human Motion Capture with External Cameras/Sensors . 20
2.3.1 Monocular 3D Human Motion Capture 20
2.3.2 Weakly-Supervised 3D Human Pose Estimation . . 21
2.3.3 Whole-Body 3D Human Motion Capture 22
2.4 Human Motion Priors for Pose Estimation 23
2.5 Scene-aware Human Pose Estimation 24

3 Background 27
3.1 Fisheye Camera Model 27
3.2 Tracking Egocentric Camera 28

4 Estimating Egocentric 3D Human Pose in Global Space 29
4.1 Introduction 29
4.2 Method 31
4.2.1 Local Pose Estimation 32
4.2.2 Local Pose Optimization 33
4.2.3 Global Pose Estimation 35
4.3 Experiments 37
4.3.1 Datasets 37
4.3.2 Evaluation Metrics 38
4.3.3 Implementation Details 38
4.3.4 Comparison with State-of-the-art Results 39
4.3.5 Ablation Study 42
4.4 Limitations 46

4.5 Conclusion 46

5 EgoPW: Estimating Egocentric 3D Human Pose in the Wild 47
5.1 Introduction 47
5.2 Method 50
5.2.1 EgoPW Dataset 51
5.2.2 Optimization for Generating Pseudo Labels 51
5.2.3 Training Egocentric Pose Estimation Network . . . 54
5.3 Experiments 57
5.3.1 Datasets 57
5.3.2 Evaluation Metrics 58
5.3.3 Pseudo Label Generation 58
5.3.4 Comparisons on 3D Pose Estimation 60
5.3.5 Ablation Study 61
5.4 Limitations 63
5.5 Conclusion 63

6 Scene-aware Egocentric 3D Human Pose Estimation 65
6.1 Introduction 65
6.2 Method 68
6.2.1 Training Dataset 69
6.2.2 Scene Depth Estimator 70
6.2.3 Scene-aware Egocentric Pose Estimator 71
6.3 Experiments 73
6.3.1 Evaluation Datasets 73
6.3.2 Evaluation Metrics 73
6.3.3 Comparisons on 3D Pose Estimation 75
6.3.4 Ablation Study 77
6.4 Limitation 80
6.5 Conclusions 80

7 EgoWholeMocap: Egocentric Whole-Body Motion Capture 83
7.1 Introduction 83
7.2 Method 86
7.2.1 Single Image Based Egocentric Pose Estimation . . 87
7.2.2 Diffusion-Based Motion Refinement 91
7.3 EgoWholeBody Dataset 92
7.4 Experiments 93
7.4.1 Datasets and Evaluation Metrics 93
7.4.2 Comparisons on Whole-Body Pose Estimation . . . 95
7.4.3 Ablation Study 96
7.5 Limitation 98
7.6 Conclusions 99

8 Conclusion 101
8.1 Insights 101
8.1.1 Motion Prior 101
8.1.2 Egocentric Dataset 102
8.1.3 Scene-aware Egocentric Motion Capture 102
8.2 Challenges and Future Directions 103
8.2.1 Egocentric Human Motion Capture 103
8.2.2 Egocentric Motion Capture + Large Language Models 105
8.2.3 Egocentric Photorealistic Avatar 105
8.2.4 Egocentric Motion Capture + Robotics 106
8.2.5 Egocentric Motion Capture + Human-Object/Scene Interactions 106
8.3 Final Conclusion 107

A Appendix for Chapter 4 109
A.1 Comparisons on Different Types of Motions 109
A.2 The Structure of RNN-based VAEs 109

B Appendix for Chapter 5 113
B.1 Quantitative Results on Different Motions 113
B.2 Qualitative Results 113
B.3 Details and Comparisons of EgoPW dataset 113
B.4 Network Architecture 115
B.4.1 Pose Estimation Network 116
B.4.2 Domain Classifier 118
B.4.3 Egocentric-external View Classifier 119
B.5 Energy Function 119

C Appendix for Chapter 6 121
C.1 Datasets 121
C.1.1 SceneEgo Test Dataset 121
C.1.2 EgoGTA Dataset 122
C.1.3 EgoPW-Scene Dataset 123
C.2 Implementation Details 123
C.2.1 Scene Depth Estimator 123
C.2.2 V2V Network 124
C.3 Evaluation Metrics 125
C.4 Ablation Study on test Dataset in Chapter 4 125

D Appendix for Chapter 7 127
D.1 Full Comparison with Existing Egocentric Pose Estimation Methods 127
D.2 Implementation Details 127

D.2.1 FisheyeViT and Pose Regressor with Pixel-Aligned 3D Heatmap 127
D.2.2 Hand Detection Network 130
D.2.3 Hand Pose Estimation Network 131
D.2.4 Diffusion-Based Motion Refinement 131
D.3 Synthetic Dataset Comparisons 132
D.4 Details of Evaluation Metrics 132
D.5 Details of Evaluation Datasets 134
D.6 The Standard Deviation of Refinement Method 134
D.7 Different Parameters in Weight Function 135
D.8 More Visualization Results 136
D.9 Comparision with networks for panorama images 136
D.10 Replacing the Pixel-Aligned 3D Heatmap to MLP 138
D.11 Compare with Gaussian Smooth 139

Bibliography 141

LIST OF FIGURES

Figure 1.1 Different human motion capture methods, including (a) marker-based mocap method (VICON), (b) markerless mocap method (Captury), and (c) single-view mocap method (Mehta et al., 2017b). 2

Figure 1.2 Different setups for egocentric human motion capture, including (a) IMU-based setup (Yi et al., 2022b), (b) third-person setup (Zhang et al., 2022), (c) inside-out vision setup (Hwang et al., 2020), and (d) inside-in vision setup. 3

Figure 4.1 Given challenging egocentric videos, the proposed method produces realistic and accurate 3D global pose sequences. 30

Figure 4.2 Overview of the method. This method takes an egocentric video as input and processes it in segments. For each segment consisting of a fixed number of consecutive frames, this method first applies an egocentric pose estimation method to obtain initial 3D local poses and 2D heatmaps which are then fed into the local pose optimization framework to get optimized local poses. Next, combined with the camera poses estimated from ORB-SLAM2, the optimized 3D local poses are transformed from the local egocentric camera space to the world coordinate space and then optimized via the global pose optimization to produce the final global poses. 32

Figure 4.3 Interpolation in the latent space. The leftmost and rightmost pose sequences (waving hands and jumping) are reconstructed from two randomly sampled latent vectors, and intermediate pose sequences are reconstructed from linear interpolation between the left and right latent vectors. . . 36

Figure 4.4 The global pose with/without global pose optimizer. Here, we zoom onto the left foot for better comparison. 37

Figure 4.5 Qualitative comparison on the accuracy of a single pose. From left to right: input image, Mo^{2}Cap^2 result projected with SLAM (green), smoothed Mo^{2}Cap^2 result projected with SLAM (green), and proposed method's result (green) overlaid on ground truth (red). Note that in order to better show the result, the estimated pose is rigidly aligned to the ground truth. 40

Figure 4.6 Global pose estimation results from a third-view camera. Top row: the input egocentric images, bottom row: the estimated 3D pose projected on an external camera. 42

Figure 4.7 Comparison between the proposed method with and without motion prior. From left to right: input image, Mo^{2}Cap^2 + SLAM (green), the result without motion prior (green) and the one with motion prior (the result of proposed method) (green) overlaid on the ground truth (red). 44

Figure 4.8 $E(\mathcal{P}_{seq})$-iteration curve of different VAEs. The proposed method gives the lowest error while keeping stable during optimization. 45

Figure 4.9 Comparison between heatmap reprojection error and conventional reprojection error. The result of heatmap reprojection error is in a green skeleton, and the result of conventional reprojection error is in a blue skeleton. 46

Figure 5.1 Compared with Mo^{2}Cap^2, the proposed method gets a more accurate egocentric pose from a single in-the-wild image, especially when the body parts are occluded. Note that the external images are only used for visualization, not the inputs to the method. 48

Figure 5.2 Overview of our method. The new EgoPW dataset (Sec. 5.2.1) is firstly collected, where pseudo labels are generated by a multi-view based optimization method (Sec. 5.2.2). Then, the proposed framework (Sec. 5.2.3) is trained, where the network is simultaneously trained with EgoPW datasets and synthetic data from Mo^{2}Cap^2. The egocentric network is further enforced to learn a better feature representation from the external view (Sec. 5.2.3.2) and finally bridge the gap between synthetic and real data with a domain classifier (Sec. 5.2.3.1). . 50

Figure 5.3 The pseudo label generation method combines the information from both the egocentric view and external view, therefore leading to more accurate pseudo labels (c). Only with the egocentric camera, the feet cannot be observed and well-tracked (b). Only with the external camera, the hands are occluded and result in the wrong result on the hand part (d). 54

Figure 5.4 The visualization of features with (b) or without (c) the adversarial supervision from external features. By supervising the training of the egocentric network with the feature representation from an external view, the egocentric network is able to focus on extracting the semantic features of the human body. 57

Figure 5.5 Qualitative comparison between the proposed method and the state-of-the-art methods. From left to right: input image, Mo^{2}Cap^2 result, xR-egopose result, the result of the proposed method, and external image. The ground truth pose is shown in red. Note that the external images are not used during inference. The input images on the top part are from the test dataset of Chapter 4, while the images on the bottom part are from the EgoPW test sequences. 59

Figure 5.6 The results of the proposed method with (c) and without external view (b). The network cannot predict accurate poses for the occluded cases without external view supervision. The external view is only for visualization and not used for predicting the pose. 62

Figure 6.1 Previous egocentric pose estimation methods like EgoPW (Chapter 5) predict body poses that may suffer from body floating issues (the first row) or body-environment penetration issues (the second row). This method predicts accurate and plausible poses complying with the scene constraints. The red skeletons are the ground truth poses and the green skeletons are the predicted poses. 66

Figure 6.2 Overview of the proposed method. First, the synthetic training dataset EgoGTA and the in-the-wild training dataset EgoPW-Scene are rendered. Both datasets contain egocentric depth maps for subsequent training process (Sec. 6.2.1). Next, an egocentric scene depth estimator is trained to predict a depth map without the human body and a depth inpainting network (Sec. 6.2.2). Finally, the 2D body pose features and scene depth map are combined into a common voxel space. The 3D body pose heatmaps are regressed from the voxel space with a V2V network and the final pose prediction is obtained with soft-argmax (Sec. 6.2.3). . 68

Figure 6.3 Qualitative comparison between the proposed method and the state-of-the-art egocentric pose estimation methods. From left to right: input image, Mo^{2}Cap^2 result, xR-egopose result, EgoPW result, and the result of the proposed method. The ground truth pose is shown in red. The input images from the first three rows are from the SceneEgo test dataset, while those in the last three rows come from the EgoPW in-the-wild test sequences (without ground-truth poses). This figure also shows the gt scene geometry of the in-the-studio data and scene geometry obtained by the SFM method for the in-the-wild data. 74

Figure 6.4 Predicted pose with different scene depth map input. The proposed network can generate different poses under different depth inputs and further disambiguate body poses under scene constraints. 77

Figure 6.5 The qualitative depth estimation results with or without the inpainting network. The depth map estimated without an inpainting network shows artifacts in the human body region (see the red box). 80
Figure 7.1 From an image sequence captured by a single head-mounted fisheye camera, the proposed method can predict accurate and temporally coherent whole-body motion, including the human body and hand poses. The SMPL-X parameters are obtained using inverse kinematics. 84
Figure 7.2 Overview of the whole-body motion capture pipeline. First, FisheyeViT is used to undistort the input image and generate image feature tokens (7.2.1.1). Next, a convolutional network is used to convert the image features to a pixel-aligned 3D heatmap and use soft-argmax and fisheye camera undistortion function to obtain the 3D body joins positions and uncertainty (7.2.1.2). The hand location is further detected, and the 3D hand poses are further regressed from the input image (7.2.1.3). Finally, the estimated hand motion and human body motion are combined and the uncertainty-aware diffusion model is applied to refine the estimated whole-body motion (7.2.2). 86
Figure 7.3 The detailed illustration of FisheyeViT (Sec. 7.2.1.1). 88
Figure 7.4 Qualitative comparison on human body pose estimations between the proposed method and the state-of-the-art egocentric pose estimation methods on in-the-studio (two rows from the top) and in-the-wild scenes (two rows from the bottom). The red skeleton is the ground truth while the green skeleton is the predicted pose. The proposed methods predict more accurate body poses compared with EgoPW (Chaper 5) and SceneEgo (Chaper 6). 93

Figure 7.5 Qualitative comparison on human hand pose estimations between the proposed method and the state-of-the-art third-view pose estimation methods. The single-view and refined hand poses from the proposed method are more accurate than the poses from Hand4Whole (Moon et al., 2022) method. The red skeleton is the ground truth while the green skeleton is the predicted pose. . 94

Figure B.1 Qualitative comparison between the proposed method and the state-of-the-art methods on the test images of the EgoPW dataset. From left to right: input image, Mo^{2}Cap^2 result, xR-egopose result, the result of the proposed method, and external image. Note that the external images are only for visualization and they are not used for predicting the pose. 116

Figure B.2 Qualitative comparison between the proposed method and the state-of-the-art methods on the test images of Mo^{2}Cap^2 work. From left to right: input image, Mo^{2}Cap^2 result, xR-egopose result, and the result of the proposed method. 117

Figure C.1 Visualization of the data collection process for the SceneEgo test dataset. The pose of the head-mounted calibration board, rigidly attached to the egocentric camera, is detected from multiple views in the studio. 122

Figure C.2 Example of the EgoGTA dataset. 122

Figure C.3 Example of the EgoPW-Scene dataset. 123

Figure D.1 Examples of the synthetic dataset EgoWholeMocap. The upper row shows the data rendered with Renderpeople models (*RenderPeople* n.d.), and the lower row shows the data rendered with SMPL-X models (Pavlakos et al., 2019). 133

Figure D.2 The weight function with different hyper-parameters k. The x-axis is the diffusion time step t and the y-axis is the weight $\mathbf{w}$. 136

Figure D.3 Qualitative comparison on human body pose estimations between the proposed method and the state-of-the-art SceneEgo (Chapter 6) method. The red skeleton is the ground truth while the green skeleton is the predicted pose. The proposed methods predict more accurate body poses. . . . 137

Figure D.4 Qualitative comparison on hand pose estimation results. The proposed single-view and refined hand poses are more accurate than the poses from the Hand4Whole (Moon et al., 2022) method. The red skeleton is the ground truth while the green skeleton is the predicted pose. 138

LIST OF TABLES

Table 2.1 Datasets for Egocentric 3D Pose Estimation. The upper part shows the dataset collected with a monocular down-facing camera. The lower part shows the dataset collected with stereo down-facing cameras. 17
Table 4.1 The experimental results on Mo^{2}Cap^2 test dataset (Xu et al., 2019) and the new test dataset. Mo^{2}Cap^2 (or xR-egopose) + Proposed is the result of the proposed method based on the predictions of Mo^{2}Cap^2 (or xR-egopose). The proposed method outperforms previous state-of-the-art Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019) in all of the three metrics. 41
Table 4.2 The quantitative results of ablation study. 43
Table 5.1 The accuracy of pseudo labels on test dataset in Chapter 4. Utilizing both egocentric and external views, the body poses from the proposed optimization method (Sec. 5.2.2) are more accurate and can serve as better pseudo labels. 58
Table 5.2 Performance of the egocentric pose estimation network (Sec. 5.2.3) on Wang *et al.*'s test dataset and Mo^{2}Cap^2 test dataset (Xu et al., 2019). This method outperforms the state-of-the-art methods, Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019), on both metrics. 60
Table 5.3 The quantitative results of ablation study. 61

Table 6.1 Performance of the proposed method on SceneEgo test dataset, test dataset in Chapter 4 and Mo^{2}Cap^2 test dataset (Xu et al., 2019). The proposed method outperforms the state-of-the-art methods EgoPW, Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019). 75

Table 6.2 Comparisons of physical plausibility on the SceneEgo test dataset. 76

Table 6.3 Results from the proposed method compared to different baselines. 78

Table 6.4 The quantitative results of depth estimator. The Abs-Rel and RMSE are evaluated on the SceneEgo test dataset following Hu et al. (2019). 80

Table 7.1 Egocentric body pose accuracy of the proposed method on SceneEgo test datasets and EgoWholeBody test dataset. The proposed method outperforms all previous state-of-the-art methods. * denotes the method trained with the datasets in Sec. 7.4.1. 95

Table 7.2 Egocentric hand pose accuracy of the proposed method. The proposed method outperforms the Hand4Whole (Moon et al., 2022) on both datasets. 96

Table 7.3 Ablation Study on SceneEgo test dataset (Chaper 6). † denotes the temporal-based method. 97

Table A.1 The BA-MPJPE of different types of motions on the indoor sequence of Mo2Cap2 dataset (Xu et al., 2019). When based on the local poses estimated by Mo^{2}Cap^2, the proposed approach improves the Mo^{2}Cap^2 (Xu et al., 2019) results by 13.8% (8.5 mm).109

Table A.2 The BA-MPJPE of different types of motions on the indoor sequence of Mo2Cap2 dataset (Xu et al., 2019). When based on the local poses estimated by xR-egopose (Tomè et al., 2019), the proposed method improves the xR-egopose results by 8.9% (4.9 mm). 110

Table A.3 The BA-MPJPE of different types of motions on the test set. When based on Mo^{2}Cap^2 (Xu et al., 2019), the proposed approach outperforms Mo^{2}Cap^2 results by 16.6% (12.4 mm). 111

Table A.4 The BA-MPJPE of different types of motions on the test set. When based on xR-egopose (Tomè et al., 2019), the proposed method outperforms xR-egopose results by 26.2% (22.9 mm). 111

Table B.1 The BA-MPJPE of different types of motions on the test set of Chapter 4. The proposed approach outperforms Mo^{2}Cap^2 results by 9.59 mm and outperforms xR-egopose results by 22.33 mm. . . 114

Table B.2 The BA-MPJPE of different types of motions on the indoor and outdoor sequence of Mo^{2}Cap^2 dataset (Xu et al., 2019). In the indoor sequence, the proposed method improves the Mo^{2}Cap^2 results by 6.62 mm and xR-egopose results by 0.65 mm; In the outdoor sequence, the proposed method improves the Mo^{2}Cap^2 results by 6.09 mm and xR-egopose results by 3.75 mm. 115

Table B.3 Comparison between the EgoPW dataset and publicly available 3D pose estimation datasets. 118

Table C.1 Results from the proposed method compared to different baselines. 125

Table D.1 Performance of the proposed method on SceneEgo test datasets (Chapter 6). The proposed method outperforms all previous state-of-the-art methods. $*$ denotes the method trained with the datasets in Sec. 7.4.1. † denotes the temporal-based methods. 128

Table D.2 Performance of the proposed method on GlobalEgoMocap test dataset (Chapter 4). The proposed method outperforms all previous state-of-the-art methods. $*$ denotes the method trained with the datasets in Sec. 7.4.1. † denotes the temporal-based methods. 129

Table D.3 Performance of the proposed method on Mo^{2}Cap^2 test dataset (Xu et al., 2019). The proposed method outperforms all previous state-of-the-art methods. $*$ denotes the method trained with the datasets in Sec. 7.4.1. † denotes the temporal-based methods. 129

Table D.4 Performance of the proposed method on the EgoWholeBody test datasets. This method outperforms all previous state-of-the-art methods. $*$ denotes the method trained with the datasets in Sec. 7.4.1. † denotes the temporal-based methods. 130

Table D.5 Comparison between different training datasets for egocentric body pose estimation. 133
Table D.6 The mean and standard deviations of the proposed refinement method. “SceneEgo-Body” and “SceneEgo-Hands” show the body and hand results on the SceneEgo dataset. “GlobalEgoMocap” and “Mo^{2}Cap^2” shows the human body results on the GlobalEgoMocap and Mo^{2}Cap^2 datasets. . . . 135
Table D.7 Comparison with different k values. 136
Table D.8 Comparison with Spherenet and Panoformer. . . 138

1 INTRODUCTION

1.1 MOTIVATION

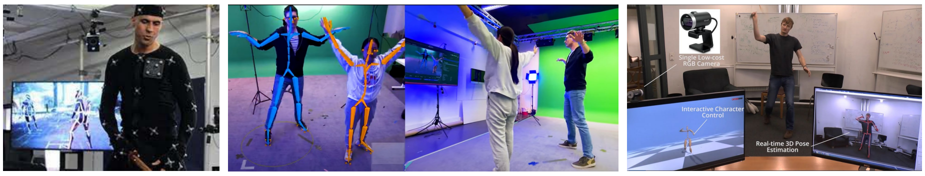
The human motion capture (mocap) systems have been extensively researched in recent years, enabling the capability to analyze, understand, and leverage human motion. These systems can facilitate various applications including filmmaking, gaming, sports analysis, and human-computer interactions.

One of the traditional and widely adopted motion capture methods is marker-based motion capture (Park and Hodgins, 2006). This method places reflective markers on anatomical landmarks of the human body and leverages high-speed cameras to track the movements of the markers with the triangulation method. The human motion can be further recovered from the location of landmarks in high precision. This method can provide highly accurate human motion captures while requiring the placement of markers on the body or wearing a special suit. This will change the appearance of the human body and sometimes can be even impractical.

Contrasting with marker-based systems, markerless multi-view motion capture (Liu et al., 2011) tracks and reconstructs the human movement without the need for physical markers. Multiple synchronized cameras capture the person from different viewpoints, and multi-view mocap algorithms are applied to create a 3D representation of the motion. Since no markers are needed to be placed on the human body, this method offers greater flexibility and can be applied in applications, such as human performance capture and photorealistic telepresence.

Recent advancements in deep learning have spurred the development of single-view motion capture systems (Mehta et al., 2020, 2017b). These methods usually leverage neural networks trained on large-scale datasets to predict 3D poses directly from 2D images or videos. Since this method can operate with a single camera, it is portable and more accessible than the aforementioned methods. This approach shows outstanding advantages in applications requiring lightweight setup and real-time motion capture, such as virtual reality applications and human-computer interactions.

Though the aforementioned methods show compelling results in the motion capture task, they still struggle with a common problem – they


Single Low-cost RGB Camera
Interactive Character Control
Real-time 3D Pose Estimation

(a) marker-based mocap (b) marker-less mocap (c) single-view mocap

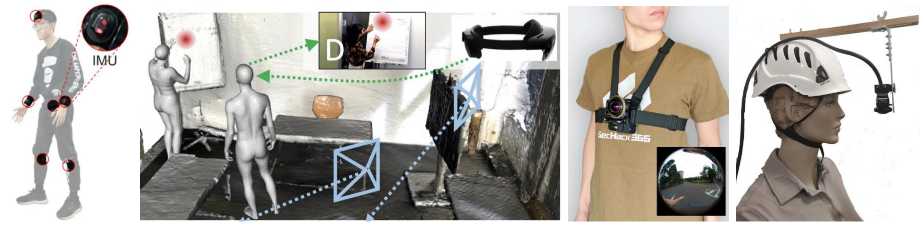
Figure 1.1: Different human motion capture methods, including (a) marker-based mocap method (VICON), (b) markerless mocap method (Captury), and (c) single-view mocap method (Mehta et al., 2017b).

can not capture human movement in large spaces since the current equipment is designed for constrained areas, like studio rooms. On the contrary, egocentric motion capture solves this by capturing movement from an individual's point of view. With wearable sensors, human motion can be captured when people roam around in a large space, building the foundation for immersive experiences in various applications.

Recent studies have proposed a variety of egocentric setups that utilize diverse sensor configurations for the egocentric motion capture. As illustrated in Fig. 1.2, these setups can be roughly classified into several categories: the Inertial Measurement Unit (IMU) setup, third-person egocentric setup, inside-out vision setup, inside-in vision setup, and hybrid approaches combining the setups mentioned above.

The Inertial Measurement Unit (IMU) setup (Guzov et al., 2021; Jiang et al., 2022a; Mollyn et al., 2023; Yi et al., 2022b) uses the IMU sensors consisting of accelerometers, gyroscopes, and magnetometers. These sensors provide data on the acceleration, angular rate, and orientation of the device. In this setup, IMUs are placed on the human body to detect limb acceleration and orientation. The data from IMUs are then processed to reconstruct the full-body motion. With a similar idea, several methods (Du et al., 2023; Jiang et al., 2022a; Winkler et al., 2022) also leverage the head and hand tracking systems on VR/AR headsets to estimate human motion. However, all of these setups need to place multiple sensors on the human body, which is inconvenient. The natural human motion can also be altered. Furthermore, the IMU-based setup suffers from the drifting issue and it does not provide information about the spatial relationship between the person and the surrounding environment.

The third-person setup (Khirodkar et al., 2023b; Zhang et al., 2023a, 2022) refers to motion capture techniques that involve a third person wearing a camera observing the motion capture subject. However, having a third person to hold the camera can be inconvenient and sometimes impractical, especially when capturing human motion through a long



(a) IMU-based (b) third-person egocentric (c) inside-out (d) inside-in

Figure 1.2: Different setups for egocentric human motion capture, including (a) IMU-based setup (Yi et al., 2022b), (b) third-person setup (Zhang et al., 2022), (c) inside-out vision setup (Hwang et al., 2020), and (d) inside-in vision setup.

sequence or in a small space. This setup is also cumbersome, which is hard to implement in a confined space.

The inside-out vision setup (Hwang et al., 2020; Li et al., 2023; Luo et al., 2021; Yuan and Kitani, 2019) captures human motion by employing wearable cameras or sensors looking front, towards the environment. This approach is commonly used in virtual reality (VR) and augmented reality (AR) systems since cameras in the headset can capture the user's surrounding environment. It's particularly effective for understanding the individual's locomotion within the environment. However, the outward-facing cameras can not directly see the movements of human body. Instead, the algorithms estimate the motion based on how the surroundings change. This can often lead to wrong motion estimations, especially for complicated movements. This limitation significantly impacts the ability to precisely capture a person's motion, which is crucial for numerous practical applications such as human telepresence and human-robot interaction.

In contrast to the inside-out setup, the inside-in vision setup (Tomè et al., 2019; Wang et al., 2024, 2022, 2021, 2023; Xu et al., 2019) mount the cameras or sensors on the human body and direct the cameras toward the person itself. By focusing inward, this inside-in setup has the potential to achieve higher accuracy since this setup can obtain high-resolution images of human motion. This setup also enables more applications, including human performance capture and human telepresence, since the human body can be directly observed.

Considering the strengths of the inside-in vision method, especially its ability to precisely capture detailed movements, this thesis adopts the inside-in setup. Following recent works Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019), this thesis captures egocentric human motion by mounting a single downward-facing fisheye camera on the head. This setup is lightweight and can be easily integrated with modern VR/AR headsets. With the wide field-of-view (FOV) provided by the fisheye lens, most parts of the human body and a large portion of the surrounding scene can be captured, enabling the awareness of the scene and enhancing the accuracy of motion capture methods in this thesis. While this setup has many advantages, it still poses several challenges where new solutions need to be found: the lack of datasets, the self-occlusion, and the fisheye distortion. This thesis aims to tackle these issues and achieve accurate and reliable whole-body and global egocentric 3D motion capture, aiming to contribute to the functionality and application of body-mounted devices.

1.2 OVERVIEW

This thesis proposes solutions that advance the research boundary of egocentric human motion capture that captures 3D human motion from a single head-mounted, downward-facing fisheye camera, as illustrated under the label (d) in Fig. 1.2. Building on the down-facing egocentric fisheye camera setup, as discussed above, several open challenges exist. This section looks into these challenges and demonstrates how the proposed solutions in this thesis can effectively address these challenges, and further enhance the performance of egocentric human motion capture.

LACK OF DATASETS. A significant challenge is the lack of datasets specifically tailored for training and validating models on egocentric motion capture with the egocentric capture setup. Most existing datasets are designed for outside-in camera setups and may not represent the unique perspectives captured by a fisheye lens mounted on the head. Some research works, such as Mo^{2}Cap^2 (Xu et al., 2019) and *x*R-egopose (Tomè et al., 2019) have introduced datasets for egocentric motion capture. However, the existing training datasets are predominantly synthetic, which results in several limitations. Firstly, these datasets lack diversity in human motion and environments. Moreover, the synthetic to real-world domain gap in these datasets presents significant challenges to real-world applications. Furthermore, insufficient annotations for surrounding environments and whole-body motions constrain the potential usage of these datasets in various applications.

To address these issues, this thesis introduces several datasets as well as new solutions: Chapter 5 presents an in-the-wild dataset named EgoPW, annotated with weak supervision from the external view. It also proposes an adversarial domain adaptation method to bridge the synthetic-to-real-world domain gap and the egocentric-view-to-external-view domain gap. Chapter 6 introduces a new synthetic dataset, EgoGTA, which includes synthetic egocentric views and precise annotations of scene geometry. Chapter 6 also proposes EgoPW-Scene with pseudo-ground truth annotations for the scene geometry in the EgoPW dataset. Chapter 7 introduces a high-quality synthetic dataset that encompasses a wide variety of whole-body human motions, including detailed movements of the human body and hands. These datasets collectively aim to enrich the field of egocentric motion capture by providing diverse and large-scale resources. By tackling the limitations of current datasets, this thesis pave the way for robust and practical egocentric motion capture methods.

GLOBAL MOTION. Another challenge is to estimate global human motion from the egocentric camera. Prior works, such as Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019) only estimate the local 3D body pose in egocentric camera space, while not being able to obtain the body pose with the global position and orientation in the world coordinate system.

To address this challenge, in Chapter 4, this thesis first introduces the Simultaneous Localization and Mapping (SLAM) method to obtain egocentric camera poses in the global coordinate system. The estimated human motion in the local space is projected into the global coordinate system to get the global motion with the camera poses. The method further refines the global motion using a learned global motion prior. The ability to capture global motion builds the foundation for many applications, such as human-scene interaction and navigation.

SELF-OCCLUSION. Since the camera is mounted on the head, a considerable portion of the lower body is occluded by the upper body. This self-occlusion issue makes capturing complete and accurate data for the entire body difficult.

This thesis introduces two strategies to mitigate this ill-posed problem. The first involves utilizing temporal information and a human motion prior, and the second employs scene geometry information to accurately position the lower body.

Following the first strategy, in Chapter 4, the thesis describes the learning of a Variational Autoencoder (VAE) based motion prior and introduces a heatmap-based 2D reprojection loss to optimize motion within the VAE latent space. In Chapter 7, a diffusion-based whole-body motion prior is learned, and an uncertainty-aware guided diffusion denoising process is implemented to refine the initial prediction of whole-body human motion.

Utilizing a motion prior in egocentric pose estimation enables a context-aware understanding of human movement, even under frequent self-occlusions which are typical for the egocentric setting.

Following the second strategy, in Chapter 6, the thesis proposes a scene-aware pose estimation network that projects the 2D image features and estimated depth map of the scene into a voxel space and regresses the 3D pose with a Voxel-to-Voxel (V2V) network. In this way, this method can learn the relative position and possible interactions between the human body joints and the environment. The ambiguity caused by self-occlusion can thus be reduced to give a better result.

FISHEYE DISTORTION. Fisheye lenses are known for their wide-angle view, which is beneficial for capturing a broad area. However, fisheye images are strongly distorted which challenges many deep neural network based estimation approaches. Simply undistorting the entire fisheye image is impractical due to the fisheye lens's large field of view (FOV). To tackle this issue, Chapter 7 propose FisheyeViT. This method initially divides the image into smaller patches, each aligned with a specific field of view (FOV) range. This method then performs patch-level undistortion, effectively rectifying the fisheye distortion while maintaining alignment with the vision transformer architecture.

In conclusion, this thesis proposes a set of innovations for the advancement of egocentric human motion capture. It addresses critical challenges including dataset limitations, self-occlusion, global motion estimation, and fisheye distortion by proposing new datasets and novel methods. These advancements enhance the accuracy, reliability, and applicability of egocentric motion capture, promising significant contributions to virtual reality, augmented reality, and beyond.

1.3 STRUCTURE

This thesis is structured as follows.

- Chapter 1 motivates the research task of egocentric human motion capture from down-facing cameras. Furthermore, it describes the structure and contributions of this thesis.
- Chapter 2 discusses the wider literature on capturing human motion from egocentric cameras mounted on the human body.
- Chapter 3 introduces the specific technical background of the motion capture and egocentric fisheye camera.
- Chapter 4 introduces a new method to accurately capture the egocentric human motion in the global coordinate system by intro-

ducing a new approach that combines concepts from SLAM and a human motion prior.

- Chapter 5 presents a new method to estimate egocentric human pose with weak supervision from an external view and an in-the-wild dataset with 3D joint pseudo-annotations obtained with supervision from an external camera system.
- Chapter 6 proposes a scene-aware egocentric pose estimation method that guides the prediction of the egocentric pose with scene constraints utilizing a joint voxel-based feature representation.
- Chapter 7 describes a new approach to capture egocentric whole-body motion, including human body and hand motion. A new version of a vision transformer, called FisheyeViT, is utilized in combination with a diffusion-based whole body motion prior to computer the final whole body motion.
- Chapter 8 concludes this thesis, summarizes the insights, and discusses possible steps of future work.

1.4 CONTRIBUTIONS

This thesis makes the following main contributions:

The contributions of Chapter 4 (published as Wang et al. (2021)) are:

- A novel framework for accurate and temporally stable global 3D human pose estimation from a monocular egocentric video;
- A new optimization algorithm with the assistance of local and global motion prior captured by an efficient convolutional network-based VAE;
- An uncertainty-aware reprojection loss to alleviate the influence of self-occlusions in egocentric settings.

The contributions of Chapter 5 (published as Wang et al. (2022)) are:

- A large in-the-wild egocentric dataset (EgoPW) captured with a head-mounted fisheye camera and an external camera;
- A new optimization method to generating pseudo labels for the in-the-wild egocentric dataset by incorporating the supervision from an external view;

- An adversarial method for training the network by learning the feature representation of egocentric images with external feature representation.

The contributions of Chapter 6 (published as Wang et al. (2023)) are:

- Synthetic and in-the-wild egocentric datasets containing egocentric pose labels and scene geometry labels;
- A new depth estimation and inpainting networks to predict the scene depth map also in regions behind the human body;
- A new egocentric pose estimation method leveraging a joint voxel-based representation of body pose features and scene geometry.

The contributions of Chapter 7 (published as Wang et al. (2024)):

- FisheyeViT for alleviating fisheye camera distortion and pose regressor using pixel-aligned 3D heatmaps for accurate egocentric body pose estimation from a single image;
- Uncertainty-aware refinement method based on motion diffusion models for correcting initial pose estimations and predicting plausible motions even under occlusion;
- EgoWholeBody, a new high-quality synthetic dataset for egocentric whole-body motion capture.

1.5 PUBLICATIONS AND PREPRINTS

The methods presented in this thesis are also publicly available in the following self-contained works:

- Wang, Jian, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt (2021). “Estimating egocentric 3d human pose in global space.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11500–11509
- Wang, Jian, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt (2022). “Estimating egocentric 3d human pose in the wild with external weak supervision.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13157–13166
- Wang, Jian, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt (2023). “Scene-aware Egocentric 3D

Human Pose Estimation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13031–13040

- Wang, Jian, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt (2024). "Egocentric Whole-Body Motion Capture with FisheyeViT and Diffusion-Based Motion Refinement." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

Further contributions were made to the following works, which are not part of this thesis:

- Akada, Hiroyasu, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik (2022). "UnrealEgo: A new dataset for robust egocentric 3d human motion capture." In: *European Conference on Computer Vision*. Springer, pp. 1–17
- Akada, Hiroyasu, Jian Wang, Vladislav Golyanik, and Christian Theobalt (2024). "3D Human Pose Perception from Egocentric Stereo Videos." In: *Computer Vision and Pattern Recognition (CVPR)*
- Millerdurai, Christen, Hiroyasu Akada, Jian Wang, Diogo Luvizon, Christian Theobalt, and Vladislav Golyanik (2024). "EventEgo3D: 3D Human Motion Capture from Egocentric Event Streams." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

2 RELATED WORK

This chapter covers the relevant background and related works on egocentric human motion capture. This chapter will discuss:

- Works related to egocentric human motion capture. (Sec. 2.1);
- Available datasets for training and evaluating egocentric human motion capture methods. (Sec. 2.2)
- The related works of weakly-supervised/whole-body 3D human pose estimation, which is related to Chapter 5 and Chapter 7. (Sec. 2.3);
- Related works that utilize human motion priors in the context of human motion capture, which is relevant for the works in Chapter 4 and Chapter 7. (Sec. 2.4);
- The related human motion capture works considering human-scene interactions, which is related to Chapter 6. (Sec. 2.5)

2.1 EGOCENTRIC HUMAN MOTION CAPTURE

As explained in Chapter 1, the works on egocentric human motion capture can be split into several different categories according to different devices for the task. This section delves into the recent advancements in inside-in egocentric motion capture techniques in detail, given their direct relevance to the configuration employed in this thesis. Additionally, methods involving alternative setups are also explored.

2.1.1 *Inside-In Egocentric Motion Capture*

Inside-in egocentric motion capture involves mounting a camera on the human body, oriented to capture the body itself. Typically, in order to ensure full coverage of the body, a fisheye camera is utilized for its wide field of view. Some studies employ stereo-fisheye cameras to enhance pose estimation accuracy and obtain the geometry of the environment. However, the use of stereo cameras adds additional challenges, including increased weight and energy demands. In contrast, this thesis adopts a single fisheye camera approach for body estimation, offering a more efficient solution. This section will cover related works of both monocular and stereo configurations.

2.1.1.1 *Egocentric Motion Capture with Monocular Fisheye Camera*

First, the inside-in egocentric motion capture methods using a single down-facing fisheye camera are discussed. Xu et al. (2019) first introduce this setup for the egocentric 3D human pose estimation. This method trains separate network modules for estimating 2D human poses and joint distances. Then, the fisheye re-projection function is employed to get the 3D human body pose. This method requires the calibration of fisheye camera models in advance. To mitigate this limitation, Zhang et al. (2021b) regressed fisheye camera parameters and 3D human pose simultaneously. Following the egocentric camera setup, Tomè et al. (2019) generated a high-quality synthetic dataset with realistic images rendered with game engines. Then, this work proposed an encoder-decoder framework that tries to recover the input image features and 3D human body pose. To address the self-occlusion issue, Park et al. (2023) leveraged the temporal information with the spatio-temporal self-attention network, and Liu et al. (2023b) introduced diffusion model to generate 3D human pose conditioned on egocentric image features. Liu et al. (2022) combined the SLAM and egocentric pose estimation methods to estimate human body poses in the world coordinate. Liu et al. (2023a) leverage the synchronized egocentric camera and external cameras to collect large-scale egocentric pose estimation datasets with pseudo-ground truth. In this work, the same device setup is adopted as described in this section, and an attempt is made to solve the challenges associated with this setup.

2.1.1.2 *Egocentric Motion Capture with a Stereo Fisheye Camera*

In this section, the setup leveraging egocentric fisheye cameras is discussed. Rhodin et al. (2016) is the first to propose the egocentric motion capture method using a lightweight stereo pair of fisheye cameras attached to a helmet or VR headset. This work combines a generative pose estimation framework with a ConvNet-based body-part detector to capture full-body motion. Cha et al. (2018) mount eight cameras on the head and leverage convolutional neural networks and a parametric-model-based approach to capture egocentric body poses, facial expressions, and scene geometry. Zhao et al. (2021) first propose a lightweight eyeglass frame with two mounted cameras. This method leverages body part information and introduces pseudo-limb masks to address self-occlusions, achieving superior results on egocentric datasets. With a similar lightweight setup, Akada et al. (2022) proposes UnrealEgo, a new large-scale synthetic dataset for egocentric 3D human pose estimation. This paper also proposes a 2D pose estimation module leveraging the stereo cameras and uses the 2D body poses to regress 3D poses. Following this work, Akada et al. (2024) propose a new transformer-based framework for improving 3D human pose estimation by leveraging scene information and temporal context. This work also extends the UnrealEgo dataset. Yang et al. (2024a) propose a two-stage pose estimation paradigm and Deformable Stereo Attention for enhancing stereo-based joint localization. Kang et al. (2023) proposed a two-path network for independent limb pose estimation using stereo heatmaps and a perspective-aware representation to estimate 3D limb orientation for accurate egocentric pose estimation. Luo et al. (2024) leverages the imitation learning to drive a simulated avatar to match the full body pose of the camera wearer. Even though the stereo camera can provide more information, this setup still suffers from extra burdens of weight and energy consumption. In light of this, this thesis focuses on the monocular egocentric motion capture setup.

2.1.2 *Egocentric Motion Capture with Other Setups*

Apart from the setups using down-facing fisheye cameras, egocentric motion capture methods using other setups are also discussed, including body-worn IMUs, VR/AR headsets, third-person egocentric motion capture, and inside-out devices like outward-facing cameras.

2.1.2.1 *IMU-Based Egocentric Motion Capture*

This section discusses the egocentric motion capture method employing body-worn inertial measurement units (IMUs). Earlier approaches utilized 17 to 18 IMUs positioned on different human limbs to achieve accurate motion capture results (Roetenberg et al., 2009; Vlasic et al., 2007). Despite their accuracy, such setups can be expensive and intrusive. To address these issues, more recent studies (Schwarz et al., 2009; Von Marcard et al., 2017) have adopted sparser IMU configurations. A pioneering study by Von Marcard et al. (2017) introduced a configuration with only six IMUs placed on the head, wrists, pelvis, and ankles, significantly reducing equipment load while maintaining motion capture fidelity. Based on this, subsequent research (Armani et al., 2024; Huang et al., 2018; Kim and Lee, 2022; Lee and Joo, 2024; Liang et al., 2023; Yi et al., 2021; Zhang et al., 2023b) has utilized deep learning to regress human motion using neural networks, even with fewer sensors.

HPS (Guzov et al., 2021) estimates the full 3D human pose and location of the human within a large 3D scene, using sparse IMUs and one extra body-mounted forward-facing camera. The initial location and body pose estimation are further optimized considering the human-scene interactions. With the same setup, Guzov et al. (2022)'s work focuses more on human-scene interactions and dynamically tracks changes in the scene made by the human. EgoLocate (Yi et al., 2023) proposes an optimization framework by incorporating the motion prior with the SLAM modules in order to improve human localization accuracy. Approaches such as TIP (Jiang et al., 2022b) and PIP (Yi et al., 2022b) make further improvements by using transformers and neural kinematics estimators to enhance motion tracking accuracy and ensure physical plausibility. Moreover, IMUPoser (Mollyn et al., 2023) used the IMUs in everyday devices like smartphones, Apple Watches, and AirPods. Lastly, Diffusion-Poser (Van Wouwe et al., 2024) explores the use of diffusion models for learning human motion priors, employing classifier-aware guidance to align generated motions with IMU signals.

While recent IMU-based motion capture systems show significant improvements, wearing the sensors can still be inconvenient, uncomfortable, and restrict natural movement.

2.1.2.2 *Inside-Out Egocentric Motion Capture*

Some research projects have experimented with mounting outward-facing cameras on the body to estimate the motion of the wearer. Shiratori et al. (2011) attached multiple cameras to all of a person's joints, using structure from motion (SfM) to localize the cameras and thus the joints. However, this multi-camera setup is cumbersome for daily use and SfM demands heavy computational resources. To address this issue, Jiang and Grauman (2017) implemented a simple approach with a single forward-facing camera mounted on the chest. This is used to directly regress the 3D positions of each body joint within the wearer's local frame. Following the same setup, Ng et al. (2020) estimated full body motions by capturing the interaction poses of a second person within the camera's view. Hwang et al. (2020) utilized a wide-view fisheye camera to capture partial views of human limbs, regressing the human body pose, body orientation, and head orientation from these partial views. Yuan and Kitani (2018, 2019) placed a front-facing camera on the head, employing imitation learning and PD control to derive a control policy that estimates current poses and predicts future ones. Luo et al. (2021) also used a front-facing camera setup, estimating physically plausible 3D motions and human-scene interactions by training a general-purpose humanoid controller that considers human kinematics, dynamics, and scene context. Li et al.

(2023) proposed a head-pose estimation algorithm and a conditional diffusion model to generate full-body motions based on the head pose trajectory. However, the reliance on outward-facing cameras means that when the human body is not in view, these methods will rely heavily on assumptions, which may not reflect accurate movements.

2.1.2.3 *Third-Person Egocentric Motion Capture*

The third-person setup refers to motion capture methods that involve a third person using cameras, typically mounted on a VR/AR headset, to observe the motion capture subject. Zhang et al. (2022) first introduced this approach, proposing a comprehensive egocentric dataset that captures high-quality 3D human motions during social interactions. They also established a benchmark for 3D human pose and shape estimation using the camera on a Hololens worn by the observer. Following this setup, Zhang et al. (2023a) introduced a scene-aware diffusion model for pose estimation in 3D environments from egocentric images and propose the physics-based collision score to guide the diffusion denoising process. Khirodkar et al. (2023a) developed an in-the-wild dataset that captures multi-human activities in unconstrained settings using egocentric devices. Their contribution also includes EgoFormer, a multi-stream transformer designed to track multiple humans from an egocentric camera perspective.

2.1.2.4 *Full-Body Pose Estimation from Head/Hand Tracking in VR/AR headsets.*

Recently, several studies have explored the potential of utilizing the head and hand tracking systems available on VR/AR headsets to estimate human motion. These systems capture the location and orientation of the head and hands to infer human motion. This setup is very similar to the IMU-based motion capture setup. Since a lot of methods adopt this specific setup, this thesis discusses it separately in the related work section. CoolMoves (Ahuja et al., 2021) pioneered this approach by employing a K-nearest-neighbor-based method to interpolate poses from a motion capture dataset. LoBSTr (Yang et al., 2021) utilized a gated recurrent unit (GRU) network to predict the lower-body pose based on past tracking signals, while the upper-body poses were determined using an inverse kinematics (IK) solver. Dittadi et al. (2021) introduced a variational autoencoder (VAE)-based optimization method designed to generate plausible and diverse human motion from various types of sparse input data. AvatarPoser (Jiang et al., 2022a) employed a transformer architecture to regress human motion, further refining predictions using an inverse kinematics (IK) solver to enhance accuracy. QuestSim (Winkler et al., 2022) integrated a reinforcement learning-based method with a physical simulator to ensure that predictions are physically plausible. QuestEnvSim (Lee et al., 2023) also utilized a reinforcement learning approach to develop a humanoid controlling strategy within a simulated environment, incorporating the surrounding geometry as an additional input signal to the control strategy. Recently, diffusion model-based methods AGRoL (Du et al., 2023) were proposed, which synthesized smooth predictions from the head and hands tracking inputs with the classifier-free guidance in the diffusion denoising model. EgoPoser (Jiang et al., 2023a) focuses on head-mounted devices(HMD)-based egocentric motion capture across large scenes. DivaTrack (Yang et al., 2024b) leverage trackers for improved foot-contact estimations and introduce a pose blending strategy that integrates upper-body predictions with lower-body generation outcomes. Dai et al. (2024) combine the use of a transformer and an LSTM to deliver a real-time pose estimation method, leveraging scalable sparse observations from HMDs and optional wearable IMUs. These methods integrate machine learning techniques with motion capture technologies, enhancing the accuracy and applicability of human pose estimation using head/hand trackers from VR/AR headsets.

2.2 DATASET FOR EGOCENTRIC HUMAN MOTION CAPTURE

Large-scale datasets are crucial for addressing deep learning challenges. However, not many such datasets are available for developing methods for egocentric pose estimation. This section summarizes the available datasets for egocentric human motion capture, which can be divided into two categories: those utilizing downward-facing cameras, as used in this thesis, and those employing other egocentric configurations. Here, we will provide a detailed introduction to the datasets using downward-facing cameras and a brief overview of other relevant datasets.

2.2.1 *Dataset with Down-facing Cameras*

Table 2.1 summarizes the key features of available egocentric pose estimation datasets, with more detailed information provided in the following sections.

The Mo^{2}Cap^2 dataset (Xu et al., 2019) is designed for egocentric motion capture using a single fisheye camera. Due to the time-consuming real-world data collection process, the Mo^{2}Cap^2 dataset is a synthetic dataset generated by combining the SMPL human body model, SURREAL textures (Doersch and Zisserman, 2019), and AMASS human

Dataset	Frames	Subject	Motion Number
Mo^{2}Cap^2 (Xu et al., 2019)	530k	700 textures	3000 motions
*x*R-egopose (Tomè et al., 2019)	383k	46 subjects	9 categories of motions
ECHP (Liu et al., 2023a)	75k	9 subjects in 20 clothes	10 categories of motions
EgoCap (Rhodin et al., 2016)	100k	8 subjects	-
EgoGlass (Zhao et al., 2021)	173k	10 subjects	6 motions
UnrealEgo (Akada et al., 2022)	900k	17 subjects	30 categories and 2.3k motions
UnrealEgo2 (Akada et al., 2024)	1250k	17 subjects	30 categories and 15k motions
UnrealEgo-RW (Akada et al., 2024)	260k	16 subjects	547 motions

Table 2.1: Datasets for Egocentric 3D Pose Estimation. The upper part shows the dataset collected with a monocular down-facing camera. The lower part shows the dataset collected with stereo down-facing cameras.

motion data (Mahmood et al., 2019). This dataset includes 530k training images with ground truth 3D joint positions.

*x*R-EgoPose (Tomè et al., 2019) is a synthetic dataset consisting of 383k frames of high-quality rendered images. It offers a rich diversity of characters, skin tones, clothing styles, backgrounds, and lighting scenarios. Featuring realistic images, the dataset is generated using Maya animation with motion capture data and V-Ray's physically based rendering setup.

The ECHP dataset (Liu et al., 2023a) is a real-world dataset comprising 75k frames. It features nine different subjects with 20 unique body textures performing ten daily actions. The authors employed two external-view cameras to capture 3D pose pseudo-annotations, achieved by triangulating the 2D pose detection results from OpenPose (Cao et al., 2017).

The first egocentric motion capture dataset for stereo egocentric cameras is EgoCap (Rhodin et al., 2016). This real-world dataset utilizes eight fixed cameras to capture ground truth 3D human motion. The egocentric images are further enhanced with background replacement and clothing color variations. The training set contains 75k fisheye images from six subjects and 25k images from two subjects for evaluation.

The UnrealEgo dataset (Akada et al., 2022) introduces a large-scale collection featuring 17 3D models and over 2,300 different actions across 14 environments. It comprises 900,000 stereo fisheye images rendered using the Unreal Engine, offering greater diversity than previous synthetic datasets. Following this, the UnrealEgo2 dataset (Akada et al., 2024) was released, providing even more diverse human motion and larger-scale renderings.

The UnrealEgo-RW dataset (Akada et al., 2024) is a real-world dataset recording a wide range of activities from 16 individuals in a multi-view motion capture studio. The dataset captures challenging motions, such as crawling and dancing. It includes 591 motion segments, resulting in over 130k stereo views (260k images).

In this thesis, in order to tackle the remaining issues with available egocentric motion capture datasets, several egocentric motion capture datasets are proposed, including EgoPW, EgoGTA, and EgoWholeMocap. The details of these datasets are introduced respectively in Chapter 5, Chapter 6, and Chapter 7.

2.2.2 *Dataset with other setups*

EGOCENTRIC DATASETS WITH IMUS The MPI08 dataset (Pons-Moll et al., 2010) is the pioneering dataset for human motion capture using IMU sensors. It features 5 IMUs attached to the human body, along with synchronized video data. Building on this, the TNT15 dataset (Von Marcard et al., 2016) includes synchronized data streams from 8 RGB cameras and 10 IMUs. This dataset contains four subjects performing five activities across 13k frames. TotalCapture (Trumble et al., 2017) represents a large-scale dataset with approximately 179k frames. It includes fully synchronized multi-view video, IMU data, and Vicon labeling. The DIP-IMU dataset (Huang et al., 2018) is the largest dataset for IMU-based human motion capture to date. It features 10 subjects equipped with 17 IMUs, yielding a total of 330k frames. The dense IMUs provide ground truth human body poses. Lastly, IMUPoser (Mollyn et al., 2023) collected a dataset using IMUs in iPhones, Apple Watches, and AirPods. This dataset comprises approximately 114k frames, gathered from 10 participants. Apart from datasets collected in the real-world environment, several methods Mollyn et al., 2023; Yi et al., 2022b, 2021 also leverage the synthetic IMU signals from the AMASS dataset (Mahmood et al., 2019).

EGOCENTRIC DATASETS WITH INSIDE-OUT SETUP Yuan and Kitani (2019) first collected a dataset with a head-mounted forward-facing camera. This dataset lasts for 8 minutes and also contains the recording from one external-view camera for evaluation. Kinpoly (Luo et al., 2021) consists of egocentric videos captured using a head-mounted camera and corresponding 3D motions captured with motion capture devices. The total motion is about 80 minutes long. GIMO (Zheng et al., 2022) is a real-world dataset consisting of egocentric video, eye gaze, 3D motions, and scanned 3D scenes. This dataset is collected using Hololens, iPhone, and motion capture suits in order to study motion prediction guided by eye gaze. This dataset contains 129k egocentric images, 11 subjects, and 217 motion trajectories in 19 different scenes. Ego-Exo4D (Grauman et al., 2024) is a large multimodal dataset containing synchronized first- and third-person videos recorded by 839 participants in 131 scenes. The dataset includes 1,422 hours of video and provides pseudo-3D body pose ground truth by using external sparse views.

EGOCENTRIC DATASETS WITH THIRD-PERSON SETUP Zhang et al. (2022) proposed a dataset for egocentric motion capture from a third-person view by gathering 125 sequences from 36 subjects across 15 indoor scenes. The dataset includes 219k external RGBD frames captured with Azure Kinect and 199k egocentric frames captured with HoloLens2. It also contains 3D human motion ground truth annotations obtained from four external RGBD views, along with ground truth 3D scene scans using an iPhone Pro Max. EgoHumans (Khirodkar et al., 2023b) collects a dataset with egocentric views obtained from Aria glasses (*Project Aria* n.d.) and external views from multi-view GoPro cameras. This dataset consists of 7 sequences from 20 subjects across 6 diverse locations, encompassing 125k RGB images, 250k greyscale images from the egocentric glasses, and 446k images from the external GoPros.

DATASETS FOR FULL-BODY POSE ESTIMATION FROM HEAD/HAND TRACKING A number of existing works (Dittadi et al., 2021; Du et al., 2023; Jiang et al., 2022a) on egocentric motion capture from head and hand tracking utilize the synthetic dataset from AMASS (Mahmood et al., 2019). DivaTrack (Yang et al., 2024b) introduces a real-world dataset featuring 6D head and hand tracking data obtained from the PICO VR headset, along with ground truth body poses from the IMU-based motion capture system. The dataset includes 772 motions and 16.5 hours of data from 22 subjects. Nymeria (Ma et al., 2024) is a large multi-modal dataset containing forward-facing camera views, 6D head/hand tracking signals, and detailed text descriptions of human motion. This dataset contains

300 hours of daily activities and 264 participants. The ground truth annotation of human motion is obtained with the IMU-based motion capture devices.

2.3 HUMAN MOTION CAPTURE WITH EXTERNAL CAMERAS/SENSORS

This section reviews related works in the area of human motion capture, a crucial area in computer vision with applications in animation, augmented reality, and human-computer interaction. The focus will be on monocular 3D human pose estimation, which is closely related to this thesis. Additionally, this section covers weakly supervised 3D pose estimation and whole-body 3D pose estimation, which is related to Sec. 5 and Sec. 7.

2.3.1 *Monocular 3D Human Motion Capture*

Monocular 3D pose estimation infers 3D positions of joints and body parts from a single camera input. This is an ill-posed task due to the monocular depth ambiguity and self-occlusion of the human body. This thesis discusses here the human 3D pose estimation from a single image and from a monocular video separately.

HUMAN 3D POSE ESTIMATION FROM A SINGLE IMAGE Some methods (Bogo et al., 2016; Guan et al., 2009; Kolotouros et al., 2021; Lassner et al., 2017; Pavlakos et al., 2019; Rempe et al., 2021; Tiwari et al., 2022; Zanfir et al., 2018) predict 2D joints and estimate 3D human pose and shape relying on iterative optimization methods.

More methods use the neural network to regress 3D poses directly. Some of the methods leverage convolutional neural networks to infer 3D human poses from a single image (Georgakis et al., 2020; Guler and Kokkinos, 2019; Kanazawa et al., 2018; Kocabas et al., 2021; Kolotouros et al., 2019a; Li and Chan, 2015; Mehta et al., 2017a; Popa et al., 2017; Rogez et al., 2017; Sun et al., 2017; Tekin et al., 2016, 2017; Tome et al., 2017; Zhang et al., 2021a; Zhou et al., 2017). Some methods train a network to perform 2D-to-3D lifting (Chen and Ramanan, 2017; Jahangiri and Yuille, 2017; Martinez et al., 2017) to regress 3D poses from 2D poses. Instead of estimating the parameters of the SMPL model, Some related approaches (Cho et al., 2022; Kolotouros et al., 2019b; Lin et al., 2021a,b) explicitly regress the vertices of the mesh with a graph convolutional network or transformer. To tackle the issue of the lack of large-scale datasets with ground truth 3D pose annotations, some works have proposed methods to generate pseudo-ground truth by using temporal information (Arnab et al., 2019), or iterative optimization in the training loop (Joo et al., 2021; Kolotouros et al., 2019a).

The proposed methods in the thesis can also fall into these categories. In Chapter 4, the thesis proposes an optimization method to get 3D poses from 2D observations. In Chapter 5, the thesis proposes a strategy to generate pseudo labels with egocentric and external views. In Chapter 6 and Chapter 7, the 3D poses are directly regressed in an end-to-end manner with deep neural networks.

HUMAN 3D POSE ESTIMATION/MOTION CAPTURE FROM MONOCULAR VIDEO Some other methods exploit temporal information and try to estimate human motion from the video. Zhou et al. (2016b) introduce EM method to estimate 3D pose from 2D predictions over the entire sequence. Mehta et al. (2017a) and Du et al. (2016) apply temporal filtering across 2D and 3D poses. Lin et al. (2017), Hossain and Little (2018), Kocabas et al. (2020), Choi et al. (2021) and Katircioglu et al. (2018) use recurrent networks to predict 3D pose sequences by leveraging previously predicted 2D and 3D poses. Pavllo et al. (2019) and Kanazawa et al. (2019) generate 3D poses with temporal convolution, while Cai et al. (2019) and Wang et al. (2020a) leverage graph convolutional networks to capture the temporal information. Luo et al. (2020) first get coarse motion with a GRU-based human motion VAE and then refine the motion with a residual estimation network. MAED (Wan et al., 2021), t-HMMR (Pavlakos et al., 2022a), and PoseBERt (Baradel et al., 2022) employ the transformer to encode the temporal information.

In this thesis, we leverage temporal information to refine human motion estimated from single-frame methods. In Chapter 4, a convolutional network-based VAE is employed to capture the motion prior and to refine the human motion. In Chapter 7, a diffusion-based refinement strategy leveraging the transformer network is introduced.

2.3.2 *Weakly-Supervised 3D Human Pose Estimation*

Recently, there has been a growing interest in developing weakly supervised 3D pose estimation methods. Weakly-supervised methods do not require datasets with paired images and 3D annotations. Some works (Novotny et al., 2019; Wang et al., 2019) leverage the non-rigid SFM to get 3D joint positions from 2D keypoint annotations in unconstrained images. Some works (Chen et al., 2019a; Chen et al., 2019b; Drover et al., 2018; Pavllo et al., 2019; Wandt and Rosenhahn, 2019) present an unsupervised learning approach to train the 3D pose estimation network with the supervision from 2D reprojections. The closest to the works in this thesis are the approaches of (Iqbal et al., 2020; Kocabas et al., 2019; Rhodin et al., 2018; Wandt et al., 2021) in that they leverage the weak supervision from multi-view images for training. Iqbal et al. (2020) and Rhodin et al. (2018) supervise the network training process by calculating the differences between Procrustes-aligned 3D poses from different views. Wandt et al. (2021) predict the camera poses and 3D body poses in a canonical form, and then supervise the training with the multi-view consistency. Kocabas et al. (2019) obtain the pseudo labels with epipolar geometry between different views and use the pseudo labels to train the 3D pose lifting network. Recently, Hua et al. (2022) propose a U-shaped graph convolutional network that can leverage the spatial configurations and cross-view correlations for 3D pose refinement. Kundu et al. (2020) enable the unsupervised training by leveraging the prior knowledge on human poses in the form of a single part-based 2D puppet model. Different from previous works, the method proposed in Chapter 5 uses a spatio-temporal optimization framework that takes egocentric and external views as input to obtain robust 3D pseudo labels for training the network. This optimization method ensures the stability of the network training process when the 2D pose estimations are inaccurate.

2.3.3 *Whole-Body 3D Human Motion Capture*

Whole-body 3D pose estimation aims to estimate the 3D human body, face, and hands parameters from input images. This task is crucial for many applications, e.g., modeling human activities and human-scene interactions. Some methods (Pavlakos et al., 2019; Xiang et al., 2019) fit the 2D body joints estimated from images with optimization algorithms, while these methods suffer from high computation overhead and can fall into local optima. Some other learning-based methods (Cai et al., 2023; Choutas et al., 2020; Feng et al., 2021; Lin et al., 2023; Rong et al., 2021; Sun et al., 2022; Zhou et al., 2021) use the neural network to regress the SMPL-X (Pavlakos et al., 2019) parameters from input images. For example, ExPose (Choutas et al., 2020) introduced body-driven attention to extract face and hand crops and used a refinement module to regress whole-body pose. OSX (Lin et al., 2023) proposes a one-stage pipeline for whole-body mesh recovery without separate networks for each part. SMPLer-X (Cai et al., 2023) proposes a foundation model for whole-body pose estimation trained with the large model and big data. Motion-X (Lin et al., 2024) proposes a large-scale dataset with precise 3D whole-body motions and corresponding text descriptions, facilitating the regression and generation of whole-body motions.

Though much progress has been made on whole-body pose estimation from an external view, the task from an egocentric view is still unexplored. Chapter 7 introduces the first whole-body 3D pose estimation method from a single egocentric image and also incorporates temporal information with diffusion-based motion refinement.

2.4 HUMAN MOTION PRIORS FOR POSE ESTIMATION

Estimating human motion, especially from a single viewpoint, is inherently an ill-posed problem. Consequently, many studies utilize motion priors to obtain plausible and most likely motions under given constraints. The human motion prior can be learned using various generative models, including Gaussian Mixture Models (GMM) (Reynolds et al., 2009), Variational Autoencoders (VAE) (Kingma and Welling, 2013), Generative Adversarial Networks (GAN) (Goodfellow et al., 2020), normalizing flows (Papamakarios et al., 2021), neural distance fields (Tiwari et al., 2022), and the recently popular diffusion denoising models (Ho et al., 2020).

Bogo et al. (2016) and Arnab et al. (2019) capture the prior to optimize the SMPL body model (Loper et al., 2015) by fitting a mixture of Gaussians to CMU mocap dataset (*CMU mocap dataset* 2008). Pavlakos et al. (2019) train a VAE to learn priors of SMPL (Loper et al., 2015) parameters on the AMASS dataset, which contains richer varieties of human motions. The motion prior is further applied to fit the SMPL parametric model on the 2D image. Humor (Rempe et al., 2021) captures the motion prior using a conditional Variational Autoencoder (VAE) trained to reconstruct the current motion from the previous one. This motion prior can be further utilized in a variety of tasks, including motion refinement, motion prediction, and motion estimation for occluded human bodies. Zanfir et al. (2020) use normalizing flow in order to avoid the compromise between KL divergence and reconstruction loss in VAE. Pose-NDF (Tiwari et al., 2022) learns the motion prior by learning a neural unsigned distance field, which learns the manifold of plausible poses as zero level set. Following a similar approach, NRDF (He et al., 2024b) proposes a novel framework for learning Neural Distance Fields (NDFs) on Riemannian manifolds. Additionally, they introduce an adaptive-step Riemannian gradient descent algorithm to accelerate convergence when mapping poses onto the manifold.

Some other methods incorporate the pose prior by training a generative adversarial network (GAN). Yang et al. (2018) develop an adversarial learning framework with a multi-source discriminator. Kanazawa et al. (2018, 2019) and Zhang et al. (2019) train discriminators for each joint rotation parameter to tell if these parameters are realistic. Kocabas et al. (2020) propose a temporal network architecture with an RNN-based discriminator for the adversarial training on the sequence of SMPL parameters.

Recently, diffusion models (Ho et al., 2020) have become popular in the human motion generation area (Dabral et al., 2023; Tevet et al., 2022; Zhang et al., 2024a) due to the high generation quality. Some methods (Choi et al., 2022; Ci et al., 2023; Foo et al., 2023; Gong et al., 2023; Holmquist and Wandt, 2023; Shan et al., 2023; Zhang et al., 2024b) have effectively applied Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) as a motion prior for the human pose estimation task. Building on the success of motion diffusion models in human pose estimation, many methods have extended this approach to egocentric pose estimation, where the human body is only partially visible from RGB cameras or VR sensors. Zhang et al. (2023a) uses a diffusion model to generate realistic human poses considering scene geometry. AGROL (Du et al., 2023) generates body motion based on head and hand 6D pose estimates from a VR headset. EgoEgo (Li et al., 2023) estimates head poses from a head-mounted front-facing camera and uses them to generate body poses. EgoHMR (Liu et al., 2023b) extracts image features and uses them as a condition for the diffusion denoising process. However, the aforementioned pose estimation methods train the *conditioned* diffusion model with image features or IMU signals. This cannot be generalized since the trained network only accepts one specific condition format and is inclined to learn domain-specific distributions of condition features. ZeDO (Jiang et al., 2023b) tackles this issue with a zero-shot diffusion-based optimization approach that doesn't require training with 2D-3D or image-3D pairs.

Different from previous methods, the method proposed in Chapter 4 captures the global motion prior learned with a lightweight sequential VAE, which enables direct optimization in the global coordinate system. Chapter 7 proposes the training of the whole-body motion diffusion model to construct the relationship between hand and body motion and leverages the uncertainty value to refine the initial motion estimation.

2.5 SCENE-AWARE HUMAN POSE ESTIMATION

In recent years, several approaches have been proposed to predict the pose of humans considering environmental and physical constraints from RGB (Pavlakos et al., 2022b; Shimada et al., 2020; Yi et al., 2022a) and inertial measurement units (IMU) (Guzov et al., 2021; Yi et al., 2023, 2022b). Some methods assume a simplified environment, such as a planar ground floor, to enforce a temporal sequence that is physically consistent with the universal law of gravity by assuming known camera poses (Shimada et al., 2021, 2020) or by tracking an object in the scene following a free flight trajectory (Dabral et al., 2021). Other approaches assume that the scene is given as input, either as a 3D reconstruction (Guzov et al., 2022, 2021; Shimada et al., 2022; Yi et al., 2023) or as geometric primitives (Yu et al., 2021), whose human motion and global location can be refined in the optimization process. Bhatnagar et al. (2022) proposed a method and dataset for human-object interactions. Taking into account the interaction between humans and furniture, holistic methods are able to estimate the position of humans and specific objects in the scene under the assumption of a planar floor (Chen et al., 2019c; Weng and Yeung, 2021; Yi et al., 2022a), or even to estimate deformations in known objects based on human poses (Li et al., 2022b). Contrary to the previous work, the method in Chapter 6 makes no strong assumptions about the objects and ground floor in the scene but instead proposes a method that learns to estimate the background scene geometry from a fisheye camera and explores the correlation between the human body and scene directly from egocentric data.

3 BACKGROUND

After discussing the broader literature in the previous chapter, this chapter focuses on the specific concepts needed for understanding the thesis. The first section introduces the fisheye camera model used in all of the following chapters. The second section introduces the way of tracking egocentric camera position, an important process when collecting real-world datasets in Chapter 6 and Chapter 7.

3.1 FISHEYE CAMERA MODEL

This section provides an overview of fisheye camera models, which is a mathematical model for the optics of fisheye cameras. The thesis employs a fisheye camera to capture human motion, taking advantage of its extensive field of view. However, this benefit comes with significant visual distortions, requiring the use of specialized camera models different from those used in perspective cameras.

This thesis employs Scaramuzza's fisheye camera model (Scaramuzza et al., 2006), chosen for its simplicity and universal applicability. The projection and reprojection functions of this model are explained as follows:

The projection function $\mathcal{P}(x, y, z)$ of a 3D point $[x, y, z]^T$ in the fisheye camera space into a 2D point $[u, v]^T$ on the fisheye image space can be written as:

$$[u, v]^T = f(\rho)\frac{[x, y]^T}{\sqrt{x^2 + y^2}} \tag{3.1}$$

where $\rho = \arctan(z/\sqrt{x^2 + y^2})$ and $f(\rho) = k_0 + k_1\rho + k_2\rho^2 + k_3\rho^3 + \ldots$ is a polynomial obtained from camera calibration.

Given a 2D point $[u, v]^T$ on the fisheye images and the distance d between the 3D point $[x, y, z]^T$ and the camera, the position of the 3D point can be obtained with the fisheye reprojection function $\mathcal{P}^{-1}(u, v, d)$:

$$[x, y, z]^T = d\frac{[u, v, f'(\rho')]^T}{\sqrt{u^2 + v^2 + (f'(\rho'))^2}} \tag{3.2}$$

where $\rho' = \sqrt{u^2 + v^2}$ and $f'(\rho) = k'_0 + k'_1\rho + k'_2\rho^2 + k'_3\rho^3 + \ldots$ is another polynomial obtained from camera calibration. The calibration of the fisheye camera and more details about the fisheye camera model can be found in Scaramuzza *et al.* (Scaramuzza et al., 2006).

Another important fisheye camera model is the Kannala-Brandt model (Kannala and Brandt, 2006). The projection function of the Kannala-Brandt model $\mathcal{P}(x, y, z)$ is:

$$[u, v]^T = d(\theta)[f_x \frac{x}{r}, f_y \frac{y}{r}]^T \tag{3.3}$$

where $r = \sqrt{x^2 + y^2}$, $\theta = \arctan(r/z)$ and $d(\theta) = \theta + k_1\theta^3 + k_2\theta^5 + k_3\theta^7 + \dots$ is a polynomial function obtained from camera calibration. This camera model is widely used and also implemented in OpenCV, but it cannot model the fisheye cameras with a field of view larger than 180°.

Note that all of the methods in our thesis do not depend on one specific fisheye camera model.

3.2 TRACKING EGOCENTRIC CAMERA

To estimate accurate egocentric camera poses and further obtain the ground truth body poses under the egocentric camera perspective, a calibration board is mounted on the head, rigidly attached to the egocentric camera. The pose of the egocentric camera can be estimated with a multi-view capturing system with the following approach.

First, the transformation matrix $\mathbf{M}_{\text{head2ego}}$ between the calibration board and the fisheye camera is estimated with hand-eye calibration (Tsai and Lenz, 1988). A second calibration board is placed on the scene in a place where it can be seen by both the egocentric camera and the studio cameras. Then the relative pose $\mathbf{M}_{\text{ego2calib}}$ between the egocentric camera and the external calibration board, the relative pose between the studio cameras and the external calibration board $\mathbf{M}_{\text{ext2calib}}$, and the relative pose between the studio cameras and the head-mounted calibration board $\mathbf{M}_{\text{ext2head}}$ are estimated. The transformation matrix $\mathbf{M}_{\text{head2ego}}$ can be finally obtained with:

$$\mathbf{M}_{\text{head2ego}} = \mathbf{M}^{-1}_{\text{ext2head}} \mathbf{M}_{\text{ext2calib}} \mathbf{M}^{-1}_{\text{ego2calib}} \tag{3.4}$$

During the data collection process, the pose of the calibration board is estimated from each single view, and the averaged calibration board poses **M**ext2head are obtained. The egocentric camera pose **M**ext2ego can be obtained with:

$$\mathbf{M}_{\text{ext2ego}} = \mathbf{M}_{\text{ext2head}} \mathbf{M}_{\text{head2ego}} \tag{3.5}$$

The egocentric camera pose enables the transformation of the ground truth pose from the studio camera coordinate system P_{ext} to the egocentric camera coordinate system $P_{\text{ego}} = P_{\text{ext}} \mathbf{M}_{\text{ext2ego}}$.

4 ESTIMATING EGOCENTRIC 3D HUMAN POSE IN GLOBAL SPACE

This chapter presents the first approach in the literature (published as Wang et al., 2021) that estimates human motion in the global space using a single egocentric camera. To achieve accurate and temporally stable global poses, a spatio-temporal optimization is performed over a sequence of frames by minimizing heatmap reprojection errors and enforcing local and global body motion priors learned from a mocap dataset. Experimental results show that this approach outperforms state-of-the-art methods both quantitatively and qualitatively.

4.1 INTRODUCTION

Traditional optical motion capture systems with external, outside-in-facing cameras are restrictive for many pose estimation applications that require the person to be able to roam around in a larger space, beyond a fixed recording volume. Examples are mobile interaction applications, pose estimation in large-scale workplace environments, and many AR/VR applications. To enable this, methods for egocentric 3D human pose estimation using head- or body-mounted cameras were researched. These methods are mobile, flexible, and have the potential to capture a wide range of daily human activities even in large-scale cluttered environments.

Some egocentric capture methods study the estimation of face (Elgharib et al., 2019, 2020; Li et al., 2015) and hand motions (Ma et al., 2016; Singh et al., 2016; Singh et al., 2017; Sridhar et al., 2015), while the estimation of the global full body pose has been less explored. Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019) use a single head-mounted fisheye camera to capture the 3D skeletal body pose in a marker-less way. Both methods have demonstrated compelling 3D pose estimation results while still suffering from an important limitation: They estimate the local 3D body pose in egocentric camera space, while not being able to obtain the body pose with global position and orientation in the world coordinate system. Henceforth, the former will be referred to as the "local pose" to distinguish it from the "global pose" defined in the world coordinate system. Local pose capture alone is insufficient for many applications. For example, captured local body poses are not enough to animate the

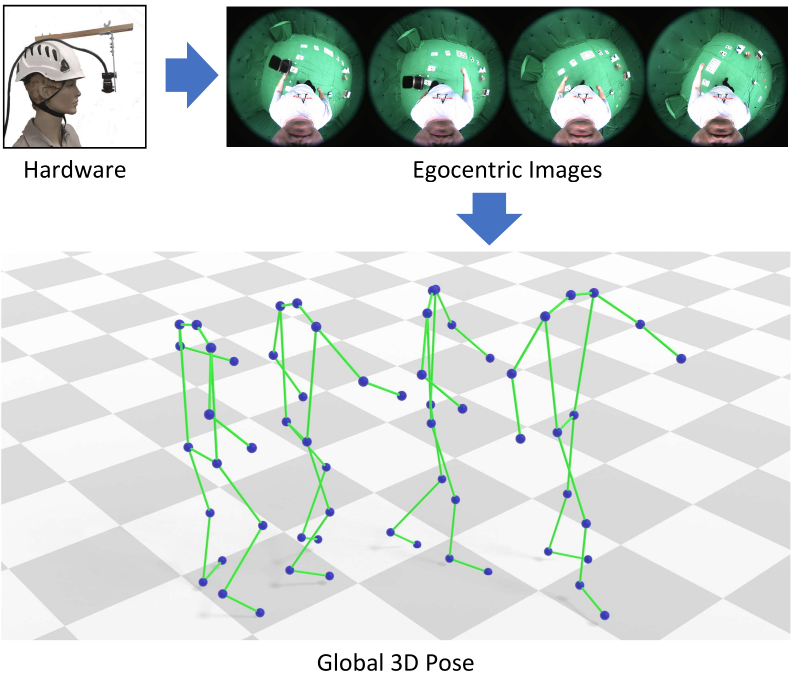


Figure 4.1: Given challenging egocentric videos, the proposed method produces realistic and accurate 3D global pose sequences.

locomotion of a virtual avatar in *x*R environments, which requires global poses.

A straightforward solution would be to simply project the local pose into the world coordinate system with the egocentric camera pose estimated by the SLAM. However, the obtained global poses exhibit significant inaccuracies. First, they show notable temporal jitters as the video frames are processed independently without taking temporal frame coherence into account. Second, they often show tracking failure due to the self-occlusion in the distorted view of the fisheye camera. Third, the obtained global poses often show unrealistic motions (such as foot sliding and global jitters) due to the inconsistency between the local pose and the estimated camera pose, which are independent of each other.

To tackle these challenges, this chapter proposes a novel approach for accurate and temporally stable egocentric global 3D pose estimation with a single head-mounted fisheye camera, as illustrated in Fig. 4.1. In order to obtain temporally smooth pose sequences, the proposed method resorts to a spatio-temporal optimization framework where we leverage the 2D and 3D keypoints from CNN detection as well as VAE-based motion priors learned from a large mocap dataset. The VAE-based motion priors have been proven effective to produce realistic and smooth motions in pose estimation methods like VIBE (Kocabas et al., 2020) and MEVA(Luo et al., 2020). However, the RNN-based VAEs in these works are less efficient and unstable due to the vanishing and exploding gradients during our optimization process. Therefore, a new convolutional VAE-based motion prior is proposed, which enables faster optimization speed and higher accuracy.

Furthermore, to reduce the error due to strong occlusion, a novel uncertainty-aware reprojection energy term is proposed by summing up the probability values at the pixels on the heatmap occupied by the projection of the 3D estimated joints rather than comparing the projection of 3D estimated joints against the predicted 2D joint position. Finally, a global pose optimizer with a separate VAE is introduced to make the local body poses consistent with the camera poses estimated by SLAM.

The method is evaluated on the dataset provided by Mo^{2}Cap^2 (Xu et al., 2019) and also a new benchmark collected with 2 subjects performing various motions. The method outperforms the state-of-the-art methods both quantitatively and qualitatively. Ablative analysis confirms the efficacy of our proposed optimization algorithm with learned motion prior and uncertainty-aware reprojection loss for improved local and global accuracy and temporal stability. To summarize, the technical contributions of this chapter are as follows:

- A novel framework for accurate and temporally stable global 3D human pose estimation from a monocular egocentric video.
- A new optimization algorithm that utilizes a local and a global motion prior encoded in an efficient convolutional network-based VAE.
- An uncertainty-aware reprojection loss to alleviate the influence of self-occlusions in egocentric settings.

The proposed method works for a wide range of motions in various environments. This method also outperforms various baselines in terms of the accuracy of the estimated global and local pose. We recommend watching the video in http://gvv.mpi-inf.mpg.de/projects/globalegomocap for better visualization.

4.2 METHOD

The goal of this method is to estimate the global body poses from a video sequence captured by a head-mounted fisheye camera. An overview of the pipeline is provided in Fig. 4.2. The video frames are split into segments with B frames each ($B = 10$ in the experiments). The pipeline

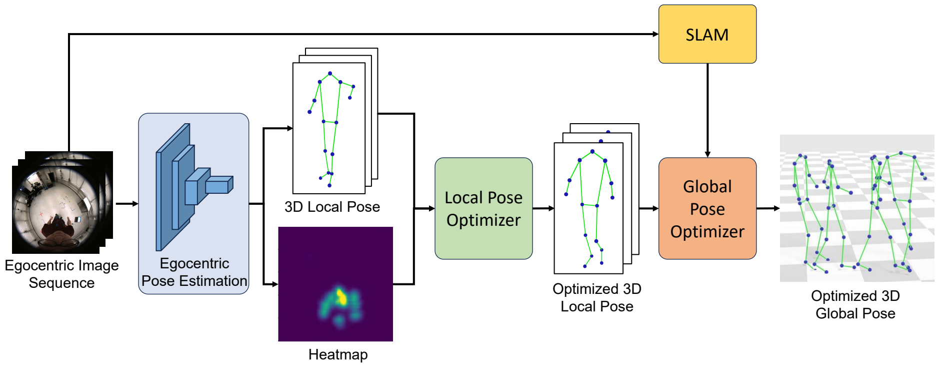


Figure 4.2: Overview of the method. This method takes an egocentric video as input and processes it in segments. For each segment consisting of a fixed number of consecutive frames, this method first applies an egocentric pose estimation method to obtain initial 3D local poses and 2D heatmaps which are then fed into the local pose optimization framework to get optimized local poses. Next, combined with the camera poses estimated from ORB-SLAM2, the optimized 3D local poses are transformed from the local egocentric camera space to the world coordinate space and then optimized via the global pose optimization to produce the final global poses.

takes one segment consisting of B consecutive frames, $\mathcal{I}_{seq} = \{\mathcal{I}_1, \ldots, \mathcal{I}_B\}$, as inputs and outputs the global poses of all the individual frames, $\mathcal{P}^g_{seq} = \{\mathcal{P}^g_1, \ldots, \mathcal{P}^g_B\}$. For each segment, this method first calculates the 3D local pose and 2D heatmap of each frame using an egocentric local body pose estimation method (Sec. 4.2.1). Next, the local motion prior is learned from local motion sequences of the AMASS dataset (Mahmood et al., 2019) with a sequential VAE (Kingma and Welling, 2013) (Sec. 4.2.2.1), and this method performs a spatio-temporal optimization with the local motion prior by minimizing the heatmap reprojection term and several regularization terms (Sec. 4.2.2.2). Given the optimized local poses, they are transformed from local fisheye camera space to the world coordinate system with camera poses estimated by a SLAM method to get initial global poses (Sec. 4.2.3.1). To improve global poses, the global pose prior is learned by training a second sequential VAE on the global motion sequences of the AMASS dataset, and the global prior is imposed in a spatio-temporal global pose optimization process(Sec. 4.2.3.2). Please refer to the supplementary materials for implementation details.

4.2.1 *Local Pose Estimation*

Given a segment containing B consecutive frames $\mathcal{I}_{seq}$, this method firstly estimate local poses represented by 15 joint locations $\widetilde{\mathcal{P}}_{seq} = \{\widetilde{\mathcal{P}}_1, \ldots, \widetilde{\mathcal{P}}_B\}$,

$\widetilde{\mathcal{P}}_i \in \mathbb{R}^{15\times 3}$, and 2D heatmaps $\mathcal{H}_{seq} = \{\mathcal{H}_1, \ldots, \mathcal{H}_B\}$ using an egocentric local pose estimation method. Note that this approach can work with any egocentric local pose estimation methods. In the experiments, the approach is evaluated on the results of two state-of-the-art methods: Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019).

4.2.2 *Local Pose Optimization*

Although Mo^{2}Cap^2 and xR-egopose can produce compelling results, both approaches suffer from limited accuracy and temporal instability, which is mainly due to depth ambiguities caused by the monocular setup and severe occlusions in a strongly distorted egocentric perspective. To improve local poses, this chapter proposes an efficient spatio-temporal optimization framework that first learns the local pose prior as a latent space with a sequential VAE (Kingma and Welling, 2013) (Sec. 4.2.2.1) and then searches for a latent vector in the learned latent space by minimizing a reprojection term and some regularization terms (Sec. 4.2.2.2).

4.2.2.1 *Learning Motion Prior*

To construct a latent space encoding local motion prior, a sequential VAE (Kingma and Welling, 2013) is trained on local motion sequences of the AMASS dataset (Mahmood et al., 2019) which are split into segments for training. A segment consisting of B consecutive poses is denoted as $\mathcal{Q}_{seq} = \{\mathcal{Q}_1, \ldots, \mathcal{Q}_B\}(\mathcal{Q}_i \in \mathbb{R}^{15\times 3})$. The sequential VAE consists of an encoder f_{enc} and a decoder f_{dec}. The encoder is used to map an input sequence of human local poses $\mathcal{Q}_{seq}$ to a latent vector z, and the decoder is used to reconstruct a pose sequence, $\widehat{\mathcal{Q}}_{seq} = \{\widehat{\mathcal{Q}}_1, \ldots, \widehat{\mathcal{Q}}_B\}(\widehat{\mathcal{Q}}_i \in \mathbb{R}^{15\times 3})$, from the latent vector. Following (Kingma and Welling, 2013), the training loss of VAE is formulated as:

$$\mathcal{L}_{total} = c_1 \left\| \widehat{\mathcal{Q}}_{seq} - \mathcal{Q}_{seq} \right\|_2^2 + c_2 KL[q(z|\mathcal{Q}_{seq}) \| \mathcal{N}(0, I)] \tag{4.1}$$

where $z = f_{enc}(\mathcal{Q}_{seq})$, $\widehat{\mathcal{Q}}_{seq} = f_{dec}(z)$, $q(z|\mathcal{Q}_{seq})$ refers to the projected distribution of $\mathcal{Q}_{seq}$ in the latent space, $\mathcal{N}(0, I)$ refers to the standard normal distribution, and $KL(.)$ refers to the Kullback–Leibler divergence.

Different from previous pose estimation methods (Kocabas et al., 2020; Luo et al., 2020) which leverage RNN-based VAEs to capture the motion prior, both the encoder f_{enc} and the decoder f_{dec} of the sequential VAE are designed as 5-layer 1D convolutional networks. Compared with RNN-based VAEs, the convolutional networks in the sequential VAE are more efficient in the optimization iterations since they can be parallelized over a time sequence. Moreover, the RNNs suffer from vanishing and exploding gradients more easily, which makes the optimization process less stable. The sequential VAE in the method has been compared with RNN-based VAEs in VIBE (Kocabas et al., 2020) and MEVA (Luo et al., 2020) in the ablation study. More details of sequential VAE are shown in the supplementary materials.

4.2.2.2 *Optimizing Local Poses with Local Motion Prior*

With the learned latent space of local motion, the task of optimizing local poses with the local motion prior can be formulated as the problem of finding a latent vector z in the learned latent space such that the reconstructed local pose sequence $\mathcal{P}_{seq} = f_{dec}(z)$ minimizes the following objective function:

$$\begin{aligned} E(\mathcal{P}_{seq}) = \lambda_R E_R(\mathcal{P}_{seq}) + \lambda_J E_J(\mathcal{P}_{seq}, \widetilde{\mathcal{P}}_{seq}) \\ + \lambda_T E_T(\mathcal{P}_{seq}) + \lambda_B E_B(\mathcal{P}_{seq}) \end{aligned} \tag{4.2}$$

where $E_R(.), E_J(.), E_T(.), E_B(.)$ are the reprojection term, pose regularization term, motion smoothness regularization term, and bone length regularization term, respectively, which will be described in detail later. In the experiment, we set the weights $\lambda_R = 0.01$, $\lambda_J = 0.01$, $\lambda_T = 1$ and $\lambda_B = 0.01$, respectively.

HEATMAP-BASED REPROJECTION: Previous works (Arnab et al., 2019; Bogo et al., 2016; Pavlakos et al., 2019; Zanfir et al., 2020) calculate the reprojection term by summing up the Euclidean distance values between the projection of estimated 3D joints and detected 2D joints. However, this calculation is sensitive to 2D joint detection errors due to the strong self-occlusions caused by the egocentric perspective. To tackle this issue, this approach defines a heatmap-based reprojection error by leveraging the uncertainty captured in the predicted 2D heatmaps, where the value at each pixel describes the probability of this pixel being a 2D joint. This new reprojection term is calculated by maximizing the summed heatmap values at the reprojected 2D joint positions:

$$E_R(\mathcal{P}_{seq}) = -\sum_{i=1}^{B} \|\mathrm{HM}_i(\Pi(\mathcal{P}_i))\|_2^2 \tag{4.3}$$

where $\mathrm{HM}_i(.)$ returns the value at a pixel on $\mathcal{H}_i$, the heatmap of i-th frame. $\Pi(.)$ refers to the projection of a 3D point. Specifically, the projection of a 3D point $[x, y, z]^T$ can be written as:

$$[u, v]^T = \frac{[x, y]^T}{\sqrt{x^2 + y^2}} \times f(\rho) \tag{4.4}$$

where $\rho = \arctan(z/\sqrt{x^2 + y^2})$ and $f(\rho) = \alpha_0 + \alpha_1\rho + \alpha_2\rho^2 + \alpha_3\rho^3 + \dots$ is a polynomial obtained from camera calibration.

POSE REGULARIZATION: To constrain the optimized pose $\mathcal{P}_i$ to stay close to the initial pose $\widetilde{\mathcal{P}}_i$, the pose regularize is defined as:

$$E_J(\mathcal{P}_{seq}, \widetilde{\mathcal{P}}_{seq}) = \sum_{i=1}^{B} \left\| \mathcal{P}_i - \widetilde{\mathcal{P}}_i \right\|_2^2 \tag{4.5}$$

MOTION SMOOTHNESS REGULARIZATION: Same as Mehta et al. (2020)'s work, the temporal smoothness regularizer (Eq. 4.6) is used to improve the temporal stability of the estimated poses, which is calculated based on the acceleration of each joint over the whole sequence:

$$E_T(\mathcal{P}_{seq}) = \sum_{i=2}^{B} \|\nabla\mathcal{P}_i - \nabla\mathcal{P}_{i-1}\|_2^2 \tag{4.6}$$

where $\nabla\mathcal{P}_i = \mathcal{P}_i - \mathcal{P}_{i-1}$.

BONE LENGTH REGULARIZATION: To explicitly enforce the constraint that each bone length stays fixed, we define the bone length regularizer as the difference between the bone length and the average bone length over the pose sequence.

$$E_B(\mathcal{P}_{seq}) = \sum_{i=1}^{B} \left\| L_{\mathcal{P}_i} - \frac{1}{B}\sum_{j=1}^{B} L_{\mathcal{P}_j} \right\|_2^2 \tag{4.7}$$

where the $L_{\mathcal{P}_i}$ is a vector composed of the length of each bone of 3D pose $\mathcal{P}_i$.

4.2.3 *Global Pose Estimation*

Based on the pose optimized by the local pose optimizer, the goal is to get the 3D pose in the global coordinate system. First, the monocular SLAM is used to get the camera pose sequence and project the local pose sequence to the global space (Sec. 4.2.3.1), then the initial global pose sequence is optimized with the global pose optimizer (Sec. 4.2.3.2).

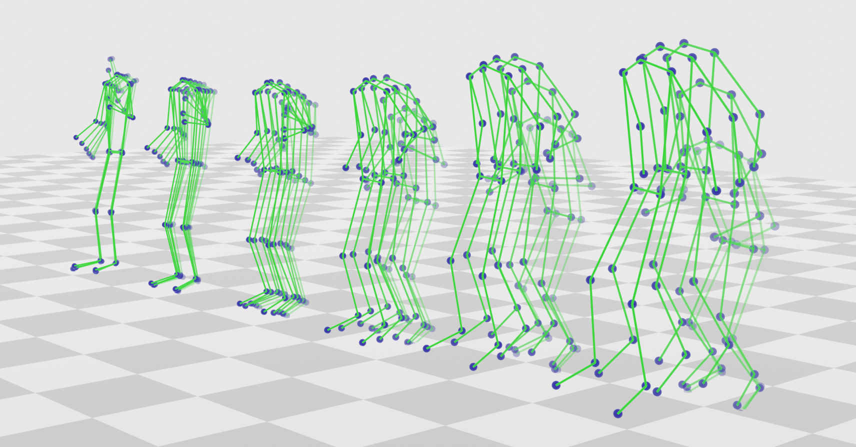


Figure 4.3: Interpolation in the latent space. The leftmost and rightmost pose sequences (waving hands and jumping) are reconstructed from two randomly sampled latent vectors, and intermediate pose sequences are reconstructed from linear interpolation between the left and right latent vectors.

4.2.3.1 *Initialization*

To obtain the initial global body poses, the camera poses are first estimated using ORB-SLAM2 (Mur-Artal and Tardós, 2017). To avoid the effects caused by the moving person in the egocentric view, a square-shaped mask that roughly covers a large portion of the body is employed to remove most of the feature points detected on the main body parts. A fixed mask is used rather than estimating a silhouette mask for each image for the sake of effectiveness and robustness.

With the estimated camera pose (R_i, t_i) $(i = 1, \cdots, B)$, the local body pose P_i can be transformed into the world coordinate space to obtain its initial global body pose $\widetilde{P}_i^g$:

$$\widetilde{\mathcal{P}}_i^g = R_i \cdot \mathcal{P}_i + t_i, \widetilde{\mathcal{P}}_i^g \in \widetilde{\mathcal{P}}_{seq}^g \tag{4.8}$$

where $\widetilde{\mathcal{P}}_{seq}^g$ is the corresponding inital global pose segment of $\mathcal{P}_{seq}$.

4.2.3.2 *Global Pose Optimizer*

Simply combining local poses with camera poses would not achieve very high-quality global poses because the optimized local body poses are not constrained to be consistent with the corresponding camera poses. For example, the initial global pose in the left part of Fig. 4.4 suffers from the foot skate artifact, which means the foot moves when it should remain in a fixed position on the ground. In order to alleviate such inconsistency

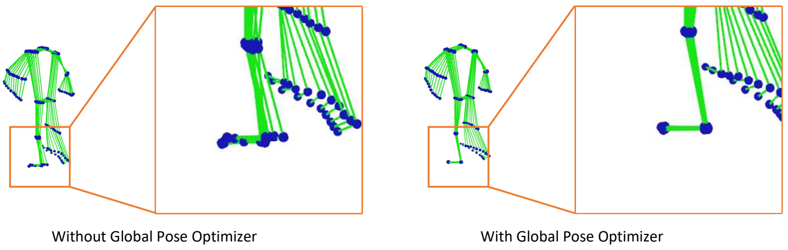


Figure 4.4: The global pose with/without global pose optimizer. Here, we zoom onto the left foot for better comparison.

errors, this method perform another spatio-temporal optimization on the initial global pose. A sequential VAE is firstly trained on global pose sequences from the AMASS dataset in the same way presented in Sec. 4.2.2.1. To measure the smoothness of our learned latent space, an experiment is conducted by interpolating two different body motions. The results shown in Fig. 4.3 demonstrate that the learned latent space is smooth, which is important for the subsequent optimization process. With the learned latent space of global motion, the goal is to find a latent vector z^g such that the global pose sequence $\mathcal{P}^g_{seq} = f^g_{dec}(z^g)$ minimizes the following objective function:

$$E(\mathcal{P}^g_{seq}) = \lambda_J E_J(\mathcal{P}^g_{seq}, \widetilde{\mathcal{P}}^g_{seq}) + \lambda_T E_T(\mathcal{P}^g_{seq}) + \lambda_B E_B(\mathcal{P}^g_{seq}) \tag{4.9}$$

where $E_J(.), E_T(.), E_B(.)$ are the same as those in 4.2.2.2, and λ_J, λ_T and λ_B are set as 0.01, 1 and 0.01, respectively. The example of the optimized result is illustrated in the right part of Fig. 4.4, where the footskate artifact is alleviated due to our global optimizer.

4.3 EXPERIMENTS

4.3.1 *Datasets*

Following Xu et al. (2019)'s and Tomè et al. (2019)'s works, the local egocentric pose estimators are trained on the synthetic dataset from Mo^{2}Cap^2. The AMASS dataset (Mahmood et al., 2019) is used to train the sequential VAEs. To make the distribution of joint position in the training data consistent with that in the real-world data, a virtual fisheye camera is attached to the forehead of the human mesh at a distance similar to the capture settings.

The method is evaluated on both the real-world dataset from Mo^{2}Cap^2 (Xu et al., 2019) and a new egocentric dataset. The new real-world egocentric dataset was captured using a head-mounted fisheye camera with a similar camera position as Mo^{2}Cap^2 (Xu et al., 2019) while the ground truth 3D poses were acquired using a multi-view motion capture system. This dataset contains around 12k frames of 2 actors wearing different clothes and performing 13 types of actions. This dataset has been made publicly available.

Compared with the Mo^{2}Cap^2 test set and the xR-egopose test set (unreleased), this new test set contains more types of actions and more data with global motions. The Mo^{2}Cap^2 test set contains 5591 frames (2 actors performing 8 types of actions). The xR-egopose test set has 10k frames (3 actors performing 6 types of actions).

4.3.2 *Evaluation Metrics*

The method is evaluated with three different metrics, namely PA-MPJPE, the bone length aligned MPJPE (BA-MPJPE), and the global MPJPE. They all calculate the Mean Per Joint Position Error (MPJPE) but use different ways of alignment to the ground truth.

For **PA-MPJPE**, the estimated pose of each frame is rigidly aligned to the ground truth pose P_{seq} using $\hat{P}_{seq}$ with Procrustes analysis (Kendall, 1989). For **BA-MPJPE**, firstly the bone length of each frame in sequences $\hat{P}_{seq}$ and P_{seq} is resized to the bone length of a standard skeleton. Then, the PA-MPJPE between the two resulting sequences is calculated. For **Global MPJPE**, all the poses of each batch (100 frames) are globally aligned to the ground truth using Procrustes analysis. Each metric has its own focus. The PA-MPJPE measures the accuracy of a single pose while BA-MPJPE eliminates the effects of body scale. The global MPJPE calculates the accuracy of global joint positions, considering the global translation and rotation.

4.3.3 *Implementation Details*

4.3.3.1 *Sequential VAE*

The input pose sequence with n frames is firstly reshaped to $(3 \times 15, n)$ and fed into the encoder with 45 input channels. The encoder has five 1D conv blocks with 64, 64, 128, 256, and 512 output channels. Each conv block contains one 1D conv layer (kernel size=3, stride=1 and padding=1), one batch norm layer, and one leaky relu layer with negative slope=0.01. The output of the encoder is sent into two linear layers giving $\mu, \sigma \in \mathbb{R}^{2048}$. The latent vector z is sampled with μ, σ with the reparameterization trick.

For the decoder, the sampled latent vector z is firstly fed into a linear layer with output dimension $n \times 512$, and five 1D de-conv blocks with

256, 128, 64, 64, and 64 output channels. Each block contains one 1D deconv layer, one batch norm layer, and one leaky relu layer with the same hyper-parameters as the encoder. The output vector with 45 channels is obtained from a final conv layer (kernel size=3, stride=1 and padding=1). The output vector is eventually reshaped to $(n, 15, 3)$, representing a pose sequence as the input.

During training, the weight of reconstruction loss and KL divergence loss is set to 1 and 5×10^{-3} respectively.

4.3.3.2 *Optimization Details*

In local and global pose optimization frameworks, the latent vector z is optimized using a PyTorch implementation and the Limited-memory BFGS optimizer (L-BFGS) (Nocedal and Wright, 2006) with strong Wolfe line search. A learning rate of 2.0 with 30 maximum iterations is used. z is initialized using the results of the single-frame egocentric pose estimation network $z = f_{enc}(\mathcal{P}_{seq})$. After optimization, the output pose sequence is reconstructed from the optimized z with a VAE decoder $f_{dec}(z)$.

Each long sequence is firstly split into several overlapping segments with length B and each segment is processed independently. After two adjacent segments are processed, the overlapping parts between these segments are merged in a linear combination way. For a segment with length $B = 10$, the local pose optimizer, running on 10-frame segments, takes 120.0 ms per segment while the global pose optimizer takes 75.7 ms per segment. The optimization process is time-efficient thanks to the simple VAE network and GPU-based optimization algorithm. All aforementioned time is measured on a computer with Xeon 6144 CPU and Tesla V100 GPU.

4.3.4 *Comparison with State-of-the-art Results*

Table 4.1 compares the approach with previous state-of-the-art single-frame-based methods on the proposed test dataset and the indoor sequences from the Mo^{2}Cap^2 dataset. Since the code or the predictions of xR-egopose are not publicly available, this method is re-implemented and used for comparison. In order to obtain the global pose for Mo^{2}Cap^2 and xR-egopose, the local predictions are rigidly transformed to the world coordinate system with the camera pose estimated by SLAM. This global pose is regarded as the main baseline and denoted as Mo^{2}Cap^2 (or xR-egopose) + SLAM. Since the camera poses from ORB-SLAM2 are ambiguous with respect to the scene scale, the scale is further estimated by calibrating the camera position with a checkerboard in the first few

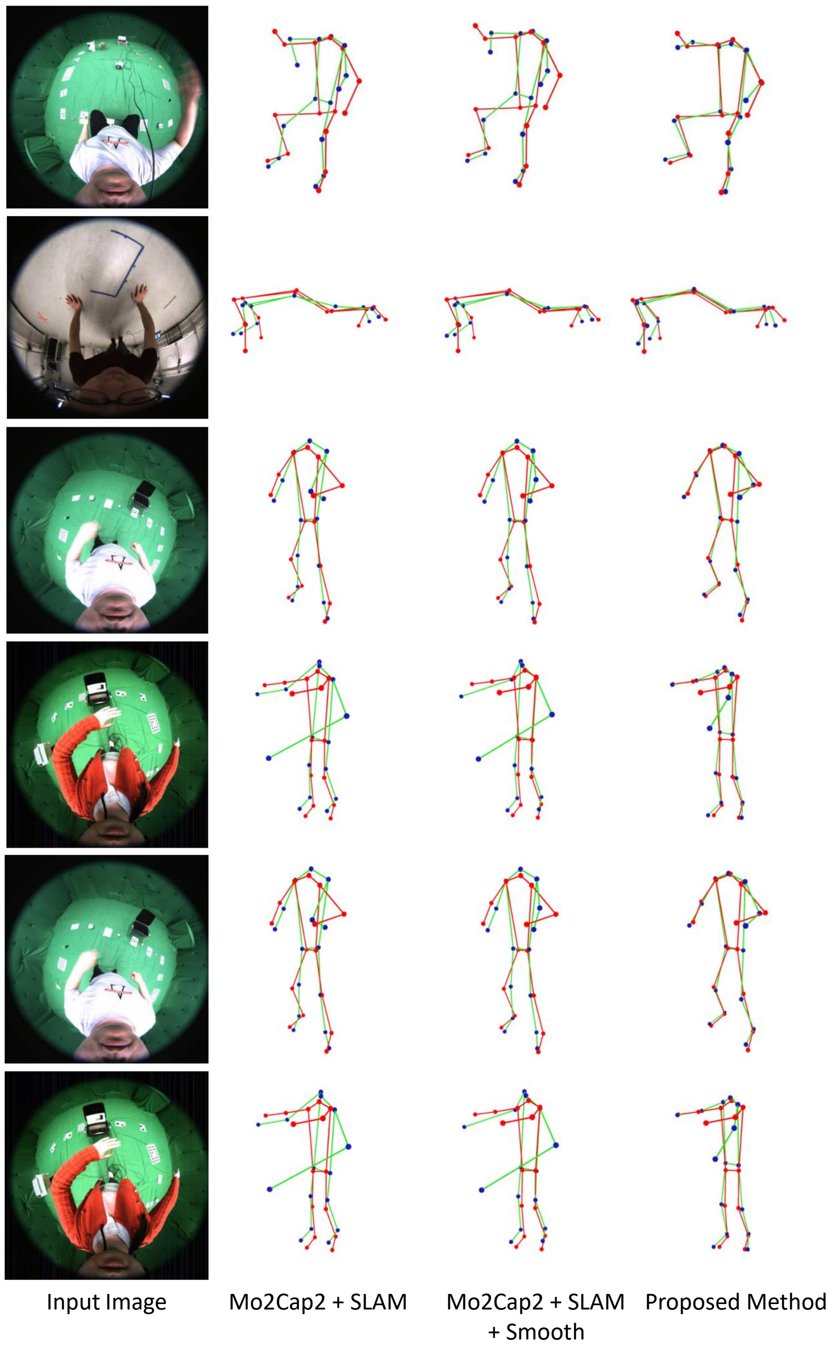


Figure 4.5: Qualitative comparison on the accuracy of a single pose. From left to right: input image, Mo^{2}Cap^2 result projected with SLAM (green), smoothed Mo^{2}Cap^2 result projected with SLAM (green), and proposed method's result (green) overlaid on ground truth (red). Note that in order to better show the result, the estimated pose is rigidly aligned to the ground truth.

Method	Global MPJPE	PA-MPJPE	BA-MPJPE
Mo^{2}Cap^2 test dataset			
Mo^{2}Cap^2+SLAM	117.4	80.48	61.40
Mo^{2}Cap^2+SLAM+Smooth	113.0	76.92	58.25
Mo^{2}Cap^2+Proposed	**110.5**	**69.87**	**52.90**
xR-egopose+SLAM	114.0	71.33	55.43
xR-egopose+SLAM+Smooth	112.2	70.27	54.03
xR-egopose+Proposed	**110.1**	**66.74**	**50.52**
The new test dataset			
Mo^{2}Cap^2+SLAM	141.8	102.3	74.46
Mo^{2}Cap^2+SLAM+Smooth	135.5	96.37	70.84
Mo^{2}Cap^2+Proposed	**119.5**	**82.06**	**62.07**
xR-egopose+SLAM	163.4	112.0	87.20
xR-egopose+SLAM+Smooth	158.1	109.6	84.70
xR-egopose+Proposed	**134.1**	**84.97**	**64.31**

Table 4.1: The experimental results on Mo^{2}Cap^2 test dataset (Xu et al., 2019) and the new test dataset. Mo^{2}Cap^2 (or xR-egopose) + Proposed is the result of the proposed method based on the predictions of Mo^{2}Cap^2 (or xR-egopose). The proposed method outperforms previous state-of-the-art Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019) in all of the three metrics.

frames of the sequence. Note that since the Mo^{2}Cap^2 dataset does not provide frames with a checkerboard, the Procrustes analysis is applied to align the trajectory estimated by SLAM with the ground truth trajectory to compute the scale. For a fair comparison, the global pose of Mo^{2}Cap^2 and xR-egopose is also smoothed with a Gaussian filter and the results are denoted as Mo^{2}Cap^2 (or xR-egopose) + SLAM + smooth. The evaluation of different types of motion is shown in Sec. A.1 of the Appendix.

From these comparisons, significant improvements are observed, proving that the method can improve the accuracy of pose estimation results from egocentric videos.

For the qualitative evaluation, the comparison between Mo^{2}Cap^2 and the proposed method (based on Mo^{2}Cap^2) is shown in Fig. 4.5. This proposed method also has the ability to estimate the global body pose, which is shown in Fig. 4.6. In Fig. 4.6, the accuracy of the global pose estimation is demonstrated by projecting the predicted global pose to an external camera.

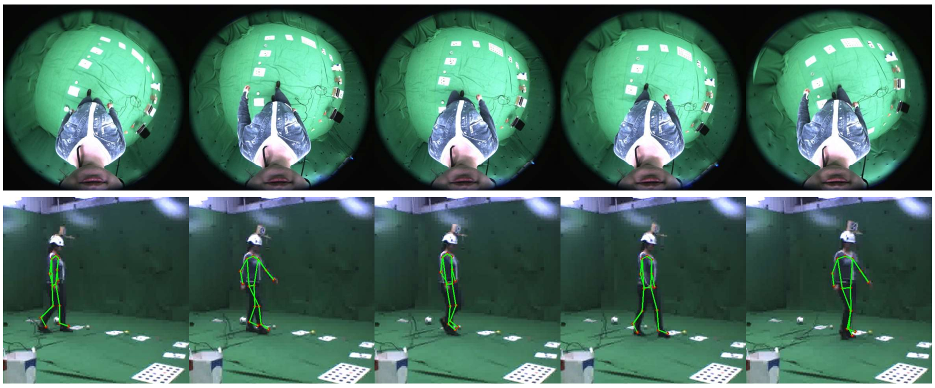


Figure 4.6: Global pose estimation results from a third-view camera. Top row: the input egocentric images, bottom row: the estimated 3D pose projected on an external camera.

4.3.5 *Ablation Study*

Further experiments are conducted to evaluate the effects of individual components of the approach. Mo^2Cap^2 is used as the local pose estimator for all ablation studies to ensure the results are comparable.

LOCAL/ GLOBAL POSE OPTIMIZER. In this experiment, to investigate the efficacy of the local and global optimizer, the method is evaluated after removing the local pose optimizer or the global pose optimizer from the entire pipeline. The results are shown in the 2nd and 3rd row of Table 4.2, demonstrating that both modules are important to the approach. The heatmap reprojection error in the local pose optimizer ensures that the optimized 3D pose conforms to the constraint of 2D predictions. The VAE prior in the global pose optimizer keeps the movement of body limbs in accordance with the global camera pose, thus improving both the global MPJPE and the local MPJPEs.

MOTION PRIORS. In order to validate the importance of motion priors, the performance of the optimization framework is tested without the motion priors by directly optimizing 3D pose P_{seq} with $E(P_{seq})$ rather than optimizing the VAE's latent vector z. The method is evaluated without motion prior on the new test dataset and shows one of the results in Fig. 4.7. In this figure, the human leg in the input image is severely occluded. The ambiguity of the image significantly reduced the accuracy of the single-frame pose estimation network. Without the motion prior, the optimization framework cannot resolve the ambiguity and the error is still large, while in the proposed method, the motion prior is able

Method	Global MPJPE	PA-MPJPE	BA-MPJPE
Mo^2Cap^2 + SLAM	141.8	102.3	74.46
w/o local optim.	134.7	96.33	70.77
w/o global optim.	123.1	84.99	64.10
w/o motion prior	128.1	92.31	68.10
w. GMM	125.0	90.12	67.50
w. single frame VAE	122.2	87.04	65.58
w. VAE in VIBE	126.7	86.48	66.46
w. VAE in MEVA	121.6	84.49	63.69
w. MLP based VAE	122.2	85.07	65.05
conventional reproj.	128.2	89.97	67.99
w_{vae}=1e-3	154.0	109.9	83.91
w_{vae}=5e-4	126.0	87.76	66.64
w_{vae}=1e-4	118.0	80.97	62.43
w_{vae}=1e-5	117.4	79.47	61.62
optimize P_{seq}	124.6	90.72	68.78
Mo^2Cap^2 + Proposed	**119.5**	**82.06**	**62.07**

Table 4.2: The quantitative results of ablation study.

to correct the estimated pose. The qualitative evaluation in the 4th row of Table 4.2 also confirms the claim in this chapter. With the motion prior, the spatio-temporal optimization framework is able to make pose predictions smoother and less ambiguous.

The VAE-based motion prior is also compared with the gaussian mixture model (GMM) prior used in Arnab et al. (2019), Bogo et al. (2016), and Kolotouros et al. (2019a)'s works and the single-frame VAE prior used in Pavlakos et al. (2019)'s work. When comparing with GMM prior, the GMM model is first trained with 8 Gaussians on the local pose sequence (local GMM) and the global pose sequence (global GMM) from the AMASS dataset. Then the local and global VAE in the method are substituted with the local and global GMM, and three MPJPEs are evaluated, as shown in the 5th row of Table 4.2. GMM prior performs worse since the VAE uses the neural network as a feature extractor, making it easier to capture priors. When comparing with single-frame-based VAE prior, a VAE network taking a single input pose is trained on the AMASS dataset, and the VAE in the local optimizer is substituted with the single-frame VAE. The evaluation result is shown in the 6th row

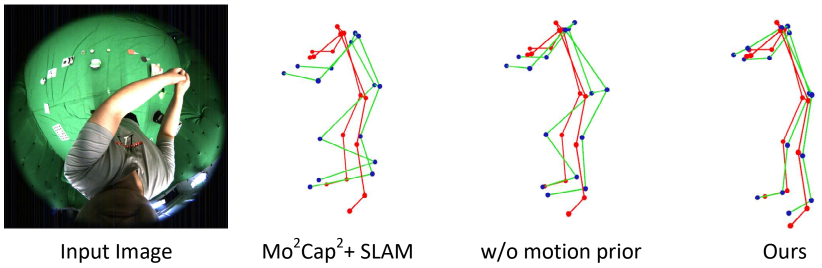


Figure 4.7: Comparison between the proposed method with and without motion prior. From left to right: input image, Mo^{2}Cap^2 + SLAM (green), the result without motion prior (green) and the one with motion prior (the result of proposed method) (green) overlaid on the ground truth (red).

of Table 4.2. The single-frame VAE cannot capture the prior over time, making it less effective than the sequential VAE proposed in this chapter.

CNN-BASED SEQUENTIAL VAE. This method uses the CNN-based sequential VAE rather than RNN-based VAE for better efficiency and optimization stability. To evaluate the advantage of the proposed method, the CNN-based sequential VAE in both the local and global optimizer is substituted with the VAEs in VIBE (Kocabas et al., 2020) or MEVA (Luo et al., 2020). The results are reported in the 7th to 9th rows of Table 4.2. The result proves that the CNN-based VAE outperforms others in terms of optimization accuracy, which can be attributed to a more stable optimization process. To demonstrate this, the $E(\mathcal{P}_{seq})$-iteration curve of the local pose optimization process (Sec. 4.2.2.2) is shown in Fig. 4.8, where RNN-based VAEs are less stable due to the gradient explosion issue. To show the efficiency of CNN-based VAE, the time for the optimization was evaluated. This method takes 195.7ms per 10-frame segment, while RNN-based VAE in VIBE and MEVA takes 552.1ms and 1139.4ms per segment respectively. The CNN-based VAE was also compared with multilayer perceptron (MLP) based VAE. According to Fig. 4.8 and the 10th row of Table 4.2, the MLP-based VAE performs worse since it is not designed to capture the temporal context of the pose sequence.

HEATMAP REPROJECTION ERROR. In this work the heatmap reprojection error is used during the optimization while a lot of previous works get the reprojection error by calculating the distance between estimated 2D joints and corresponding projected 3D joints (Arnab et al., 2019; Bogo et al., 2016; Pavlakos et al., 2019; Zanfir et al., 2020). To evaluate the advantage of heatmap reprojection error, the heatmap reprojection error in the pipeline is substituted with the conventional reprojection error in Bogo

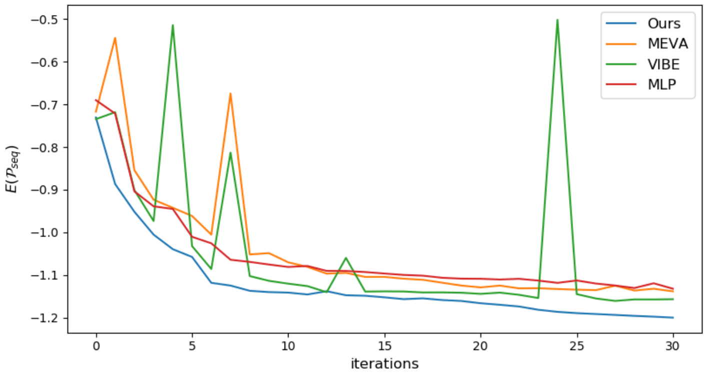


Figure 4.8: $E(\mathcal{P}_{seq})$-iteration curve of different VAEs. The proposed method gives the lowest error while keeping stable during optimization.

et al. (2016)'s work. In the qualitative evaluation shown in Fig. 4.9, the 2D pose estimation gives wrong results for the right-hand position while the ground truth hand position is still covered by the heatmap. The heatmap reprojection error can leverage such uncertainty in the heatmap and give better results than the conventional reprojection error. The quantitative results are also shown in the 10th row of Table 4.2. These results prove the advantage of the heatmap reprojection error.

WITH VAE PRIOR LOSS Different from previous optimization methods based on motion priors (Pavlakos et al., 2019; Zanfir et al., 2020), the prior error $E_{prior} = \|z\|_2$ is not applied in this method. That is because the prior error encourages the latent vector z closer to zero, which would make the pose stay close to a single mean pose, thus resulting in unnecessary bias. To prove this, the prior error is added with several different weights w_{vae} in the energy function, and the MPJPEs are shown in the 3rd to 6th row of Table 4.2. From the experimental result, all three errors rise as the prior weight increases. The errors converge to the proposed method when the prior weight approaches zero. This verifies the claim that the VAE prior error E_{prior} is harmful to the proposed optimization algorithm.

OPTIMIZATION OF P_{seq} In the optimization algorithm, the latent vector of VAE z is optimized, and the final prediction P_{seq} is obtained with the VAE decoder f_{dec}. An alternative optimization strategy is to optimize pose sequence P_{seq} directly, calculate the latent vector z with VAE encoder f_{enc} and incorporate prior term with $E_{prior} = \|z\|_2$. To compare these approaches, the direct optimization result is reported in the 7th row of Table 4.2. The result is better when the optimization is performed in the latent space, consistent with the previous research (Zanfir et al., 2020).

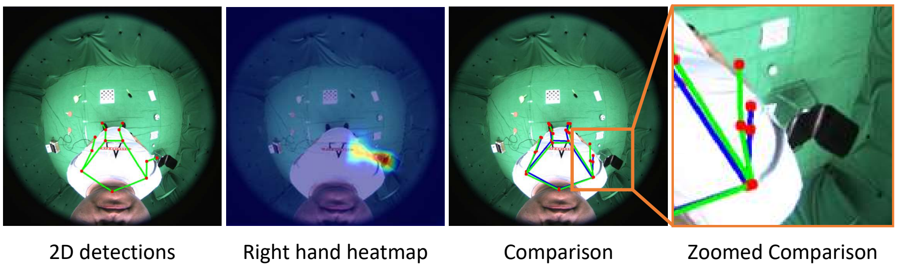


Figure 4.9: Comparison between heatmap reprojection error and conventional reprojection error. The result of heatmap reprojection error is in a green skeleton, and the result of conventional reprojection error is in a blue skeleton.

4.4 LIMITATIONS

As a common limitation of monocular SLAM methods, global camera pose estimation requires an environment with rich visual features. Featureless scenes such as white walls and green screens can lead to unreliable camera poses. This problem can be alleviated by using the RGBD-based SLAM method. Furthermore, the accuracy of this method remains sub-optimal, which is not comparable to methods based on external views. This is mostly caused by the inaccurate single-frame pose estimation method, which is used to provide the 2d heatmap supervision and the initial pose for optimization.

4.5 CONCLUSION

This chapter proposes a method for estimating global poses with a single head-mounted fisheye camera. This is achieved by employing a novel spatio-temporal optimization framework including: a sequential VAE to effectively capture the body motion prior; a global motion prior to ensure consistency between the local body motion and the camera poses; and a heatmap-based reprojection error term to leverage the uncertainty in predicted heatmaps. Experiments show that the proposed method outperforms state-of-the-art methods. This work will be referred to as **"GlobalEgoMocap"** in this thesis.

As mentioned in the limitation section, the performance of this method is limited by the single-frame egocentric pose estimation method. To solve this, the next chapter will present a new method for predicting accurate poses from in-the-wild egocentric images.

5 ESTIMATING EGOCENTRIC 3D HUMAN POSE IN THE WILD WITH EXTERNAL WEAK SUPERVISION

The last chapter presents GlobalEgoMocap, an optimization-based method for capturing egocentric motion in the global space. However, the performance of the GlobalEgoMocap is still constrained by the performance of single-frame pose estimation methods, including Mo^{2}Cap^2 (Xu et al., 2019) and *x*R-EgoPose (Tomè et al., 2019). These single-frame methods struggle to estimate the pose, especially from in-the-wild egocentric images, because they can only be trained on synthetic data due to the unavailability of large-scale in-the-wild egocentric datasets. Furthermore, these methods easily fail when the body parts are occluded by or interacting with the surrounding scene. To address the shortage of in-the-wild data, this Chapter proposes a large-scale in-the-wild egocentric dataset called *Egocentric Poses in the Wild Dataset (EgoPW dataset)*. This dataset is captured by a head-mounted fisheye camera and an auxiliary external camera, which provides an additional observation of the human body from a third-person perspective during training. This Chapter also presents a new egocentric pose estimation method, which can be trained on the new dataset with weak external supervision. Specifically, pseudo labels for the EgoPW dataset are first generated with a spatio-temporal optimization method by incorporating the external-view supervision. The pseudo labels are then used to train an egocentric pose estimation network. To facilitate network training, this Chapter proposes a novel learning strategy to supervise the egocentric features with the high-quality features extracted by a pretrained external-view pose estimation model. The experiments show that the EgoPW method predicts accurate 3D poses from a single in-the-wild egocentric image and outperforms the state-of-the-art methods both quantitatively and qualitatively.

5.1 INTRODUCTION

Egocentric motion capture using head- or body-mounted cameras has recently become popular because traditional motion capture systems with outside-in cameras have limitations when the person is moving around in a large space and thus restrict the scope of applications. Different from traditional systems, the egocentric motion capture system is mobile, flexible, and has no requirements on recording space, which enables

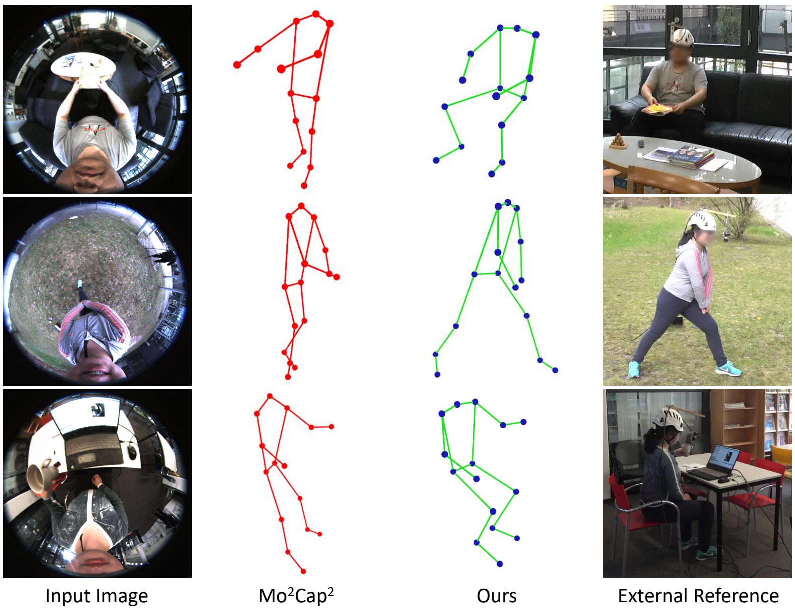


Figure 5.1: Compared with Mo^{2}Cap^2, the proposed method gets a more accurate egocentric pose from a single in-the-wild image, especially when the body parts are occluded. Note that the external images are only used for visualization, not the inputs to the method.

capturing a wide range of human activities for many applications, such as wearable medical monitoring, sports analysis, and *x*R.

This Chapter focuses on estimating the full 3D body pose from a single head-mounted fisheye camera. The most related works are Mo^{2}Cap^2 (Xu et al., 2019) and *x*R-egopose (Tomè et al., 2019). While these methods have produced compelling results, they are only trained on synthetic images as limited real data exists and, therefore, suffer from significant performance drops in real-world scenarios. Furthermore, these methods often struggle with the cases when parts of the human body are occluded by or interacting with the surrounding scene (see the Mo^{2}Cap^2 results in Fig. 5.1). This is due to the domain gap between synthetic and real data, but also due to their limited capability of handling occlusions.

To address the issue of the limited real egocentric data, a large-scale in-the-wild egocentric dataset called *Egocentric Poses in the Wild (EgoPW)* is captured. This is currently the largest egocentric in-the-wild dataset, containing more than 312k frames and covering 20 different daily activities in 8 everyday scenes. To obtain the supervision for the network training, one possibility is using a multi-view camera setup to capture training data with ground truth 3D body poses or apply multi-view weak supervision. However, this setup is impractical for recording in an environment with limited space (e.g. in the small kitchen shown in Fig. 5.3), which is a common recording scenario. Therefore, considering a trade-off between flexibility and 3D accuracy, a new device setup is proposed consisting of an egocentric camera and a single auxiliary external camera. The external view can provide additional supervision during training, especially for the highly occluded regions in the egocentric view (e.g. the lower body part).

To handle occlusions and estimate accurate poses, a new egocentric pose estimation method is proposed for training on the EgoPW dataset in a weakly supervised way. Specifically, a spatio-temporal optimization method is introduced to generate accurate 3D poses for each frame in the EgoPW dataset. The generated poses are further used as pseudo labels for training an egocentric pose estimation network (Xu et al., 2019). To improve the network performance, the training of the egocentric pose estimation network is facilitated with the extracted features from the external pose estimation network which has been trained on a large in-the-wild body pose dataset. Specifically, the feature extracted from these two views is enforced to be similar by fooling a discriminator not being able to detect which view the features are from. To further improve the performance of the pose estimation network, besides the EgoPW dataset, a synthetic dataset (Xu et al., 2019) is also leveraged to train the network and adopt a domain adaptation strategy to minimize the domain gap between synthetic and real data.

The proposed method is evaluated on the test data provided by Chapter 4 and Xu et al. (2019). This method significantly outperforms the state-of-the-art methods both quantitatively and qualitatively. Qualitative results are also shown on various in-the-wild images, demonstrating that this method can predict accurate 3D poses on very challenging scenes, especially when the body joints are seriously occluded (see the results in Fig. 5.1). To summarize, the contributions of this Chapter are presented as follows:

- A new method to estimate egocentric human pose with weak supervision from an external view, which significantly outperforms existing methods on in-the-wild data, especially when severe occlusions exist;
- A large in-the-wild egocentric dataset (EgoPW) captured with a head-mounted fisheye camera and an external camera;
- A new optimization method to generating pseudo labels for the in-the-wild egocentric dataset by incorporating the supervision from an external view;

- An adversarial method for training the network by learning the feature representation of egocentric images with external feature representation.

We recommend watching the video in https://people.mpi-inf.mpg.de/~jianwang/projects/egopw for better visualization. The EgoPW dataset is also available on the same webpage.

5.2 METHOD

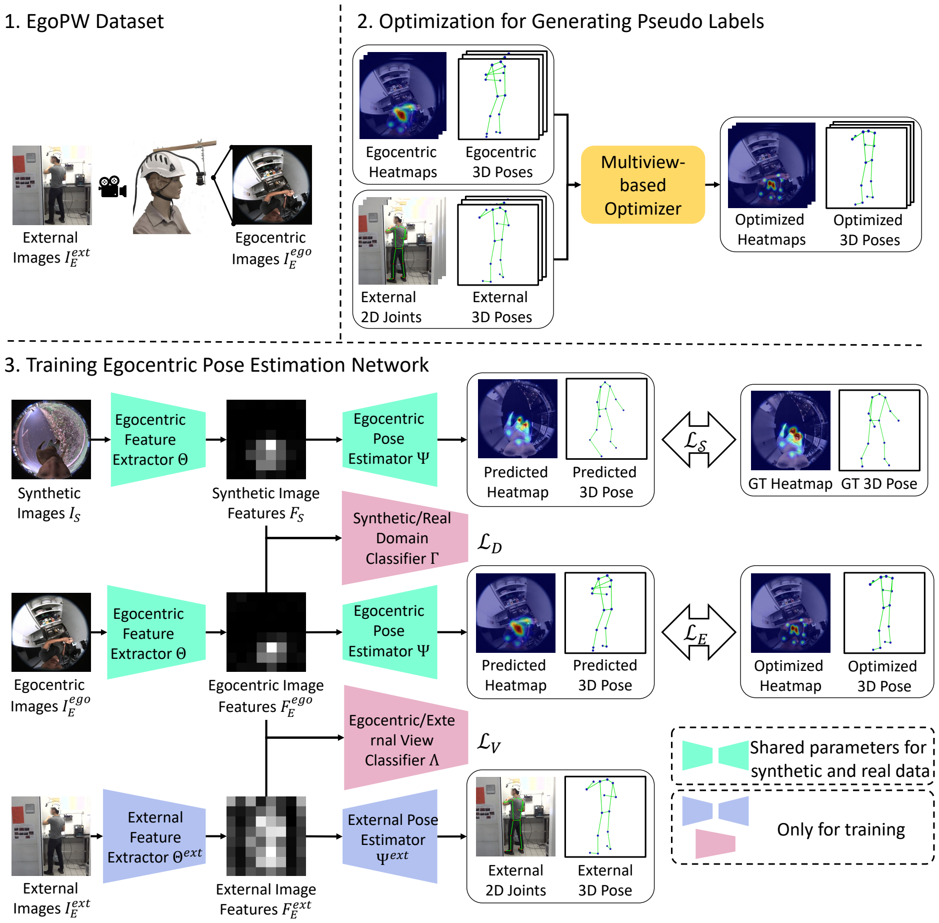

1. EgoPW Dataset
External Images I_E^{ext}
Egocentric Images I_E^{ego}
2. Optimization for Generating Pseudo Labels
Egocentric Heatmaps
Egocentric 3D Poses
External 2D Joints
External 3D Poses
Multiview-based Optimizer
Optimized Heatmaps
Optimized 3D Poses
3. Training Egocentric Pose Estimation Network
Synthetic Images I_S
Egocentric Feature Extractor Θ
Synthetic Image Features F_S
Egocentric Pose Estimator Ψ
Predicted Heatmap
Predicted 3D Pose
$\mathcal{L}_S$
GT Heatmap
GT 3D Pose
Synthetic/Real Domain Classifier Γ
$\mathcal{L}_D$
Egocentric Images I_E^{ego}
Egocentric Feature Extractor Θ
Egocentric Image Features F_E^{ego}
Egocentric Pose Estimator Ψ
Predicted Heatmap
Predicted 3D Pose
$\mathcal{L}_E$
Optimized Heatmap
Optimized 3D Pose
Egocentric/External View Classifier Λ
$\mathcal{L}_V$
Shared parameters for synthetic and real data
Only for training
External Images I_E^{ext}
External Feature Extractor Θ^{ext}
External Image Features F_E^{ext}
External Pose Estimator Ψ^{ext}
External 2D Joints
External 3D Pose

Figure 5.2: Overview of our method. The new EgoPW dataset (Sec. 5.2.1) is firstly collected, where pseudo labels are generated by a multi-view based optimization method (Sec. 5.2.2). Then, the proposed framework (Sec. 5.2.3) is trained, where the network is simultaneously trained with EgoPW datasets and synthetic data from Mo^{2}Cap^2. The egocentric network is further enforced to learn a better feature representation from the external view (Sec. 5.2.3.2) and finally bridge the gap between synthetic and real data with a domain classifier (Sec. 5.2.3.1).

This section proposes a new approach to train a neural network on the in-the-wild dataset with weak supervision from egocentric and external views. The overview of the proposed approach is illustrated in Fig. 5.2. A large-scale egocentric in-the-wild dataset is firstly captured, called *EgoPW*, which contains synchronized egocentric and external image sequences (Sec. 5.2.1). Next, the pseudo labels of the EgoPW dataset are generated with an optimization-based framework. This framework takes as input a sequence in a time window with B frames of egocentric images $\mathcal{I}_{seq}^{ego} = \{\mathcal{I}_1^{ego}, \ldots, \mathcal{I}_B^{ego}\}$ and external images $\mathcal{I}_{seq}^{ext} = \{\mathcal{I}_1^{ext}, \ldots, \mathcal{I}_B^{ext}\}$ and outputs egocentric 3D poses $\mathcal{P}_{seq}^{ego} = \{\mathcal{P}_1^{ego}, \ldots, \mathcal{P}_B^{ego}\}$ as the pseudo labels (Sec. 5.2.2).

Next, the egocentric pose estimation network is trained on the synthetic data from Mo^{2}Cap^2 (Xu et al., 2019) and on the EgoPW dataset with pseudo labels $\mathcal{P}_{seq}^{ego}$. In the training process, the feature representation from an on-the-shelf external pose estimation network (Xiao et al., 2018) is leveraged to enforce our egocentric network to learn a better feature representation in an adversarial way (Sec. 5.2.3.2). An adversarial domain adaptation strategy is also used to mitigate the domain gap between synthetic and real datasets (Sec. 5.2.3.1).

5.2.1 *EgoPW Dataset*

This section first describes the newly collected *EgoPW* dataset, which is the first large-scale in-the-wild human performance dataset captured by an egocentric camera and an external camera (Sony RX0), both synchronized. EgoPW contains a total of 318k frames, which are divided into 97 sequences of 10 actors in 20 clothing styles performing 20 different actions.

All personal data is collected with IRB approval. 3D poses are generated as pseudo labels using the egocentric and external images, which will be elaborated on later. In terms of size, the EgoPW dataset is larger than existing in-the-wild 3D pose estimation datasets, like 3DPW (Von Marcard et al., 2018), and has similar scale to the existing synthetic egocentric datasets, including the Mo^{2}Cap^2 (Xu et al., 2019) and the xR-egopose (Tomè et al., 2019) datasets.

5.2.2 *Optimization for Generating Pseudo Labels*

This section presents an optimization method based on Chapter 4 to generate pseudo labels for EgoPW. An input sequence is split into segments containing B consecutive frames. For the egocentric frames I_{seq}^{ego}, the 3D poses represented by 15 joint locations are estimated in the coordinate system of the egocentric camera (called "egocentric poses") $\widetilde{\mathcal{P}}_{seq}^{ego} = \{\widetilde{\mathcal{P}}_1^{ego}, \dots, \widetilde{\mathcal{P}}_B^{ego}\}$, $\widetilde{\mathcal{P}}_i^{ego} \in \mathbb{R}^{15\times3}$, and 2D heatmaps $H_{seq}^{ego} = \{H_1^{ego}, \dots, H_B^{ego}\}$ using the Mo^{2}Cap^2 method (Xu et al., 2019). Aside from egocentric poses, the transformation matrices between the egocentric camera poses of two adjacent frames $[R_{seq}^{SLAM} \mid t_{seq}^{SLAM}] = \{[R_1^2 \mid t_1^2], \dots, [R_{B-1}^B \mid t_{B-1}^B]\}$ are also estimated using ORB-SLAM2 (Mur-Artal and Tardós, 2017). For the external frames I_{seq}^{ext},the 3D poses (called "external poses") $\mathcal{P}_{seq}^{ext} = \{\mathcal{P}_1^{ext}, \dots, \mathcal{P}_B^{ext}\}$, $\mathcal{P}_i^{ext} \in \mathbb{R}^{15\times3}$ are estimated using VIBE (Kocabas et al., 2020) and 2D joints $\mathcal{J}_{seq}^{ext} = \{\mathcal{J}_1^{ext}, \dots, \mathcal{J}_B^{ext}\}$, $\mathcal{J}_i^{ext} \in \mathbb{R}^{15\times2}$ are obtained using openpose (Cao et al., 2017).

Next, following Chapter 4, a latent space is learned to encode an egocentric motion prior with a sequential VAE which consists of a CNN-based encoder f_{enc} and decoder f_{dec}. The egocentric pose is further optimized by finding a latent vector z such that the corresponding pose sequence $P_{seq}^{ego} = f_{dec}(z)$ minimizes the objective function:

$$\begin{aligned} E(\mathcal{P}_{seq}^{ego}, R_{seq}, t_{seq}) = {} & \lambda_R^{ego} E_R^{ego} + \lambda_R^{ext} E_R^{ext} + \lambda_J^{ego} E_J^{ego} \\ & + \lambda_J^{ext} E_J^{ext} + \lambda_T E_T + \lambda_B E_B \\ & + \lambda_C E_C + \lambda_M E_M. \end{aligned} \tag{5.1}$$

In this objective function, E_R^{ego}, E_J^{ego},E_T, and E_B are egocentric reprojection term, egocentric pose regularization term, motion smoothness regularization term and bone length regularization term, which are the same as those defined in Chapter 4. E_R^{ext}, E_J^{ext}, E_C, and E_M are the external reprojection term, external 3D body pose regularization term, camera pose consistency term, and camera matrix regularization term, which will be described later. Please see the Sec. B.5 in the Appendix B for a detailed definition of each term.

Note that since the relative pose between the external camera and the egocentric camera is unknown, the relative egocentric camera pose with respect to the external camera pose still needs to be optimized for each frame, i.e. the rotations $R_{seq} = R_1, \dots, R_B$ and translations $t_{seq} = t_1, \dots, t_B$.

EXTERNAL REPROJECTION TERM. In order to supervise the optimization process with the external 2D pose, the external reprojection term is proposed to minimize the difference between the projected 3D pose with the external 2D joints. The energy term is defined as:

$$E_R^{ext}(\mathcal{P}_{seq}^{ego}, R_{seq}, t_{seq}) = \sum_{i=1}^{B} \left\| \mathcal{J}_i^{ext} - K\left[R_i \mid t_i\right] \mathcal{P}_i^{ego} \right\|_2^2, \tag{5.2}$$

where K is the intrinsic matrix of the external camera; $[R_i \mid t_i]$ is the pose of the egocentric camera in the i th frame w.r.t the external camera position. In Eq. 5.2, the egocentric body pose $\mathcal{P}_i^{ego}$ is first projected to the 2D body pose in the external view with the egocentric camera pose $[R_i \mid t_i]$ and the intrinsic matrix K. Then, the projected body poses are compared with the 2D joints estimated by the openpose (Cao et al., 2017). Since the relative pose between the external camera and the egocentric camera is unknown at the beginning of the optimization, the egocentric camera pose $[R_i \mid t_i]$ is simultaneously optimized when optimizing the egocentric body pose $\mathcal{P}_{seq}^{ego}$. In order to make the optimization process converge faster, the egocentric camera pose $[R_i \mid t_i]$ is initialized with the Perspective-n-Point algorithm (Gao et al., 2003).

CAMERA POSE CONSISTENCY. The accurate 3D pose cannot be obtained only with the external reprojection term because the egocentric camera pose and the optimized body pose can be arbitrarily changed without violating the external reprojection constraint. To alleviate this ambiguity, the camera consistency term E_C is introduced as follows:

$$\begin{aligned} E_C(R_{seq}, t_{seq}) = \sum_{i=1}^{B-1} \Bigg\| \begin{bmatrix} R_i & t_i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_i^{i+1} & t_i^{i+1} \\ 0 & 1 \end{bmatrix} \\ - \begin{bmatrix} R_{i+1} & t_{i+1} \\ 0 & 1 \end{bmatrix} \Bigg\|_2, \end{aligned} \tag{5.3}$$

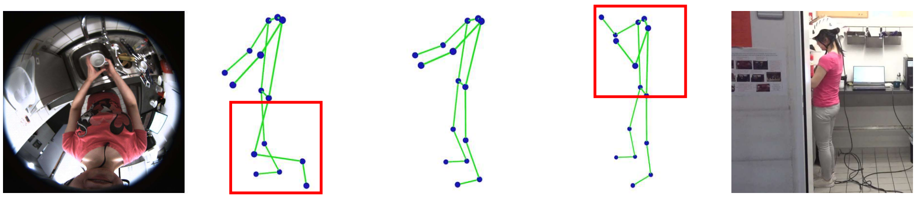
It enforces the egocentric camera pose at $(i+1)$ th frame $[R_{i+1} \mid t_{i+1}]$ to be consistent with the pose obtained by transforming the egocentric camera pose at the i th frame $[R_i \mid t_i]$ with the relative pose between the i th and $(i+1)$ th frame.

EXTERNAL 3D BODY POSE REGULARIZATION. Besides the external reprojection term, the external 3D body poses are also leveraged to supervise the optimization of the egocentric 3D body pose. the external 3D pose term which measures the difference between the external and the egocentric body poses after a rigid alignment is defined as follow:

$$E_J(\mathcal{P}_{seq}^{ego}, \mathcal{P}_{seq}^{ext}) = \sum_{i=1}^{B} \left\|\mathcal{P}_i^{ext} - \left[R_i^{pa} \mid t_i^{pa}\right] \mathcal{P}_i^{ego}\right\|_2^2, \tag{5.4}$$

where $\left[R_i^{pa} \mid t_i^{pa}\right]$ is the transformation matrix calculated with Procrustes analysis, which rigidly aligns the external 3D pose estimation $\mathcal{P}_i^{ext}$ and the egocentric 3D pose $\mathcal{P}_i^{ego}$.

By combining the body poses estimated from the egocentric view and external view, accurate pseudo labels can be finally reconstructed. As



(a) Egocentric Image (b) Only Egocentric (c) Our Pseudo Label (d) Only External (e) External Image

Figure 5.3: The pseudo label generation method combines the information from both the egocentric view and external view, therefore leading to more accurate pseudo labels (c). Only with the egocentric camera, the feet cannot be observed and well-tracked (b). Only with the external camera, the hands are occluded and result in the wrong result on the hand part (d).

shown in Fig. 5.3, the hands of the person are occluded in the external view, resulting in the tracking of the hands failing in the external view (Fig. 5.3, b), however, the hands can be clearly seen and tracked in the egocentric view (Fig. 5.3, d); on the other hand, the feet cannot be observed in the egocentric view and thus fail to be tracked in this view (Fig. 5.3, b), but can be easily viewed and tracked in the external view (Fig. 5.3, d). By joining the information from both views, accurate 3D poses can be accurately predicted as the pseudo labels (Fig. 5.3, c). Note that the external camera is only used for generating the pseudo labels. At test time, only the egocentric camera is used.

CAMERA MATRIX REGULARIZATION. The camera rotation matrix R_i is constrained to be orthogonal:

$$E_J(R_{seq}) = \sum_{i=1}^{B} \left\| R_i^T R_i - I \right\|_2^2 . \tag{5.5}$$

Different from previous single-view pose estimation methods which leverage the weak supervision from multiple views Iqbal et al., 2020; Kocabas et al., 2019; Rhodin et al., 2018; Wandt et al., 2021, this spatio-temporal optimization method generates the pseudo labels under the guidance of learned motion prior, making it robust to noisy and inaccurate 2D pose estimations which is common for the 2D pose estimation results from the egocentric view.

5.2.3 *Training Egocentric Pose Estimation Network*

Through the optimization framework in Sec. 5.2.2, the accurate 3D pose pseudo labels $\mathcal{P}_{seq}^{ego}$ can be obtained for each egocentric frame in the EgoPW dataset. The 3D pose pseudo labels are further processed into the 2D heatmap H_E and the distance between joints and egocentric camera D_E with the fisheye camera model (Scaramuzza and Ikeuchi, 2014) described in Chapter 3.

Afterward, a single-image-based egocentric pose estimation network is trained on both the synthetic dataset from Mo^{2}Cap^2 and the EgoPW dataset, as shown in the right part of Fig. 5.2. The pose estimation network contains a feature extractor Θ which encodes an image into a feature vector and a pose estimator Ψ which decodes the feature vector to 2D heatmaps and a distance vector. The 3D pose can be reconstructed from them with the fisheye camera model. Here, the synthetic dataset is noted as $S = \{I_S, H_S, D_S\}$ including synthetic images I_S along with their corresponding heatmaps H_S and distance labels D_S from Mo^{2}Cap^2 dataset. The EgoPW dataset is noted as $E = \{I_E^{ego}, H_E, D_E, I_E^{ext}\}$ including egocentric in-the-wild images I_E^{ego} along with pseudo heatmaps H_E, distance labels D_E and corresponding external images I_E^{ext}. During the training process, the egocentric pose estimation network is trained with two reconstruction loss terms and two adversarial loss terms. The reconstruction losses are defined as the mean squared error (MSE) between the predicted heatmaps/distances and heatmaps/distances from labels:

$$\begin{aligned} L_S &= \mathrm{mse}(\hat{H}_S, H_S) + \mathrm{mse}(\hat{D}_S, D_S) \\ L_E &= \mathrm{mse}(\hat{H}_E, H_E) + \mathrm{mse}(\hat{D}_E, D_E), \end{aligned} \tag{5.6}$$

where

$$\begin{aligned} \hat{H}_S, \hat{D}_S &= \Psi(F_S), F_S = \Theta(I_S); \\ \hat{H}_E, \hat{D}_E &= \Psi(F_E^{ego}), F_E^{ego} = \Theta(I_E^{ego}). \end{aligned} \tag{5.7}$$

Two adversarial losses are separately designed for learning egocentric feature representation and bridging the domain gap between synthetic and real datasets. These two losses are described as follows.

5.2.3.1 *Adversarial Domain Adaptation*

To bridge the domain gap between the synthetic and real data domains, following Tzeng et al. (2017), this chapter introduces an adversarial discriminator Γ which takes as input the feature vectors extracted from a synthetic image and an in-the-wild image, and determines if the feature is extracted from an in-the-wild image. The adversarial discriminator Γ is trained with a cross-entropy loss:

$$\mathcal{L}_D = -E[\log(\Gamma(F_S))] - E[\log(1 - \Gamma(F_E^{ego}))]. \tag{5.8}$$

Once the discriminator Γ has been trained, the feature extractor Θ maps the images from different domains to the same feature space such that the classifier Γ cannot tell if the features are extracted from synthetic images or real images. Therefore, the pose estimator Ψ can predict more accurate poses for the in-the-wild data.

5.2.3.2 *Supervising Egocentric Feature Representation with External View*

Although the EgoPW training dataset is large, the variation of identities in the dataset is still relatively limited (20 identities) compared with the existing large-scale external-view human datasets (thousands of identities). Generally speaking, the representations learned with these external-view datasets are of higher quality due to the large diversity of the datasets. To improve the generalizability of our network and prevent overfitting to the training identities, the egocentric representation in the proposed network is further supervised by leveraging the high-quality third-person view features. From a transfer learning perspective, although following Mo^{2}Cap^2 (Xu et al., 2019), the egocentric network is pretrained on the third-person-view datasets, it can easily "forget" the learned knowledge while being finetuned on the synthetic dataset. The supervision from third-person-view features can prevent the egocentric features from deviating too much from those learned from large-scale real human images.

However, directly minimizing the distance between egocentric features F_E^{ego} and external features F_E^{ext} will not enhance the performance since the intermediate features of the egocentric and external view should be different from each other due to significant difference on the view direction and camera distortions. To tackle this issue, the adversarial training strategy is applied to align the feature representation from egocentric and external networks. Specifically, an adversarial discriminator Λ is introduced to take the feature vectors extracted from an egocentric image and the corresponding in-the-wild images and predict if the feature is from egocentric or external images. The adversarial discriminator Λ is trained with a cross-entropy loss:

$$L_V = -E[\log(\Lambda(F_E^{ego}))] - E[\log(1 - \Lambda(F_E^{ext}))], \tag{5.9}$$

where $F_E^{ext} = \Theta^{ext}(I_E^{ext})$ and Θ^{ext} is the feature extractor of external pose estimation network that shares exactly the same architecture as the egocentric pose estimation network. The parameters of the features extractor Θ^{ext} and the pose estimator Ψ^{ext} of the external pose estimation network are obtained from the pretrained model in Xiao et al. (2018)'s work and remain fixed during the training process.

Note that the deep layers of the pose estimation network usually represent the global semantic information of the human body (Chu et al., 2017), the output feature of the 4th res-block of ResNet-50 network (He

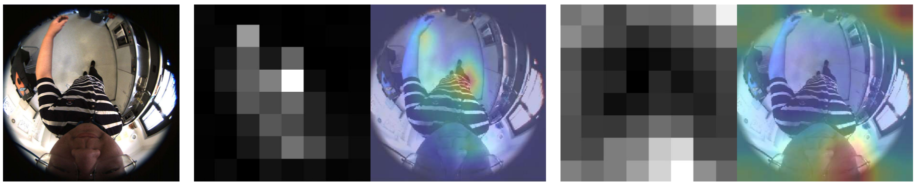

a) Input
b) With external feature supervision
c) Without external feature supervision

Figure 5.4: The visualization of features with (b) or without (c) the adversarial supervision from external features. By supervising the training of the egocentric network with the feature representation from an external view, the egocentric network is able to focus on extracting the semantic features of the human body.

et al., 2016) is used as the input to the discriminator Λ. Furthermore, the spatial position of the joints is quite different in the egocentric view and the external view, which will make the discriminator Λ easily learn the difference between egocentric and external features. To solve this, an average pooling layer is introduced in the discriminator Λ to spatially aggregate features, thus further eliminating the influence of spatial distribution between egocentric and external images. Please refer to the Sec. B.4 in the Appendix B for further details.

During the training process, the egocentric pose estimation network is trained to produce the features F_E^{ego} to fool the domain classifier Λ such that it cannot distinguish whether the feature is from an egocentric or external image.

To achieve this, the egocentric network learns to pay more attention to the relevant parts of the input image, i. e., the human body, which is demonstrated in Fig. 5.4.

5.3 EXPERIMENTS

5.3.1 *Datasets*

The finetuned network is quantitatively evaluated on the real-world datasets from Mo^{2}Cap^2 (Xu et al., 2019) and Chapter 4. The real-world dataset in Mo^{2}Cap^2 (Xu et al., 2019) contains 2.7k frames of two people captured in indoor and outdoor scenes, and that in Chapter 4 contains 12k frames of two people captured in the studio. To measure the accuracy of the pseudo labels, the optimization method (Sec. 5.2.2) is evaluated only on the dataset from Chapter 4 since the Mo^{2}Cap^2 dataset does not include the external view.

To evaluate the proposed method on the in-the-wild data, a qualitative evaluation is also conducted on the test set of the EgoPW dataset. The

Method	PA-MPJPE	BA-MPJPE
Mo^2Cap^2	102.3	74.46
xR-egopose	112.0	87.20
Method in Chapter 4	83.40	63.88
VIBE (Kocabas et al., 2020)	68.13	52.99
Proposed Optimizer	**57.19**	**46.14**

Table 5.1: The accuracy of pseudo labels on test dataset in Chapter 4. Utilizing both egocentric and external views, the body poses from the proposed optimization method (Sec. 5.2.2) are more accurate and can serve as better pseudo labels.

EgoPW dataset has been made publicly available, and more details and comparisons to other datasets are included in Sec. B.3 of Appendix B.

5.3.2 *Evaluation Metrics*

The results of the proposed method, as well as other baseline methods, are measured using two metrics, PA-MPJPE and BA-MPJPE. For **PA-MPJPE**, the estimated pose $\hat{\mathcal{P}}$ of each frame is rigidly aligned to the ground truth pose $\mathcal{P}$ using Procrustes analysis (Kendall, 1989). In order to eliminate the influence of the body scale, the **BA-MPJPE** scores are also reported. In this metric, the bone lengths of each predicted body pose $\hat{\mathcal{P}}$ and ground truth body pose $\mathcal{P}$ are first resized to the bone length of a standard skeleton. Then, the PA-MPJPE is calculated between the two resulting poses.

5.3.3 *Pseudo Label Generation*

In this paper, the pseudo labels are first generated with the optimization framework (Sec. 5.2.2) and further leveraged to train the network (Sec. 5.2.3). Thus, pseudo-labels with higher accuracy generally lead to better network performance. In this experiment, the accuracy of pseudo labels is evaluated on Wang *et al.*'s dataset and the results are shown in Table 5.1. This table shows that the proposed method outperforms all the baseline methods by leveraging both the egocentric view and external view during optimization. Note that though compared in Table 5.1, it is impossible to use any external-view-based pose estimation method, e. g. VIBE (Kocabas et al., 2020) and 3DPW (Von Marcard et al., 2018), for training the egocentric pose estimation network. This is because the

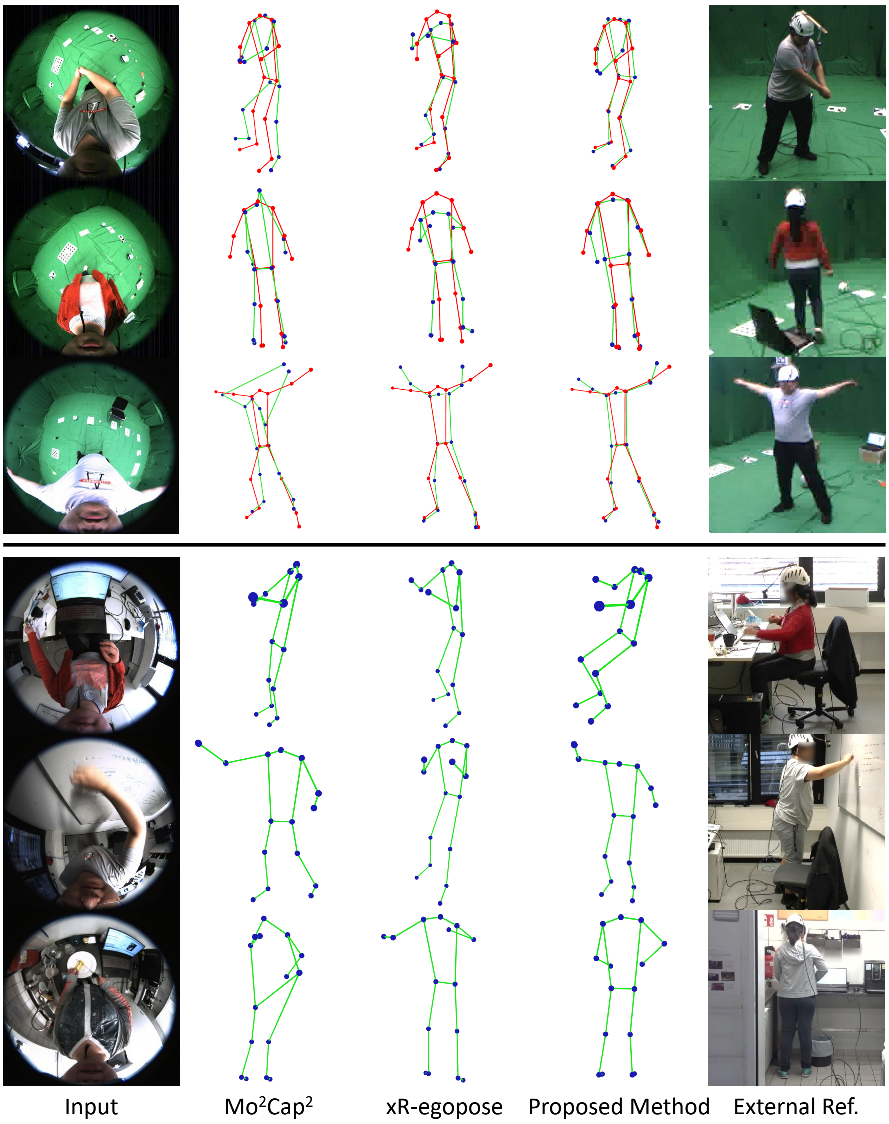


Figure 5.5: Qualitative comparison between the proposed method and the state-of-the-art methods. From left to right: input image, Mo^2Cap^2 result, xR-egopose result, the result of the proposed method, and external image. The ground truth pose is shown in red. Note that the external images are not used during inference. The input images on the top part are from the test dataset of Chapter 4, while the images on the bottom part are from the EgoPW test sequences.

Method	PA-MPJPE	BA-MPJPE
Chapter 4's test dataset		
Rhodin et al. (2018)	89.67	73.56
Mo^{2}Cap^2 (Xu et al., 2019)	102.3	74.46
xR-egopose (Tomè et al., 2019)	112.0	87.20
Ours	**81.71**	**64.87**
Mo^{2}Cap^2 test dataset		
Rhodin et al. (2018)	97.69	76.92
Mo^{2}Cap^2 (Xu et al., 2019)	91.16	70.75
xR-egopose (Tomè et al., 2019)	86.85	66.54
Ours	**83.17**	**64.33**

Table 5.2: Performance of the egocentric pose estimation network (Sec. 5.2.3) on Wang *et al.*'s test dataset and Mo^{2}Cap^2 test dataset (Xu et al., 2019). This method outperforms the state-of-the-art methods, Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019), on both metrics.

relative pose between the external and egocentric camera is unknown, making it impossible to obtain the egocentric body pose only from the external view. Compared with the proposed optimization approach, the method in Chapter 4 performs worse due to the lack of external-view supervision.

5.3.4 *Comparisons on 3D Pose Estimation*

In this section, the egocentric pose estimation network trained in Sec. 5.2.3 is compared with previous single-frame-based methods on the test dataset from Chapter 4 under the "Wang *et al.*'s test dataset" in Table 5.2. Since the code or the predictions of xR-egopose are not publicly available, the reimplementation of xR-egopose is used instead. On this dataset, the proposed method outperforms Mo^{2}Cap^2 by 20.1% and xR-egopose by 27.0% respectively. The proposed method is also compared with previous methods on the Mo^{2}Cap^2 test dataset and the results are shown under the "Mo^{2}Cap^2 test dataset" in Table 5.2. On the Mo^{2}Cap^2 test dataset, the proposed method performs better than Mo^{2}Cap^2 and xR-egopose by 8.8% and 4.2%, respectively.

Method	PA-MPJPE	BA-MPJPE
w/o external view	90.05	68.99
w/o learning representation	85.46	67.01
w/o domain adaptation	84.22	66.48
Unsupervised DA	91.56	69.17
Ours	**81.71**	**64.87**

Table 5.3: The quantitative results of ablation study.

The results in Table 5.2 show that our approach outperforms all previous methods on the single-frame egocentric pose estimation task. More quantitative results on each type of motion are available in Sec. B.1 of Appendix B. For the qualitative comparison, the results of our method on the studio dataset and in-the-wild dataset are shown in Fig. 5.5. The method performs much better compared with Mo^{2}Cap^2 and xR-egopose, especially for the in-the-wild cases where the body parts are occluded. Please refer to Sec. B.2 of Appendix B for more qualitative results.

The proposed method is also compared with Rhodin et al. (2018)'s method, which uses weak supervision from multiple views to supervise the training of a single view pose estimation network. In the EgoPW dataset, there is only one egocentric and one external view. Thus, the 3D pose estimation network for the external view is fixed, and only the egocentric pose estimation network is trained. Following Rhodin et al. (2018), the predictions from the egocentric and external view are aligned with Procrustes analysis, and the loss proposed by Rhodin et al. (2018) is calculated. The results in Table 5.2 show the proposed method performs better. This is mainly because the proposed spatio-temporal optimization method predicts accurate and temporally stable 3D poses as pseudo labels, while other methods suffer from inaccurate egocentric pose estimations.

5.3.5 *Ablation Study*

SUPERVISION FROM THE EXTERNAL VIEW. In this work, the external view is introduced as supervision for training the network. The external view enables generating accurate pseudo labels, especially when the human body parts are occluded in the egocentric view but can be observed in the external view. Without the external view, the obtained pseudo labels are less accurate and will further affect the network performance. In order to demonstrate this, The 3D poses are first generated

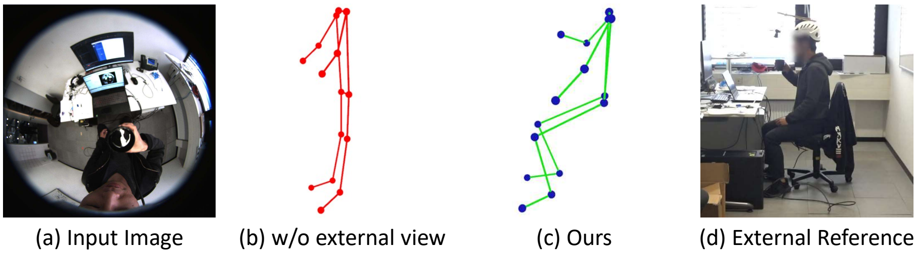


Figure 5.6: The results of the proposed method with (c) and without external view (b). The network cannot predict accurate poses for the occluded cases without external view supervision. The external view is only for visualization and not used for predicting the pose.

as pseudo labels with the method in Chapter 4, i. e.without any external supervision, and then train the pose estimation network on these new pseudo labels. The result is shown in the "w/o external view" row of Table 5.3. The qualitative results with and without external-view supervision are also shown in Fig. 5.6. Both the qualitative and quantitative results demonstrate that with external supervision, the performance of the pose estimation network is significantly better, especially in occluded cases.

LEARNING EGOCENTRIC FEATURE REPRESENTATION AND BRIDGING THE DOMAIN GAP WITH ADVERSARIAL TRAINING. In this Chapter, the pose estimation network is trained with two adversarial components to learn the feature representation of the egocentric human body (Sec. 5.2.3.2) and bridge the domain gap between synthetic and real images (Sec. 5.2.3.1). To demonstrate the effectiveness of both modules, the domain classifier Λ is removed from our training process, and the results are shown in the row of "w/o learning representation" in Table 5.3. The domain classifier Γ was also removed, and the network was trained without L_D. The quantitative results are shown in the row of "w/o domain adaptation" in Table 5.3. After moving any of the two components, the proposed method suffers from a performance drop, which demonstrates the effectiveness of both the feature representation learning module and the domain adaptation module.

COMPARISON WITH ONLY USING UNSUPERVISED DOMAIN ADAPTATION. In this experiment, our approach is compared with the unsupervised adversarial domain adaptation method Tzeng et al., 2017 that is commonly used for transfer learning tasks. The network is trained only with the L_S and L_D in the adversarial domain adaptation module (Sec. 5.2.3.1). The results are shown in the "Unsupervised DA" of Table 5.3. The proposed approach outperforms the unsupervised domain adaptation method due to the high-quality pseudo labels.

5.4 LIMITATIONS

The accuracy of pseudo labels in this method is constrained by the in-the-wild capture system, which contains only one egocentric view and one external view. The use of inaccurate pseudo labels also limits the network's performance. One future solution is to fuse different sensors, including IMUs and external-view depth cameras, for capturing the ground truth for the in-the-wild dataset.

Another limitation of this method is that the method is still trained on the synthetic dataset from Mo^2Cap^2 (Xu et al., 2019). The images in this dataset are far from realistic, suffering from a large domain gap with the real-world egocentric human motion dataset. It is more challenging for the domain adaptation strategy in the proposed method to bridge the domain gap between synthetic and real-world datasets. Future methods can synthesize realistic datasets with high-quality articulated human models like renderpeople model (*RenderPeople* n.d.) or SMPL model Loper et al., 2015 with clothing simulations Black et al., 2023.

The proposed method still suffers from the self-occlusion problem. To solve this, future research can constrain the egocentric human motion capture on the explicit scene geometry. The scene geometry will provide rich clues by avoiding the human body penetrating the scene or floating in the air.

5.5 CONCLUSION

This paper proposes a new approach for egocentric human pose estimation using a single head-mounted fisheye camera. A new in-the-wild egocentric dataset (EgoPW) was collected, and a new optimization method was designed to generate accurate egocentric poses as pseudo labels. The egocentric pose estimation network is then supervised with the pseudo labels and the features from the external network. The experiments show that this method outperforms all of the state-of-the-art methods both qualitatively and quantitatively and it also works well under severe occlusion. In the following chapters, we will refer to the work presented in this chapter as **"EgoPW"**.

As mentioned in the limitation section, this proposed method still suffers from the self-occlusion issue. The next chapter tries to solve this by first estimating the scene geometry from a single egocentric frame and then predicting the egocentric pose by combining the image features and scene geometry.

6 SCENE-AWARE EGOCENTRIC 3D HUMAN POSE ESTIMATION

Egocentric 3D human pose estimation with a single head-mounted fisheye camera has recently gained attention due to its numerous applications in virtual and augmented reality. Existing methods, including the ones from previous chapters, still struggle with challenging poses where the human body is highly occluded or closely interacting with the scene. To address this issue, a scene-aware egocentric pose estimation method is proposed, guiding the prediction of the egocentric pose with scene constraints.

An egocentric depth estimation network is first introduced to predict the scene depth map from a wide-view egocentric fisheye camera while mitigating the occlusion of the human body with a depth-inpainting network. Next, a scene-aware pose estimation network is proposed, projecting the 2D image features and estimated depth map of the scene into a voxel space and regressing the 3D pose with a V2V network. The voxel-based feature representation provides a direct geometric connection between 2D image features and scene geometry, further facilitating the V2V network to constrain the predicted pose based on the estimated scene geometry.

To enable the training of these networks, a synthetic dataset called EgoGTA and an in-the-wild dataset based on EgoPW, called EgoPW-Scene, were generated. The experimental results of the new evaluation sequences show that the predicted 3D egocentric poses are accurate and physically plausible in terms of human-scene interaction, demonstrating that this method outperforms state-of-the-art methods both quantitatively and qualitatively.

6.1 INTRODUCTION

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

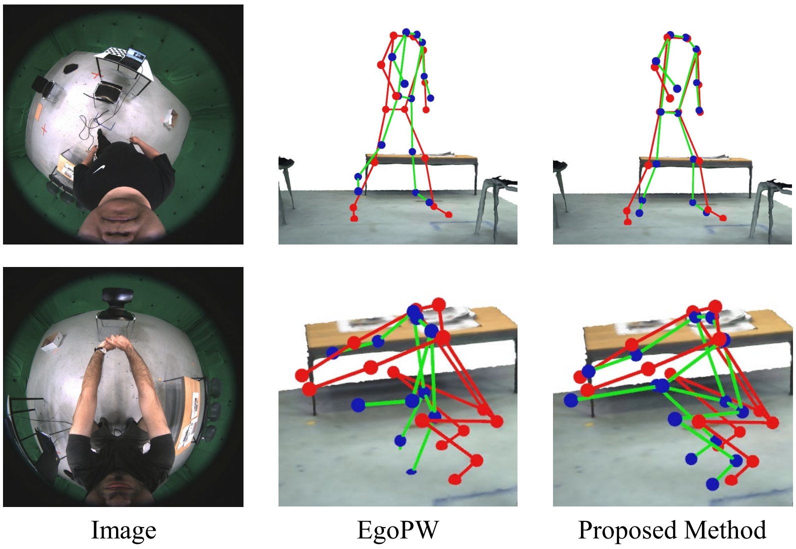


Figure 6.1: Previous egocentric pose estimation methods like EgoPW (Chapter 5) predict body poses that may suffer from body floating issues (the first row) or body-environment penetration issues (the second row). This method predicts accurate and plausible poses complying with the scene constraints. The red skeletons are the ground truth poses and the green skeletons are the predicted poses.

This Chapter proposes a method to estimate the full 3D body pose from a single head-mounted fisheye camera. A number of works have been proposed, including Mo^{2}Cap^2 Xu et al., 2019, *x*R-egopose Tomè et al., 2019, the method in Chapter 4 and EgoPW in Chapter 5. These methods have made significant progress in estimating egocentric poses. However, when taking account of the interaction between the human body and the surrounding environment, they still suffer from artifacts that contrast the physics plausibility, including body-environment penetrations or body floating (see the EgoPW results in Fig. 6.1), which is mostly ascribed to the ambiguity caused by the self-occluded and highly distorted human body in the egocentric view. This problem will render restrictions on subsequent applications including action recognition, human-object interaction recognition, and motion forecasting.

To address this issue, this Chapter proposes a scene-aware pose estimation framework that leverages the scene context to constrain the prediction of an egocentric pose. This framework produces accurate and physically plausible 3D human body poses from a single egocentric image, as illustrated in Fig. 5.1. Thanks to the wide-view fisheye camera mounted on the head, the scene context can be easily obtained even with only one egocentric image. To this end, an egocentric depth estimator is trained to predict the depth map of the surrounding scene. To mitigate the occlusion caused by the human body, the depth map including the visible human is predicted, and a depth-inpainting network is used to recover the depth behind the human body.

Next, the projected 2D pose features and scene depth are combined in a common voxel space, and the 3D pose heatmaps are regressed with a V2V network Moon et al., 2018. The 3D voxel representation projects the 2D poses and depth information from the distorted fisheye camera space to the canonical space, and further provides direct geometric connection between 2D image features and 3D scene geometry. This aggregation of 2D image features and 3D scene geometry facilitates the V2V network to learn the relative position and potential interactions between the human body joints and the surrounding environment and further enables the prediction of plausible poses under the scene constraints.

Since no available dataset can be used for train these networks, this Chapter proposes EgoGTA, a synthetic dataset based on the motion sequences of GTA-IM Cao et al., 2020, and EgoPW-Scene, an in-the-wild dataset based on EgoPW (Chapter 5. Both of the datasets contain body pose labels and scene depth map labels for each egocentric frame.

To better evaluate the relationship between the estimated egocentric pose and scene geometry, a new test dataset containing ground truth joint positions in the egocentric view was collected. The evaluation results on the new dataset, along with results on datasets in Chapter 4 and Mo^{2}Cap^2 Xu et al., 2019 demonstrate that the proposed method significantly outperforms existing methods both quantitatively and qualitatively. The proposed method is also qualitatively evaluated on in-the-wild images. The predicted 3D poses are accurate and plausible even in challenging real-world scenes. To summarize, the contributions in this Chapter are listed as follows:

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

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

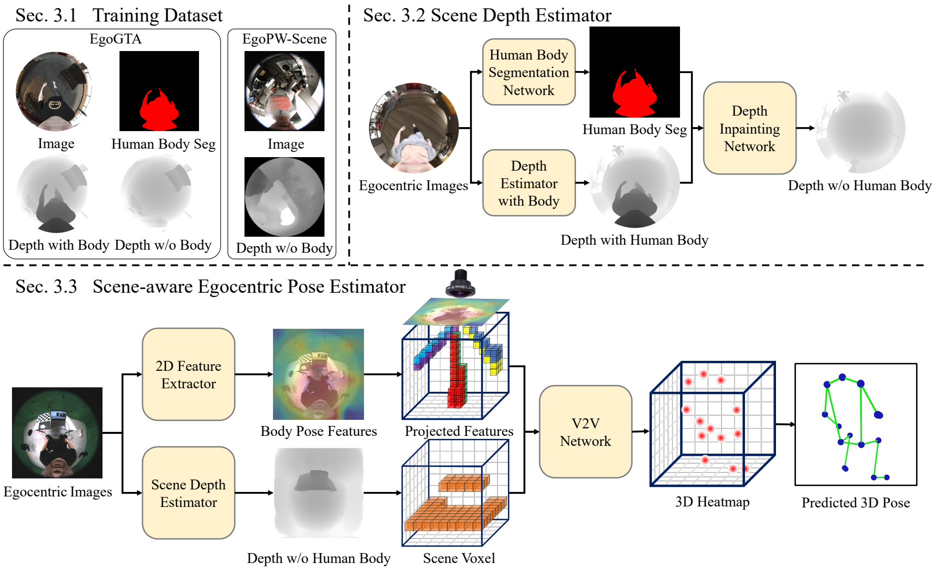


Figure 6.2: Overview of the proposed method. First, the synthetic training dataset EgoGTA and the in-the-wild training dataset EgoPW-Scene are rendered. Both datasets contain egocentric depth maps for subsequent training process (Sec. 6.2.1). Next, an egocentric scene depth estimator is trained to predict a depth map without the human body and a depth inpainting network (Sec. 6.2.2). Finally, the 2D body pose features and scene depth map are combined into a common voxel space. The 3D body pose heatmaps are regressed from the voxel space with a V2V network and the final pose prediction is obtained with soft-argmax (Sec. 6.2.3).

previous approaches and generates plausible poses considering the scene context.

6.2 method

A new method is proposed for predicting accurate egocentric body pose by leveraging the estimated scene geometry. An overview of this method is shown in Fig. 6.2. To train the scene-aware network, a synthetic dataset based on the GTA-IM dataset (Cao et al., 2020), called EgoGTA, and an in-the-wild dataset based on the EgoPW dataset (Chapter 5), called EgoPW-Scene (Sec. 6.2.1), are first generated. Next, a depth estimator is trained to estimate the geometry of the surrounding scene and the depth-inpainting network is introduced to estimate the depth behind the human body (Sec. 6.2.2). Finally, the 2D features and scene geometry are combined in a common voxel space and predict the egocentric pose with a V2V network (Moon et al., 2018) (Sec. 6.2.3).

6.2.1 *Training Dataset*

Although many training datasets for egocentric pose estimation (Tomè et al., 2019; Xu et al., 2019) have been proposed, they cannot yet train the scene-aware egocentric pose estimation network due to the lack of scene geometry information. To address this, the EgoGTA dataset and EgoPW-Scene dataset are introduced (both have been made publicly available). These datasets contain pose labels and depth maps of the scene for each egocentric frame, facilitating the training process. Examples from both datasets are shown in Fig. 6.2.

6.2.1.1 *EgoGTA Dataset*

To obtain precise ground truth human pose and scene geometry for training, a new synthetic egocentric dataset based on GTA-IM (Cao et al., 2020), which contains various daily motions and ground truth scene geometry, is devised. First, the SMPL-X model is fitted on the 3D joint trajectories from GTA-IM. Next, a virtual fisheye camera is attached to the forehead of the SMPL-X model, and images, semantic labels, and depth maps of the scene with and without the human body are rendered. In total, 320 K frames are obtained in 101 different sequences, each with a different human body texture. Here, the EgoGTA dataset is denoted as $\mathcal{S}_G = \{I_G, S_G, D_G^B, D_G^S, P_G\}$, including synthetic images I_G and their corresponding human body segmentation maps S_G, depth map with human body D_G^B, depth map of the scene without human body D_G^S, and egocentric pose labels P_G.

6.2.1.2 *EgoPW-Scene Dataset*

To generalize the method to real-world data, the EgoPW dataset (Chapter 5) is also extended as the training dataset. First, the scene geometry is reconstructed from the egocentric image sequences of the EgoPW training dataset using a Structure-from-Motion (SfM) algorithm (Hartley and Zisserman, 2003). This step provides a dense reconstruction of the background scene. The global scale of the reconstruction is recovered from known objects present in the sequences, such as laptops and chairs. The depth maps of the scene are further rendered in the egocentric perspective based on the reconstructed geometry. The EgoPW-Scene dataset contains 92 K frames in total, which are distributed in 30 sequences performed by 5 actors. The number of frames in the EgoPW-Scene dataset is less than the frames of the EgoPW dataset since SfM fails on some sequences. Here, the EgoPW-Scene dataset is denoted as $\mathcal{S}_E = \{I_E, D_E^S, P_E\}$, including in-the-wild images I_E and their corresponding depth map of the scene without human body D^S_E, and egocentric pose labels P_E.

6.2.2 *Scene Depth Estimator*

This section proposes a depth estimation method to capture the scene geometry information from an egocentric perspective. Available depth estimation methods Fu et al. (2018), Hu et al. (2019), and Lee et al. (2019) can only generate depth maps with the human body, but are not able to infer the depth information behind the human, i. e., the background scene depth. However, the area occluded by the human body, e. g. the areas of foot contact, are crucial for generating plausible poses, as demonstrated in Sec. 6.3.4. To predict the depth map of the scene behind the human body, a two-step approach is adopted. First, the depth map, including the human body, and the semantic segmentation of the human are estimated using two separate models. Then, a depth inpainting network is used to recover the depth behind the human body. This two-step strategy is necessary because the visual evidence of the human in the RGB images is too strong to be ignored by the depth estimator, making it easier to train the scene depth estimation as separate tasks.

First, the depth estimator network $\mathcal{D}$ is trained. It takes as input a single egocentric image I and predicts the depth map with human body $\hat{D}^B$. The network architecture of $\mathcal{D}$ is the same as Hu et al. (2019)'s work. To minimize the influence of the domain gap between synthetic and real data, the network is initially trained on the NYU-Depth V2 dataset (Nathan Silberman and Fergus, 2012) following Hu et al. (2019)'s work, and further fine-tuned on the EgoGTA dataset.

Next, the segmentation network $\mathcal{S}$ is trained for segmenting the human body. The network takes the egocentric image I as input and predicts the segmentation mask for the human body $\hat{S}$ as output. Following Yuan et al. (2020a), HRNet is adopted as the segmentation network. Similarly, to reduce the domain gap, the network is pretrained on the LIP dataset (Gong et al., 2017) and finetuned on the EgoGTA dataset. The networks $\mathcal{D}$ and $\mathcal{S}$ are not trained on the EgoPW-Scene dataset since the dataset lacks the ground truth segmentation maps and depth maps with the human body.

Finally, a depth inpainting network $\mathcal{G}$ is proposed for generating the final depth map of the scene without a human body. The masked depth map $\hat{D}^M = (1 - \hat{S}) \odot \hat{D}^B$ is generated as a Hadamard product between the background segmentation and the depth map with the human body. Then, the masked depth map $\hat{D}^M$ and the segmentation mask $\hat{S}$ are fed into the inpainting network $\mathcal{G}$, which predicts the final depth map $\hat{D}^S$. The inpainting network $\mathcal{G}$ is trained and the depth estimation network

$\mathcal{D}$ is finetuned on both the EgoGTA and EgoPW-Scene datasets. During training, the differences between the predicted depth maps and the ground truth depth of the background scene are penalized with L^S, and consistency of the depth map in the non-human body regions is maintained with L^C. Specifically, the loss function is defined as follows:

$$\begin{aligned} L &= \lambda^S L^S + \lambda^C L^C, \quad \text{with} \\ L^S &= \left\| \hat{D}_G^S - D_G^S \right\|_2^2 + \left\| \hat{D}_E^S - D_E^S \right\|_2^2, \quad \text{and} \\ L^C &= \left\| (\hat{D}_G^S - \hat{D}_G^B)(1 - \hat{S}_G) \right\|_2^2 \\ &\quad + \left\| (\hat{D}_E^S - \hat{D}_E^B)(1 - \hat{S}_E) \right\|_2^2, \end{aligned} \tag{6.1}$$

where

$$\begin{aligned} \hat{D}_G^S &= \mathcal{G}(\hat{D}_G^M, \hat{S}_G); & \hat{D}_E^S &= \mathcal{G}(\hat{D}_E^M, \hat{S}_E); \\ \hat{D}_G^B &= \mathcal{D}(I_G); & \hat{D}_E^B &= \mathcal{D}(I_E); \\ \hat{S}_G &= \mathcal{S}(I_G); & \hat{S}_E &= \mathcal{S}(I_E), \end{aligned} \tag{6.2}$$

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

6.2.3 *Scene-aware Egocentric Pose Estimator*

This pose estimator relies on the prior knowledge that human bodies are mostly in contact with the scene. However, explicitly estimating this contact from a single egocentric image is very challenging. Therefore, a data-driven approach is employed by learning a model that predicts a plausible 3D pose based on the estimated scene geometry and features extracted from the input image. To achieve this goal, the EgoPW body joints heatmap estimator is first leveraged to extract 2D body pose features F and use the scene depth estimator from Sec. 6.2.2 to estimate the depth map of the scene without human body $\hat{D}^S$. Afterward, the body pose features and depth map are projected into a 3D volumetric space considering the fisheye camera projection model. After obtaining the volumetric representation of human body features V_{body} and scene depth V_{scene}, the 3D body pose $\hat{P}$ is predicted from the volumetric representation with a V2V network (Moon et al., 2018).

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

6.2.3.1 *Scene and Body Encoding as a 3D Volume*

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

The 3D coordinates in V are projected into the egocentric image space with the fisheye camera model (Scaramuzza and Ikeuchi, 2014): $V_{\text{proj}} = \mathcal{P}(V)$, where $V_{\text{proj}} \in R^{N,N,N,2}$ and $\mathcal{P}$ is the fisheye camera projection function. The volumetric representation V_{body} of the human body is obtained by filling a cube $V_{\text{body}} \in R^{N,N,N,K}$ by bilinear sampling from the feature maps F with K channels using 2D coordinates in V_{proj}:

$$V_{\text{body}} = F\{V_{\text{proj}}\} \tag{6.3}$$

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

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

$$\left\| \left(\frac{xL}{N} - \frac{L}{2}, \frac{yL}{N} - \frac{L}{2}, \frac{zL}{N}\right) - (x_p, y_p, z_p) \right\| < \epsilon \tag{6.4}$$

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

6.2.3.2 *Predicting 3D Body Pose with V2V Network*

The volumetric representation aggregated from V_{body} and V_{scene} are fed into the volumetric convolutional network $\mathcal{N}$, which has a similar architecture as Moon et al. (2018). The V2V network produces the 3D heatmaps of the body joints:

$$V_{\text{heatmap}} = \mathcal{N}(V_{\text{body}}, V_{\text{scene}}) \tag{6.5}$$

Following Iskakov et al. (2019), the soft-argmax of V_{heatmap} is computed across the spatial axes to infer the body pose $\hat{P}$. The predicted pose $\hat{P}$ is finally compared with the ground truth pose P_G from the EgoGTA dataset and P_E from the EgoPW-Scene dataset with the MSE loss.

6.3 EXPERIMENTS

In this section, the method is evaluated using both existing and new datasets for egocentric monocular 3D human pose estimation. For implementation details, please refer to the Sec. C.2 of Appendix C.

6.3.1 *Evaluation Datasets*

Evaluating human-scene interaction requires precise annotations for camera pose and scene geometry. However, such information is not available in existing datasets for egocentric human pose estimation. To address this issue, a new real-world dataset called the **SceneEgo test dataset** was collected using a head-mounted fisheye camera combined with a calibration board. The ground truth scene geometry is obtained with SfM method (Hartley and Zisserman, 2003) from a multi-view capture system with 120 synced 4K resolution cameras and the ground truth egocentric camera pose is obtained by localizing a calibration board rigidly attached to the egocentric camera. This dataset contains around 28K frames of two actors wearing 4 different clothes, performing various human-scene interacting motions such as sitting, reading a newspaper, and using a computer. This dataset is evenly split into training and testing splits. The method was finetuned on the training split before the evaluation. This dataset will be made publicly available and additional details of it are shown in the supplementary materials.

Besides the new SceneEgo test dataset, the proposed methods are also evaluated on the test datasets from Chapter 4 and Mo^{2}Cap^2 (Xu et al., 2019). The real-world dataset in Mo^{2}Cap^2 (Xu et al., 2019) contains 2.7K frames of two people captured in indoor and outdoor scenes, and the dataset in Chapter 4 contains 12K frames of two people captured in the studio.

6.3.2 *Evaluation Metrics*

The accuracy of the estimated body pose is measured using MPJPE and PA-MPJPE. For the test dataset in Chapter 4 and Mo^{2}Cap^2 (Xu et al., 2019), PA-MPJPE and BA-MPJPE (Xu et al., 2019) are evaluated since

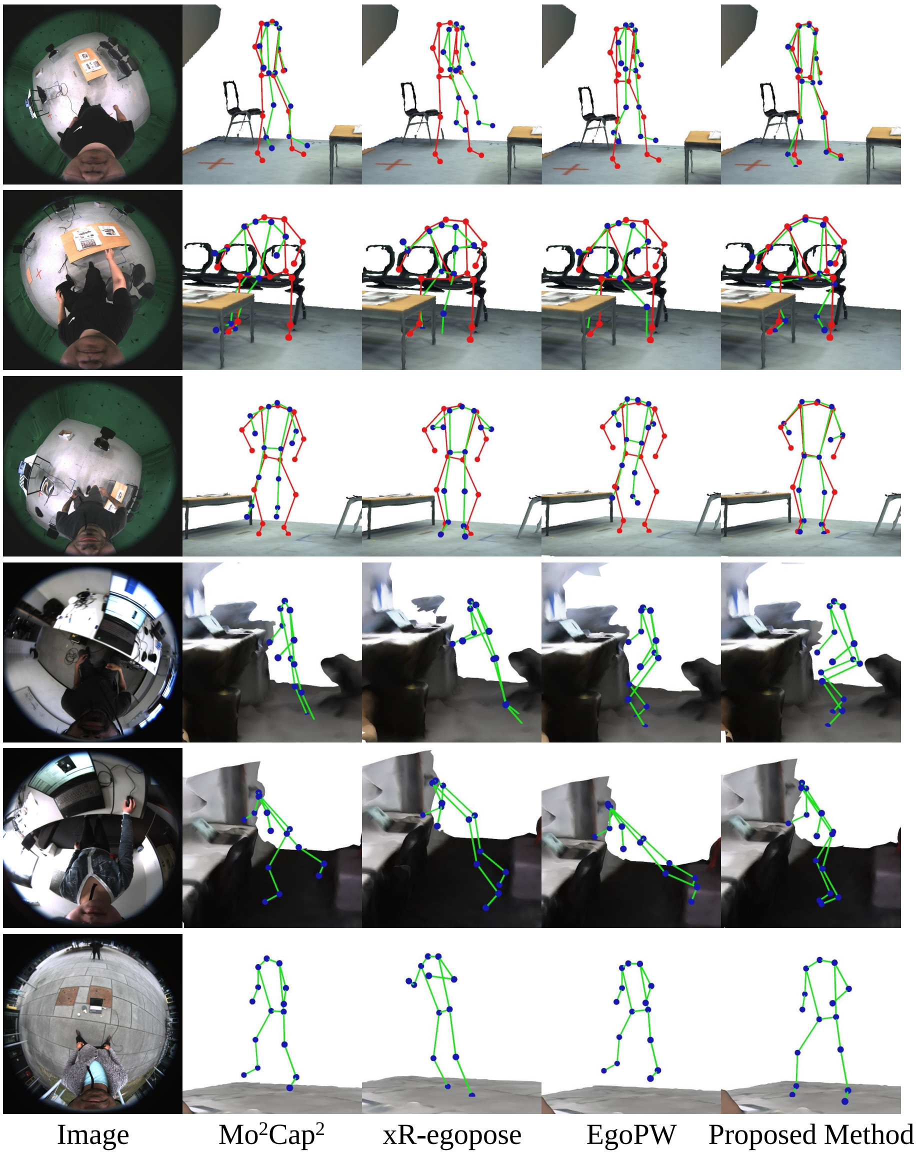


Figure 6.3: Qualitative comparison between the proposed method and the state-of-the-art egocentric pose estimation methods. From left to right: input image, Mo^2Cap^2 result, *x*R-egopose result, EgoPW result, and the result of the proposed method. The ground truth pose is shown in red. The input images from the first three rows are from the SceneEgo test dataset, while those in the last three rows come from the EgoPW in-the-wild test sequences (without ground-truth poses). This figure also shows the gt scene geometry of the in-the-studio data and scene geometry obtained by the SFM method for the in-the-wild data.

Method	MPJPE	PA-MPJPE
SceneEgo test dataset		
Mo^{2}Cap^2 (Xu et al., 2019)	200.3	121.2
xR-egopose (Tomè et al., 2019)	241.3	133.9
EgoPW (Chapter 5)	189.6	105.3
Proposed Method	**118.5**	**92.75**
Method	PA-MPJPE	BA-MPJPE
Test dataset in Chapter 4		
Mo^{2}Cap^2 (Xu et al., 2019)	102.3	74.46
xR-egopose (Tomè et al., 2019)	112.0	87.20
EgoPW (Chapter 5)	81.71	64.87
Proposed Method	**76.50**	**61.92**
Mo^{2}Cap^2 test dataset (Xu et al., 2019)		
Mo^{2}Cap^2 (Xu et al., 2019)	91.16	70.75
xR-egopose (Tomè et al., 2019)	86.85	66.54
EgoPW (Chapter 5)	83.17	64.33
Proposed Method	**79.65**	**62.82**

Table 6.1: Performance of the proposed method on SceneEgo test dataset, test dataset in Chapter 4 and Mo^{2}Cap^2 test dataset (Xu et al., 2019). The proposed method outperforms the state-of-the-art methods EgoPW, Mo^{2}Cap^2 (Xu et al., 2019) and xR-egopose (Tomè et al., 2019).

the ground truth poses in the egocentric camera space are not provided. Further details of the metrics are shown in the supplementary materials.

6.3.3 *Comparisons on 3D Pose Estimation*

This section compares the proposed method with previous single-frame-based methods, including EgoPW, xR-egopose (Tomè et al., 2019) and Mo^{2}Cap^2 (Xu et al., 2019) on SceneEgo test dataset under the "SceneEgo test data" in Table 6.1. Since the code for xR-egopose has not been released, a re-implemented version is used for the evaluation. In the SceneEgo dataset, the proposed method outperforms the previous state-

Method	Non pene.	Contact
Mo^{2}Cap^2 (Xu et al., 2019)	69.6%	23.1%
xR-egopose (Tomè et al., 2019)	64.5%	38.3%
EgoPW (Chapter 5)	71.7%	38.8%
Proposed Method	**84.1%**	**89.4%**

Table 6.2: Comparisons of physical plausibility on the SceneEgo test dataset.

of-the-art methods, EgoPW by 37.5% on MPJPE and 11.9% on PA-MPJPE. The proposed method is also compared with previous methods on the test dataset in Chapter 4 and Mo^{2}Cap^2 test dataset (Xu et al., 2019) and show the results in Table 6.1. On the test dataset in Chapter 4, the proposed method performs better than EgoPW by 6.4%. On the Mo^{2}Cap^2 test dataset, this method performs better than EgoPW by 7.8%.

The physical plausibility of the predictions is also evaluated by calculating the percentage of predicted poses that are in contact with the scene and do not penetrate the scene, as shown in Table 6.2. A body pose is defined as being in contact with the scene if any body joint is less than 5 cm from the scene mesh. A body pose suffers from the body floating issue if it is not in contact with the scene. Compared with previous approaches, the proposed method generates body poses that are more physically plausible considering the constraints of the scene.

From the results in Table 6.1 and Table 6.2, the proposed approach outperforms all previous methods on the single-frame egocentric pose estimation task. For the qualitative comparison, the results of this method on the studio dataset and in-the-wild sequences are shown in Fig. 6.3. The predicted poses are physically plausible under the scene constraint, whereas other methods generate poses suffering from body floating and penetration issues.

To further demonstrate that this method can predict poses according to the constraints of the scene geometry, the input image is fixed, and the scene depth input is changed to the depth map corresponding to the standing pose, squatting pose, and sitting pose. The results are presented in Fig. 6.4 and show that the predicted poses change to standing, squatting, and sitting to better adapt to the input changes of the scene geometry. This shows the method's ability to disambiguate poses under different scene constraints.

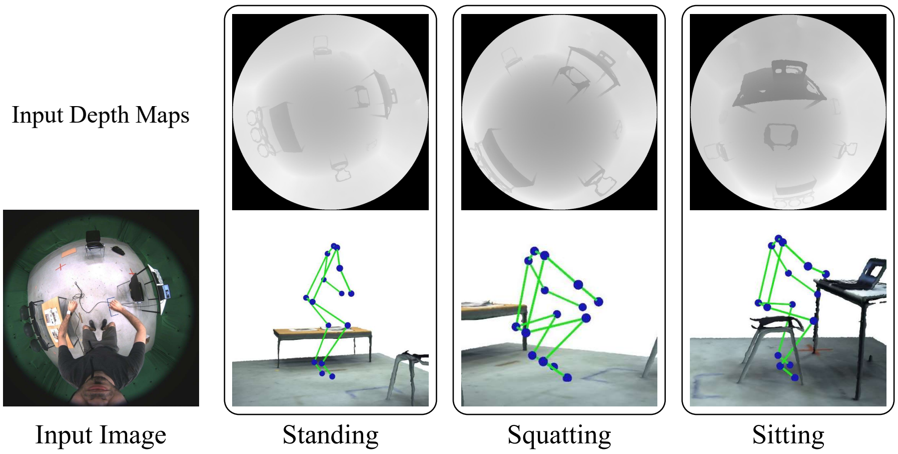


Figure 6.4: Predicted pose with different scene depth map input. The proposed network can generate different poses under different depth inputs and further disambiguate body poses under scene constraints.

6.3.4 *Ablation Study*

SIMPLE COMBINATION OF 2D FEATURES AND DEPTH MAPS. In Sec. 6.2.3, the volumetric representation of egocentric 2D features and scene depth map is claimed to be important for understanding the interaction between the human body and the surrounding scene. To provide evidence for this claim, an experiment is conducted comparing the proposed method with baseline methods that simply combine the 2D image features and scene depth map. Two baseline methods are used since there are two types of egocentric pose estimation methods, i. e.direct regression of 3D poses (xR-egopose (Tomè et al., 2019)) and prediction of 2D position and depth for each joint (Mo^{2}Cap^2 (Xu et al., 2019) and EgoPW). In the baseline method "xR-egopose + Depth", 2D heatmaps and scene depth maps are concatenated as the input to the 3D pose regression network in xR-egopose. In the baseline "EgoPW + Depth", 2D features and the scene depth map are concatenated and input into the joint depth prediction network.

From the evaluation results shown in Table 6.3, both of the baseline methods perform worse than the proposed method. In "xR-egopose + Depth", simply combining the scene depth and 2D heatmaps cannot provide direct geometric supervision for the 3D pose regression network. In "EgoPW + Depth", though the joint depth estimation network performs better with the help of scene depth information, the 2D pose estimation network does not benefit from it. Both of the experiments demonstrate the effectiveness of the volumetric representation of 2D features and scene

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

Table 6.3: Results from the proposed method compared to different baselines.

geometry, which provides direct geometry supervision for the physical plausibility of predicted egocentric poses.

OPTIMIZATION. This experiment compares the proposed method with an optimization baseline that refines a 3D pose considering the scene constraint. Similar to the method in Chapter 4 and EgoPW method, a VAE consisting of a CNN-based encoder f_{enc} and decoder f_{dec} is first trained to learn an egocentric motion prior. Then, the egocentric pose P is optimized by finding a latent vector z such that the corresponding pose $P = f_{\text{dec}}(z)$ minimizes the objective function $E(P) = \lambda_R E_R + \lambda_J E_J + \lambda_C E_C$, where E_R is the egocentric reprojection term, E_J is the egocentric pose regularization term, and E_C is the contact term. The latent vector z is initialized with the estimated pose from the EgoPW method. The E_R and E_J are the same as those defined in Chapter 4. Denote the nth joint in egocentric pose P as $P_n, n \in [0, N]$, where N is the number of joints, and the mth point in scene point cloud C as $C_m, m \in [0, M]$, where M is the number of points in the point cloud. The contact term E_C is defined as:

$$\begin{aligned} E_C &= \sum_{n\in[0,N]} d_n^2, \quad \text{if } d_n \leq \epsilon, \text{ otherwise } 0, \text{ and} \\ d_n &= \min_{m\in[0,M]} \|P_n - C_m\|_2. \end{aligned} \tag{6.6}$$

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

The result of the optimization method is shown as "EgoPW+Optimizer" in Table 6.3, which demonstrates that the optimization framework is less effective than the proposed method. This is because the accuracy of the optimization method relies heavily on the initial pose. If the initial pose is not accurate, it will be difficult to determine the contact labels for each joint with the fixed distance margin. Without accurate contact labels, the optimization framework might force the joint that does not contact the scene to keep in contact, eventually resulting in wrong poses.

SCENE DEPTH ESTIMATOR. The proposed estimates the depth of the surrounding scene and infers the depth behind the human body with a depth inpainting network. To validate the effectiveness of the scene depth map, the input depth map is removed from the V2V network and the results are shown as "w/o Depth" in Table 6.3. This baseline increases the MPJPE by about 70 mm, which is evidence of the relevant extra information provided by the scene depth.

To demonstrate the benefits of recovering the scene depth behind the human body, the proposed model is evaluated using the estimated depth, including the human body, as input to the V2V network. The human body area is also removed from the depth maps and the resulting depth maps are used as input to the V2V network. The results are shown in "w/ Depth with Body" and "w/ Depth w/o Body" in Table 6.3. Both of the baseline methods perform worse than the proposed method because the area in the scene occluded by the human body can provide clues for generating plausible poses.

The proposed method with ground truth depth maps is also evaluated in "w/ GT Depth" in Table 6.3, which further improves over the estimated depth by 7.2% in terms of MPJPE and 4.2% in terms of PA-MPJPE. This demonstrates that the accuracy of the predicted pose benefits from better depth maps, but still the estimated scene depth already provides a significant improvement over the baselines.

Finally, the proposed method is compared with a baseline method without depth inpainting, *i.e.*, the human body is removed from the input image, and the scene depth map is predicted directly with the network $\mathcal{D}$ in Sec. 6.2.2. The pose estimation accuracy is shown in "w/ Depth w/o Inpainting" of Table 6.3 and the depth accuracy is shown in Table 6.4. The proposed method outperforms the baseline as it is more challenging to simultaneously estimate and inpaint the depth. Moreover, the network can be influenced by the segmented part in the input image as some extinct object, as shown in Fig 6.5.

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

Table 6.4: The quantitative results of depth estimator. The Abs-Rel and RMSE are evaluated on the SceneEgo test dataset following Hu et al. (2019).

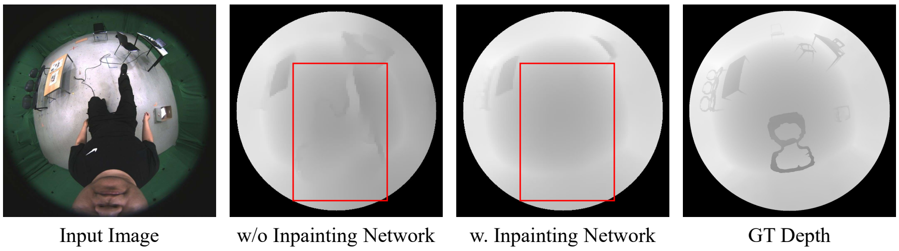


Figure 6.5: The qualitative depth estimation results with or without the inpainting network. The depth map estimated without an inpainting network shows artifacts in the human body region (see the red box).

6.4 LIMITATION

The accuracy of voxel-based pose estimation network is constrained by the accuracy of estimated depth, especially where the scene is occluded by the human body. In the future, one solution is to leverage the temporal information to get a full view of the surrounding environment by using SLAM or SFM method. Another possible solution is to leverage the stereo SLAM or RGBD SLAM to obtain accurate scene geometry information.

Furthermore, the efficiency of the network inference is also influenced by the slow inference speed of the V2V network. In order to accelerate the speed, the tri-plane representation can be used instead to encode the image features and scene geometry in the 3D space.

6.5 CONCLUSIONS

In this paper, a new approach to estimate egocentric human pose under scene constraints is proposed. First, a depth inpainting network is trained to estimate the depth map of the scene without the human body. Next, the egocentric 2D features and scene depth map are combined in a volumetric space, and the egocentric pose is predicted with a V2V network. The experiments demonstrate that the proposed method outperforms all baseline methods both qualitatively and quantitatively, and it can predict physically plausible poses in terms of human-scene interaction. In the following chapters, this work will be referred to as **"SceneEgo"**.

Though hands can be observed in the egocentric view, existing egocentric human motion capture models cannot track the hand motion simultaneously with the human body motion. To solve this issue, the next chapter introduces a novel method for egocentric whole-body motion capture, which simultaneously captures both human body and hand movements.

7 EGOCENTRIC WHOLE-BODY MOTION CAPTURE WITH FISHEYEVIT AND DIFFUSION-BASED MOTION REFINEMENT

The existing egocentric motion capture methods in the literature, as well as those discussed in previous chapters, focus only on human body motion rather than hand motion. This limitation restricts a wide range of applications, including human-object interactions, photorealistic telepresence, and so on. This Chapter explores egocentric whole-body motion capture using a single fisheye camera, which simultaneously estimates human body and hand motion. This task presents significant challenges due to three factors: the lack of high-quality datasets, fisheye camera distortion, and human body self-occlusion. To address these challenges, a novel approach, is proposed that leverages FisheyeViT to extract fisheye image features, which are subsequently converted into pixel-aligned 3D heatmap representations for 3D human body pose prediction. For hand tracking, dedicated hand detection and hand pose estimation networks are incorporated to regress 3D hand poses. Finally, a diffusion-based whole-body motion prior model is developed to refine the estimated whole-body motion while accounting for joint uncertainties.

To train these networks, a large synthetic dataset, EgoWholeBody, comprising 840,000 high-quality egocentric images captured across a diverse range of whole-body motion sequences, is collected. Quantitative and qualitative evaluations demonstrate the effectiveness of the method in producing high-quality whole-body motion estimates from a single egocentric camera.

7.1 INTRODUCTION

Egocentric 3D human motion estimation using head-mounted devices (Tomè et al., 2019; Xu et al., 2019) has garnered significant traction in recent years, driven by its diverse applications in VR/AR. Immersed in a virtual world, users can traverse virtual environments, interact with virtual objects, and even simulate real-world interactions. To fully capture the intricacies of human motion during such interaction, understanding both body and hand movements is essential. While existing egocentric motion capture methods (Liu et al., 2023b; Tomè et al., 2019; Xu et al., 2019) focus solely on body motion, neglecting the hands, this work proposes the task

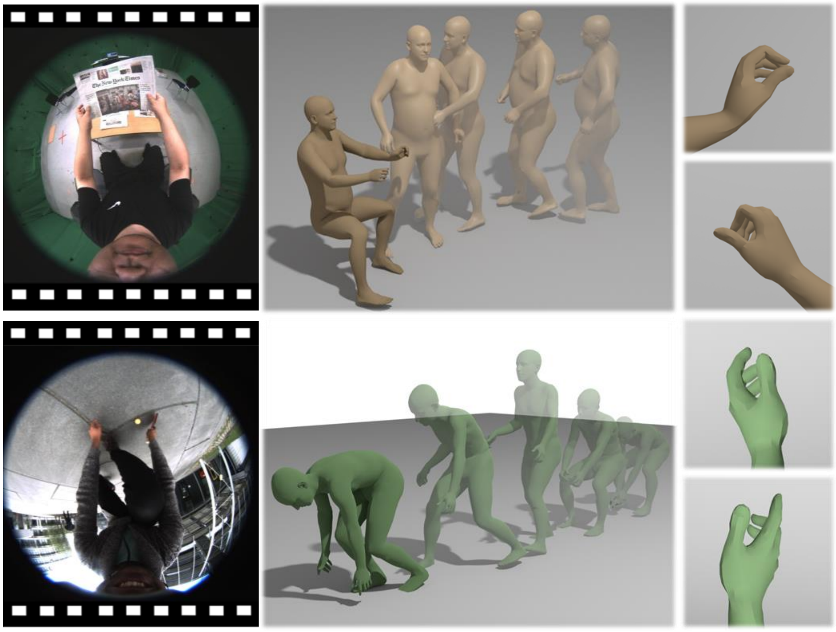


Figure 7.1: From an image sequence captured by a single head-mounted fisheye camera, the proposed method can predict accurate and temporally coherent whole-body motion, including the human body and hand poses. The SMPL-X parameters are obtained using inverse kinematics.

of egocentric *whole-body* motion capture, i. e.simultaneous estimation of the body motion and hand motion from a single head-mounted fisheye camera (shown in Fig. 5.1). This task is extremely challenging due to three factors: First, the fisheye image introduces significant distortion, making it difficult for existing networks, which are designed for non-distorted images, to extract features. Second, the egocentric camera perspective frequently leads to the occlusion of body parts, such as the feet and hands, further complicating the task of whole-body motion capture. Lastly, large-scale training data with ground truth annotations for both body and hand poses is absent in existing datasets including UnrealEgo (Akada et al., 2022), Mo^{2}Cap^2 (Xu et al., 2019),*x*R-Egopose (Tomè et al., 2019), EgoFish3D (Liu et al., 2023a) and EgoPW (Chapter 5).

This Chapter proposes a novel egocentric whole-body motion capture method to address the aforementioned challenges. To effectively address fisheye distortion, *FisheyeViT* is proposed for extracting image features, along with a joint regressor employing *pixel-aligned 3D heatmap* for predicting 3D body poses. Instead of attempting to undistort the entire fisheye image, which is impractical due to the fisheye lens's large field of view (FOV), the image is partitioned into smaller patches aligned with a specific FOV range. This approach enables individual patch-level undistortion and seamlessly aligns with the vision transformer architecture that is employed for extracting the complete image feature map. An egocentric 3D pose regressor utilizing 3D heatmap representations is further proposed. Unlike the existing approach in Chapter 6 that projects image features into 3D space through fisheye reprojection functions and regresses 3D heatmaps with V2V networks (Moon et al., 2018)–leading to intricate network learning and high computational complexity–the proposed egocentric pose regressor adopts a simpler approach. It employs deconvolutional layers to obtain pixel-aligned 3D heatmaps. Notably, the voxels in the 3D heatmap directly correspond to pixels in 2D features, subsequently linking to image patches in FisheyeViT. This streamlined approach significantly simplifies network training. Joint locations from the pixel-aligned 3D heatmap are finally transformed with the fisheye camera model to obtain the 3D human body poses. Due to the large size difference between the body and hands, a hand detection network and a hand pose estimation network are trained to accurately regress 3D hand poses.

To overcome the challenges posed by self-occlusion and improve the accuracy of pose estimation, a novel method is proposed for refining the whole-body motion predictions by incorporating temporal context and a motion prior. The proposed method learns a whole-body motion prior with the diffusion model (Ho et al., 2020) from a collection of diverse human motion sequences, capturing intrinsic correlations between hand and body movements. Following this, the joint uncertainties are extracted from the pixel-aligned 3D heatmap and utilize them to guide the refinement of the whole-body motion. The joint uncertainties act as indicators of the trustworthiness of the pose regressor's predictions. By conditioning on joints with low uncertainty, the whole-body motion diffusion model selectively refines joints with high uncertainty. This strategy substantially improves the quality of whole-body pose estimations and effectively mitigates the effects of self-occlusion.

In response to the absence of the egocentric whole-body motion capture datasets, the Chapter presents *EgoWholeBody*, a new large-scale high-quality synthetic dataset. This dataset encompasses a wide range of whole-body motions, comprising over 870k frames, which significantly surpasses the size of previous egocentric training datasets. EgoWholeBody could serve as a valuable resource for advancing research in egocentric whole-body motion capture.

A thorough evaluation across a range of datasets, including SceneEgo (Chaper 6), GlobalEgoMocap (Chapter A) and Mo^{2}Cap^2 (Xu et al., 2019), has demonstrated the remarkable improvements of the proposed method in estimating egocentric whole-body motion compared to pre-

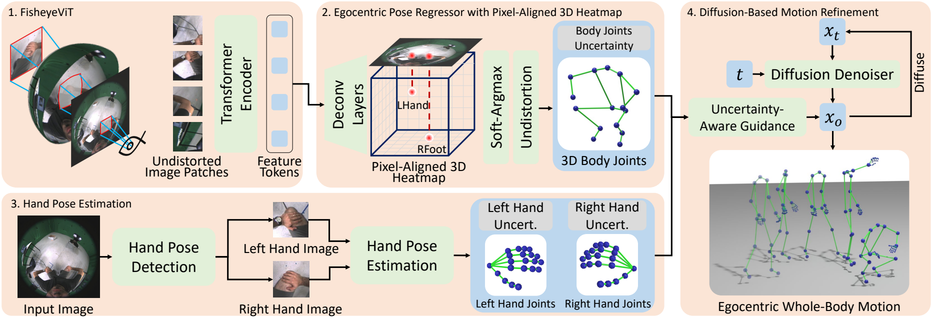


Figure 7.2: Overview of the whole-body motion capture pipeline. First, FisheyeViT is used to undistort the input image and generate image feature tokens (7.2.1.1). Next, a convolutional network is used to convert the image features to a pixel-aligned 3D heatmap and use soft-argmax and fisheye camera undistortion function to obtain the 3D body joins positions and uncertainty (7.2.1.2). The hand location is further detected, and the 3D hand poses are further regressed from the input image (7.2.1.3). Finally, the estimated hand motion and human body motion are combined and the uncertainty-aware diffusion model is applied to refine the estimated whole-body motion (7.2.2).

vious approaches. This substantiates the effectiveness of the proposed approach in addressing the special challenges encountered in egocentric views, including fisheye distortion and self-occlusion.
In summary, the key contributions of this Chapter are the following:

- The first egocentric whole-body motion capture method that predicts accurate and temporarily coherent egocentric body and hand motion;
- FisheyeViT for alleviating fisheye camera distortion and pose regressor using pixel-aligned 3D heatmaps for accurate egocentric body pose estimation from a single image;
- Uncertainty-aware refinement method based on motion diffusion models for correcting initial pose estimations and predicting plausible motions even under occlusion;
- *EgoWholeBody*, a new high-quality synthetic dataset for egocentric whole-body motion capture.

7.2 METHOD

This section proposes a new method for predicting accurate egocentric whole-body poses from egocentric image sequences. An overview of the proposed approach is shown in Fig. 7.2.

7.2.1 *Single Image Based Egocentric Pose Estimation*

7.2.1.1 *FisheyeViT*

In this section, FisheyeViT is introduced, which is specially designed to alleviate the fisheye distortion issue. Instead of undistorting the entire fisheye image, undistorted image patches are extracted from the fisheye image and then used as tokens in the transformer network (Dosovitskiy et al., 2020). To obtain the undistorted patches, the fisheye image is first warped to a unit semi-sphere. Then, the patches are obtained using the gnomonic projection (see Fig.7.2).

The FisheyeViT can be split into five steps, the first four of which are illustrated in Fig. 7.3.

Step 1. Given an input image $\mathbf{I}$ with size $H \times W$, $N \times N$ patch center points are first evenly sampled: $\{\mathbf{C}_{ij} = (u_i, v_j) = \left(\frac{H}{N}(i + \frac{1}{2}), \frac{W}{N}(j + \frac{1}{2})\right) | i, j \in 0, ..., N-1\}$. Then, the patch center points $\mathbf{C}_{ij}$ are projected onto a unit sphere with the fisheye reprojection function: $\mathbf{P}^c_{ij} = (x^c_{ij}, y^c_{ij}, z^c_{ij}) = \mathcal{P}^{-1}(u_i, v_j, 1)$. The fisheye camera model is described in Chapter 3. Given a point $\mathbf{P}^c_{ij}$ on the unit sphere, the tangent plane $\mathbf{T}_{ij}$ that passes through the point is defined by the normal vector $\mathbf{v}^c_{ij} = (x^c_{ij}, y^c_{ij}, z^c_{ij})$. In the following steps, the gnomonic projection is implemented by sampling grid points in the plane and projecting them back onto the fisheye image.

Step 2. In this step, the orientation of the grid points in the tangent plane is determined, ensuring that the grid points from different tangent planes $\mathbf{T}_{ij}$ have the same orientation when projected back onto the fisheye image. To achieve this, this method selects a 2D point $\mathbf{U}_{ij} = (u_i + d, v_j)$ in the fisheye image space that is d pixels to the right of the patch center point and project it to the unit sphere using the fisheye reprojection function: $\mathbf{P}^u_{ij} = (x^u_{ij}, y^u_{ij}, z^u_{ij}) = \mathcal{P}^{-1}(u_i + d, v_j, 1)$. The intersection point $\mathbf{P}^x_{ij}$ is then calculated between the vector $\mathbf{v}^u_{ij} = (x^u_{ij}, y^u_{ij}, z^u_{ij})$ that is passing the origin and the tangent plane $\mathbf{T}_{ij}$:

$$\mathbf{P}^x_{ij} = \frac{\left\langle \mathbf{P}^c_{ij}, \mathbf{v}^c_{ij} \right\rangle}{\left\langle \mathbf{v}^u_{ij}, \mathbf{v}^c_{ij} \right\rangle} \mathbf{v}^u_{ij} = \frac{1}{\left\langle \mathbf{v}^u_{ij}, \mathbf{v}^c_{ij} \right\rangle} \mathbf{v}^u_{ij}, \tag{7.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

Step 3. Based on the center point $\mathbf{P}^c_{ij}$ and intersection point $\mathbf{P}^x_{ij}$ on the tangent plane $\mathbf{T}_{ij}$, a coordinate system is built with the x axis: $\mathbf{v}^x_{ij} = \mathrm{Norm}(\mathbf{P}^x_{ij} - \mathbf{P}^c_{ij})$, the z axis: $\mathbf{v}^z_{ij} = \mathrm{Norm}(\mathbf{v}^c_{ij})$ and the y axis: $\mathbf{v}^y_{ij} = \mathbf{v}^z_{ij} \times \mathbf{v}^x_{ij}$, where Norm denotes the normalize operation. $M \times M$ points are grid-sampled in a $l \times l$ square on the x-y plane:

$$\{\mathbf{P}^{mn}_{ij} = \mathbf{P}^c_{ij} + (l\frac{m}{M}\mathbf{v}^x_{ij}, l\frac{n}{M}\mathbf{v}^y_{ij})\} \tag{7.2}$$

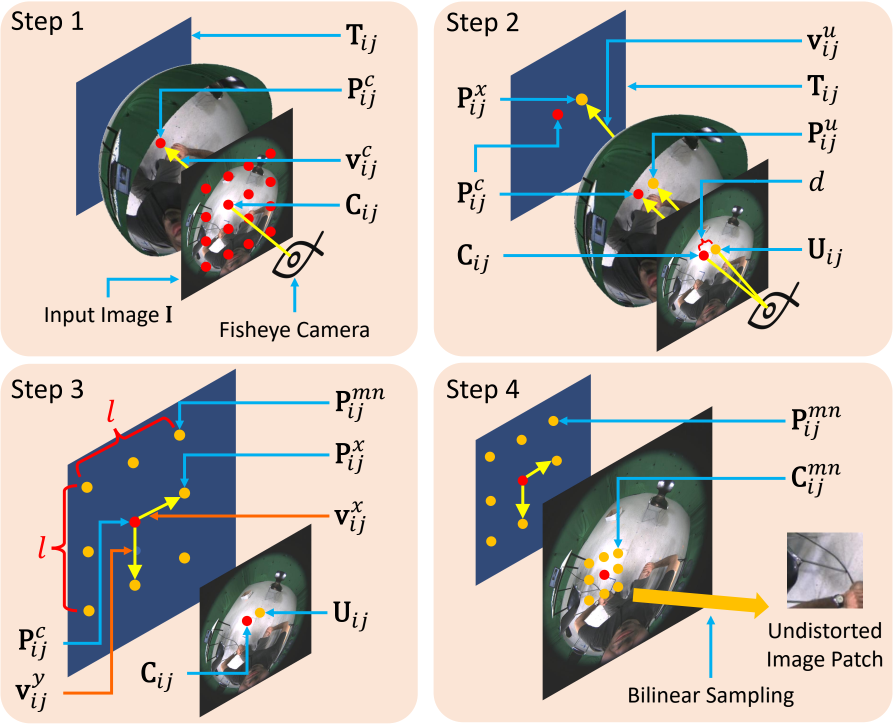


Figure 7.3: The detailed illustration of FisheyeViT (Sec. 7.2.1.1).

where $m, n \in -\frac{1}{2}(M-1), ..., -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, ... \frac{1}{2}(M-1)$.

Step 4. The points $\mathbf{P}_{ij}^{mn}$ are projected back to the fisheye image with the fisheye projection function: $\mathbf{C}_{ij}^{mn} = \mathcal{P}(\mathbf{P}_{ij}^{mn})$. The bilinear sampling is then applied to obtain the colors at points $\mathbf{C}_{ij}^{mn}$ of the input image $\mathbf{I}$, yielding the undistorted image patch $\mathbf{I}_{ij}^{\text{undis}}$. Please also see the supplementary video for a visual demonstration of undistorted image patches and their movement on the fisheye image.

Step 5. The image patches $\{\mathbf{I}_{ij}^{\text{undis}}\}$ are sent to a ViT transformer network (Dosovitskiy et al., 2020) to obtain the feature tokens $\{\mathbf{F}_{ij}\}$. The feature token is further reshaped in $i \times j$ matrix and obtain the image feature $\mathbf{F}$. In the FisheyeViT, $N = 16; M = 16; d = 8; l = 0.2m$ is chosen given the image size $H = W = 256$.

Note that $\mathbf{C}_{ij}^{mn}$ is independent of the image $\mathbf{I}$. This means that, given a fixed fisheye camera model, $\mathbf{C}_{ij}^{mn}$ can be pre-computed for all combinations of m, n and i, j in advance. This significantly speeds up both the training and evaluation processes. Furthermore, the number and dimensions of image patches $\{\mathbf{I}_{ij}^{\text{undis}}\}$ match exactly with those in the traditional ViT network. This compatibility allows us to finetune existing ViT networks on the egocentric datasets. The sampling strategy ensures that each image patch $\mathbf{I}_{ij}^{\text{undis}}$ corresponds to the same FOV range in the fisheye camera. The ablation study in Sec. 7.4.3 shows that FisheyeViT enhances the performance of the pose estimation network when applied to egocentric fisheye images.

7.2.1.2 *Pose Regressor with Pixel-Aligned 3D Heatmap*

After collecting image features with FisheyeViT, a 3D heatmap-based network is utilized to estimate the body poses. The existing 3D heatmap-based pose regressors (Moon et al., 2022; Sun et al., 2018) are designed for the weak-perspective cameras and predict the 3D heatmap in *xyz* space. Directly applying these regressors will result in misalignment between 3D heatmap features in *xyz* space and 2D image features in the fisheye image space. Therefore, the proposed method introduces a novel egocentric pose regressor that relies on the pixel-aligned 3D heatmap, tailored to address the needs of fisheye cameras. The idea is to regress the 3D heatmap in *uvd* space rather than traditional *xyz* space, where *uv* corresponds to the fisheye image *uv* space. Specifically, given a feature map $\mathbf{F} \in \mathbb{R}^{C\times N\times N}$, where C is the channel number, N is feature map height and width, two deconvolutional layers are firstly used to convert the feature map $\mathbf{F}$ into shape $(D_h \times J, H_h, W_h)$, and further reshape it to pixel-aligned 3D heatmap $\mathbf{H} \in \mathbb{R}^{J\times D_h\times H_h\times W_h}$, where J is the joint number and D_h, H_h, W_h is the 3D heatmap depth, height and width. The illustration of pixel-aligned 3D heatmap is shown in Fig. 7.2. Next, the max-value positions $\tilde{\mathbf{J}}_b = \{(u_i, v_i, d_i) \mid i \in 0, 1, 2, ..., J\}$ is obtained from $\mathbf{H}$ by the differentiable soft-argmax operation (Sun et al., 2018). Here, it is worthy to note that u_i and v_i correspond to the uv-coordinate of the 3D body joint projected in the fisheye image space, and d_i denotes the distance of the joint to the fisheye camera. Finally, the 3D body joints $\hat{\mathbf{J}}_b = \{(x_i, y_i, z_i) \mid i \in 0, 1, 2, ..., J\}$ are recovered with the fisheye reprojection function: $(x_i, y_i, z_i) = \mathcal{P}^{-1}(u_i, v_i, d_i)$. The predicted body pose $\hat{\mathbf{J}}_b$ is finally compared with the ground truth body pose $\mathbf{J}_b$ with the MSE loss. By first regressing 3D body poses in *uvd* space and then reprojecting it, the proposed method ensures that the 3D heatmap is pixel-aligned with the end-to-end training.

With the pixel-aligned heatmap, the proposed 3D pose regressor solves problems in all three types of previous egocentric joint regressors. First, Mo^{2}Cap^2 (Xu et al., 2019) employs separate networks to predict 2D joint positions and joint distances. However, this method can yield unrealistic joint estimations because small errors in 2D joints can result in large errors in 3D joints due to the projection effect. Second, *x*R-egopose (Tomè et al., 2019) and EgoHMR (Liu et al., 2023b) directly regress the 3D joint positions. However, this method is agnostic to the fisheye camera parameters, making it suitable only for a specific camera configuration (*e.g.*, camera parameters, head-mounted position, and so on). Third, SceneEgo (Chapter 6) projects 2D features into 3D voxel space and uses a V2V network to regress 3D poses. Because of these, the SceneEgo method suffers from low accuracy and large computation overhead. Different from previous methods, the pose regressor with pixel-aligned 3D heatmap is versatile and efficient since it directly estimates 3D joints while also incorporating an explicitly parametrized fisheye camera model. Moreover, it can preserve the uncertainty of the estimated joints, which will be used in the uncertainty-aware motion refinement method (Sec. 7.2.2.2). Detailed comparison with other pose prediction heads is shown in Table 7.3.

7.2.1.3 *Egocentric Hand Pose Estimation*

In this section, a network is first trained to detect hand pose locations, followed by training a 3D hand pose estimation network to regress 3D hand poses. Then, the process of integrating the estimated hand and body poses is described.

HAND DETECTION. Given an input image $\mathbf{I}$, the HRNet (Wang et al., 2020b) network is first finetuned to regress the 2D hand poses of left hand $\mathbf{J}_{lh}^{2d}$ and right hand $\mathbf{J}_{rh}^{2d}$. Next, the center point of left hand $\mathbf{C}_{lh}$ and right hand $\mathbf{C}_{rh}$, along with the bounding box sizes, d_{lh} and d_{rh} are obtained from the hand poses. Given the center points and bounding boxes, the approach described in Sec. 7.2.1.1 is further used to compute undistorted image patches of left $\mathbf{I}_{lh}$ and right hands $\mathbf{I}_{rh}$.

HAND POSE ESTIMATION. Given the cropped image $\mathbf{I}_{lh}$ or $\mathbf{I}_{rh}$, the 3D hand poses $\hat{\mathbf{J}}_{lh}^{loc}$ and $\hat{\mathbf{J}}_{rh}^{loc}$ can be regressed with the Hand4Whole (Moon et al., 2022) network, which is fine-tuned on the EgoFullBody dataset.

INTEGRATION OF BODY AND HAND POSES. It is not straightforward to integrate the hand poses with the body pose in the egocentric camera view primarily due to the fisheye camera's perspective effects. Take the left hand as an example. Following Step 3 in Sec. 7.2.1.1, a local coordinate system is established on the tangent plane of the left-hand image with XYZ axes as follows: $x : \mathbf{v}_{lh}^{x}; y : \mathbf{v}_{lh}^{y}; z : \mathbf{v}_{lh}^{z}$. A rotation matrix $\mathbf{R}$ is defined to represent the transformation between the root coordinate system and the local coordinate system on the tangent plane. The estimated hand pose is first rotated with the rotation matrix $\hat{\mathbf{J}}_{lh} = \mathbf{R}\hat{\mathbf{J}}_{lh}^{loc}$ and then translated to align the wrist location of the human body. This same process is also applied to the right hand to get the right hand pose $\hat{\mathbf{J}}_{rh}$. The whole-body joints $\hat{\mathbf{J}}$ are obtained by combining $\hat{\mathbf{J}}_{b}$, $\hat{\mathbf{J}}_{lh}$, and $\hat{\mathbf{J}}_{rh}$. The uncertainty of whole-body joints $\hat{\mathbf{U}}$ is also obtained from the maximal value of the 3D heatmap in pose estimation modules.

7.2.2 *Diffusion-Based Motion Refinement*

The single-frame estimations in Sec. 7.2.1 suffer from inaccuracies and temporal instabilities. This section proposes a diffusion-based motion refinement method to tackle this problem. The whole-body motion prior is first learned with the motion diffusion model in Sec. 7.2.2.1. Then, an uncertainty-aware zero-shot motion refinement method is introduced in Sec. 7.2.2.2 to refine the initial whole-body motion estimations.

7.2.2.1 *Whole-Body Motion Diffusion Model*

The DDPM (Ho et al., 2020) is used as the diffusion approach to capture the whole-body motion prior $q(\mathbf{x})$. DDPM learns a distribution of whole-body motion $\mathbf{x}$ through a forward diffusion process and an inverse denoising process. The forward diffusion process is a Markov process of adding Gaussian noise over $t \in \{0, 1, ..., T-1\}$ steps:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)I) \tag{7.3}$$

where $\mathbf{x}_t$ denotes the whole-body motion sequence at step t, the variance $(1-\alpha_t) \in (0, 1]$ denotes a constant hyperparameter increases with t.

The inverse process uses a denoising network $D(\cdot)$ to remove the added Gaussian noise at each time step t. Here, the transformer-based framework in EDGE (Tseng et al., 2023) is adopted as the motion-denoising network $D(\cdot)$. This method follows Ramesh et al. (2022) to make the network predict the original signal itself, i. e.$\hat{\mathbf{x}}_0 = D(\mathbf{x}_t, t)$ and train it with the simple objective (Ho et al., 2020):

$$\mathcal{L}_{\text{simple}} = E_{\mathbf{x}_0 \sim q(\mathbf{x}_0), t \sim [1,T]} \left[||\mathbf{x}_0 - D(\mathbf{x}_t, t)||_2^2 \right] \tag{7.4}$$

7.2.2.2 *Uncertainty-Aware Motion Refinement*

Given the learned whole-body motion prior, the uncertainty value for each pose is leveraged to guide the diffusion denoising process using the classifier-guided diffusion sampling (Dhariwal and Nichol, 2021). Given an initial sequence of whole-body pose estimation $\mathbf{x}_e = \{\hat{\mathbf{J}}_i\}$ and the uncertainty value for each pose $\mathbf{u} = \{\hat{\mathbf{U}}_i\}$, where i denotes the ith pose in the sequence, the joints with low uncertainty are kept but the diffusion model is applied to generate joints with high uncertainty conditioned on the low-uncertainty joints. Specifically, in the tth sampling step of the diffusion process, the denoising network predicts $\hat{\mathbf{x}}_0 = D(\mathbf{x}_t, t)$, which is noised back to $\mathbf{x}_{t-1}$ by sampling from the Gaussian distribution:

$$\mathbf{x}_{t-1} \sim \mathcal{N}(\hat{\mathbf{x}}_0 + \mathbf{w}(\mathbf{x}_e - \hat{\mathbf{x}}_0), \Sigma_t) \tag{7.5}$$

where Σ_t is a scheduled Gaussian distribution in DDPM (Ho et al., 2020) and $\mathbf{w}$ controls the weight of a specific joint between the predicted motion $\hat{\mathbf{x}}_0$ and the estimated motion $\mathbf{x}_e$. Generally, $\mathbf{w} \rightarrow \overrightarrow{0}$ is expected when $t \rightarrow 0$ such that the temporal stability is guaranteed through the generation of the denoising process, and $\mathbf{w} \rightarrow \overrightarrow{1}$ when $t \rightarrow T$ such that the denoising process is initialized by the estimated motion $\mathbf{x}_e$. $w_{ij} = \mathbf{w}[i][j]$, which is the weight of jth joints in the ith pose, is also expected to be smaller when the uncertainty value $u_{ij} = \mathbf{u}[i][j]$ of the jth joints in the ith pose is large. Based on this requirement, $\mathbf{w}$ is designed as:

$$\mathbf{w} = 1 / \left(1 + e^{-k(t - T\mathbf{u})}\right) \tag{7.6}$$

where T is the overall diffusion steps, k is a hyperparameter which is empirically set to 0.1. The experimental results in Sec. 7.4 demonstrate the effectiveness of uncertain-aware motion refinement and the uncertainty-guided diffusion sampling strategy.

7.3 EGOWHOLEBODY DATASET

This section introduces EgoWholeBody, a large-scale high-quality synthetic dataset built for the task of egocentric whole-body motion capture. The EgoWholeBody dataset is organized into two sections. The first part, containing over 700k frames, is rendered with 14 different rigged Renderpeople (*RenderPeople* n.d.) models driven by 2367 Mixamo (*Mixamo* n.d.) motion sequences. The second part focuses on hand motions and contains 170k frames with the SMPL-X model. This data is constructed from 24 different shapes and textures, driven by 262 motion sequences selected from the GRAB (Taheri et al., 2020) and TCDHandMocap dataset (Hoyet et al., 2012). Synthetic test sequences have also been generated, comprising 133k images rendered using 3 Renderpeople models and Mixamo motions.

During the rendering process, a virtual fisheye camera is initially attached to the forehead of the human body models. Blender (*Blender* n.d.) is then used to render the images, semantic labels, and depth maps. The EgoWholeMocap dataset is larger and more diverse than previous egocentric training datasets–see Sec. D.3 in the Appendix D for a detailed comparison.

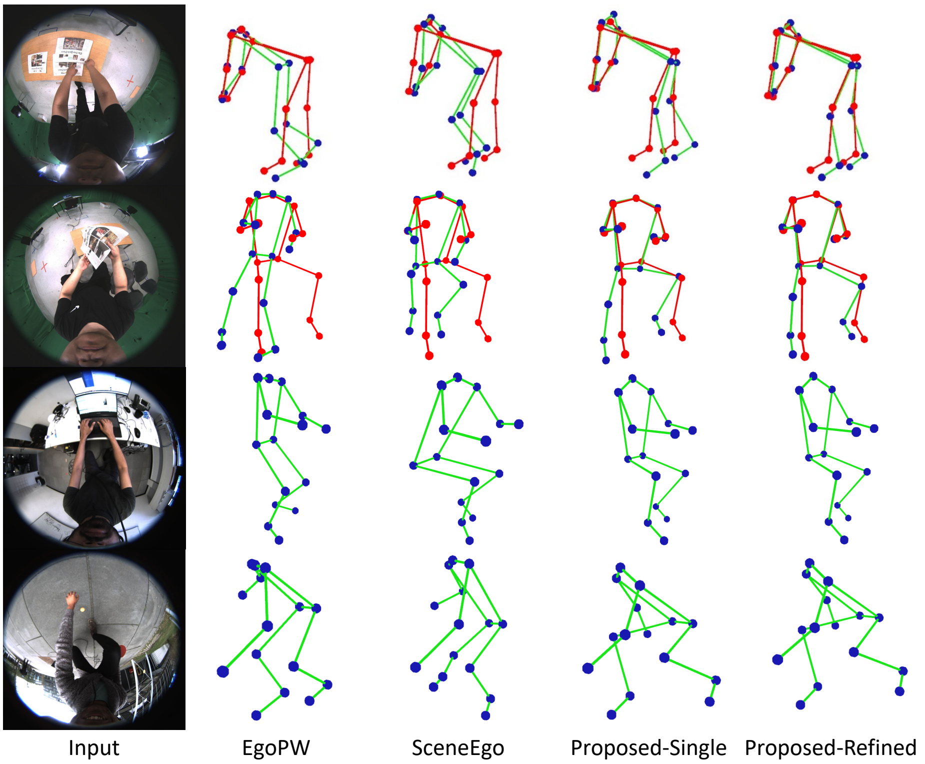


Figure 7.4: Qualitative comparison on human body pose estimations between the proposed method and the state-of-the-art egocentric pose estimation methods on in-the-studio (two rows from the top) and in-the-wild scenes (two rows from the bottom). The red skeleton is the ground truth while the green skeleton is the predicted pose. The proposed methods predict more accurate body poses compared with EgoPW (Chaper 5) and SceneEgo (Chaper 6).

7.4 EXPERIMENTS

7.4.1 *Datasets and Evaluation Metrics*

TRAINING DATASETS. To train the body pose estimation module (Sec. 7.2.1.1 and Sec. 7.2.1.2), the EgoWholeBody dataset and the EgoPW dataset (Chaper 5) are used. Additionally, the EgoWholeBody dataset is used to train the hand pose estimation module in Sec. 7.2.1.3. For training the whole-body diffusion model (Sec. 7.2.2), a combined motion capture dataset is utilized, which includes EgoBody (Zhang et al., 2022), Mixamo (*Mixamo* n.d.), TCDHandMocap dataset (Hoyet et al., 2012) and GRAB dataset (Taheri et al., 2020).

EVALUATION DATASETS. The experiments in this chapter evaluate the proposed methods on four datasets: the GlobalEgoMocap test datasets (Chap-

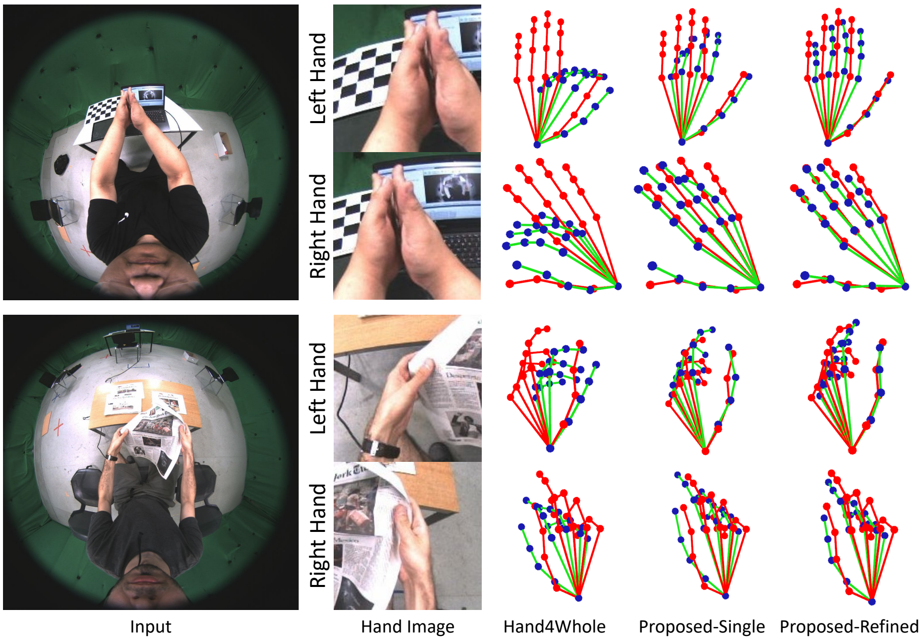


Figure 7.5: Qualitative comparison on human hand pose estimations between the proposed method and the state-of-the-art third-view pose estimation methods. The single-view and refined hand poses from the proposed method are more accurate than the poses from Hand4Whole (Moon et al., 2022) method. The red skeleton is the ground truth while the green skeleton is the predicted pose.

ter 4), the Mo^{2}Cap^2 test dataset (Xu et al., 2019), the SceneEgo test dataset (Chaper 6) and out EgoWholeBody test dataset. The details of the datasets are shown in Sec. D.5 of supplementary materials. Note that evaluating whole-body poses requires accurate annotations for human hands, which is absent in real-world datasets. To resolve the issue, a multi-view motion capture system is used to obtain the hand motion from the multi-view videos of the SceneEgo test dataset (Chaper 6). The hand pose annotations will be made publicly available.

EVALUATION METRICS. MPJPE and PA-MPJPE are adopted to evaluate the precision of human body poses on the SceneEgo test dataset (Chapter 6). PA-MPJPE and BA-MPJPE are evaluated for the GlobalEgoMocap test dataset (Chapter 4) and Mo^{2}Cap^2 test dataset (Xu et al., 2019), where egocentric camera poses are unavailable. For hand pose accuracy, the predicted and ground truth hand poses are aligned at the root position, followed by computing MPJPE and PA-MPJPE. Detailed explanations of these metrics are in Sec. D.4 of the supplementary materials. All reported metrics are in millimeters.

Method	MPJPE	PA-MPJPE
SceneEgo test dataset (Chaper 6)		
EgoPW (Chaper 5)	189.6	105.3
SceneEgo (Chaper 6)	118.5	92.75
EgoPW* (Chaper 5)	90.96	64.33
SceneEgo* (Chaper 6)	89.06	70.10
Proposed Method-Single	64.19	50.06
Proposed Method-Refined	**57.59**	**46.55**
EgoWholeBody test dataset		
EgoPW* (Chaper 5)	84.21	63.02
SceneEgo* (Chaper 6)	87.57	69.46
Proposed Method-Single	66.28	43.14
Proposed Method-Refined	**60.32**	**40.35**

Table 7.1: Egocentric body pose accuracy of the proposed method on SceneEgo test datasets and EgoWholeBody test dataset. The proposed method outperforms all previous state-of-the-art methods. * denotes the method trained with the datasets in Sec. 7.4.1.

7.4.2 *Comparisons on Whole-Body Pose Estimation*

For a fair comparison with existing methods focusing solely on body or hand pose, the evaluation is split into two parts, reporting results of body poses in Table 7.1 and hand pose in Table 7.2. First, the accuracy of the human body poses from the proposed method is compared with state-of-the-art methods, including EgoPW (Chapter 5) and SceneEgo (Chapter 6), on the EgoWholeBody and SceneEgo test datasets. The comparison with more previous methods and on more evaluation datasets are shown in Sec. D.1 of the Appendix D. Since the motion refinement method incorporates random Gaussian noise, five samples are generated, and the average MPJPE values are calculated. The standard deviation is low ($<$ 0.01mm) and is discussed in Sec. D.6 of supplementary materials. Results are presented in Table 7.1, where the single-frame results are labeled as "Proposed Method-Single" and the refinement results are labeled as "Proposed Method-Refined". The single-frame body pose estimation method outperforms all previous methods by a large margin. The diffusion-based motion refinement method can further improve the accuracy of body poses estimated by the single-frame methods.

Note that previous methods in this thesis and literature (Tomè et al., 2019; Xu et al., 2019) use training datasets different from each other.

Method	MPJPE	PA-MPJPE
SceneEgo test dataset (Chaper 6)		
Hand4Whole (Moon et al., 2022)	49.66	13.85
Proposed Method-Single	23.63	9.59
Proposed Method-Refined	**19.37**	**9.05**
EgoWholeBody test dataset		
Hand4Whole (Moon et al., 2022)	52.85	35.04
Proposed Method-Single	33.10	19.68
Proposed Method-Refined	**28.29**	**14.51**

Table 7.2: Egocentric hand pose accuracy of the proposed method. The proposed method outperforms the Hand4Whole (Moon et al., 2022) on both datasets.

For a fair comparison, previous methods are re-trained with the EgoWholeBody training datasets in Sec. 7.4.1 and show the results with "*" in Table 7.1. This retraining led to significant improvements across all previous methods, demonstrating the EgoWholeBody dataset's broad applicability. However, these methods still underperformed compared to the proposed method, highlighting its superiority.

To evaluate the accuracy of the hand pose estimation method, the hand images are first cropped with the hand detection method in Sec. 7.2.1.3. Then Table 7.2 shows the results of the single-frame hand pose estimation (labeled as "Proposed Method-Single") and whole-body motion refinement methods (labeled as "Proposed Method-Refined"). The single-frame hand pose estimation method outperforms the state-of-the-art method Hands4Whole (Moon et al., 2022), demonstrating the effectiveness of training the network on the EgoWholeBody dataset. The whole-body motion refinement method can also enhance the accuracy of hand motion.

For a qualitative comparison, the body and hand poses of the proposed method are compared with existing methods on the SceneEgo dataset and the in-the-wild EgoPW (Chaper 5) evaluation sequences. The results are shown in Fig. 7.4 and Fig. 7.5, showing that the proposed method can predict high-quality whole-body poses from an egocentric camera.

7.4.3 *Ablation Study*

EGOWHOLEBODY DATASET. Compared to existing egocentric datasets, the EgoWholeBody dataset contains diverse body and hand motions, larger quantity of images, and higher image quality. This is demonstrated by training the body pose estimation network without the EgoWholeBody

Method	MPJPE	PA-MPJPE
Body Pose Results		
w/o EgoWholeBody	75.10	58.62
w/o FisheyeViT	67.36	53.44
w/ Mo^{2}Cap^2 (Xu et al., 2019) head	87.47	65.10
w/ *x*R-egopose (Tomè et al., 2019) head	116.5	95.78
w/ SceneEgo (Chaper 6) head	77.73	62.69
Proposed Method-Single	**64.19**	**50.06**
w/ GlobalEgoMocap (Chapter 4)†	69.83	56.73
w/o uncert. guidance†	62.16	48.40
Only body diffusion	58.95	47.03
Proposed Method-Refined†	**57.59**	**46.55**
Hand Pose Results		
Only hand diffusion	21.69	9.24
Proposed Method-Refined	**19.37**	**9.05**

Table 7.3: Ablation Study on SceneEgo test dataset (Chaper 6). † denotes the temporal-based method.

dataset, using the Mo^{2}Cap^2 (Xu et al., 2019) and EgoPW (Chaper 5)training dataset. The results, labeled as "w/o EgoWholeBody" in Table 7.3, show that performance without the EgoWholeBody dataset is inferior to the proposed method. This highlights that training with the EgoWholeBody dataset enhances the performance of the pose estimation method. This result can also be compared with the evaluation results of existing methods on the SceneEgo test set (Table 7.1). Trained without EgoWholeBody, the proposed approach still outperforms previous methods, showing the effectiveness of the proposed method.

FISHEYEVIT AND POSE REGRESSOR WITH PIXEL-ALIGNED 3D HEATMAP. To assess the individual contributions of FisheyeViT and the pixel-aligned 3D heatmap in the single-frame pose estimation pipeline, experiments are conducted to measure their impact on the overall performance. First, the FisheyeViT module in the single-frame pose estimation method is substituted to ViT (Dosovitskiy et al., 2020). The result is shown in “w/o FisheyeViT” in Table 7.3 and it is worse than the proposed full method. This demonstrates the effectiveness of FisheyeViT in addressing fisheye distortion and feature extraction.

Next, the performance of the single-frame pose estimation network is analyzed when substituting the pose regressor based on the pixel-aligned 3D heatmap with the pose estimation heads from previous works (Tomè et al., 2019; Xu et al., 2019) and SceneEgo (Chaper 6). The results of the three experiments, labeled as "w/ Mo^{2}Cap^2 head", "w/ xR-egopose head" and "w/ SceneEgo head", show a performance drop compared to the proposed full method. This emphasizes the crucial role of the pixel-aligned 3D heatmap in accurately estimating egocentric 3D body joint positions.

DIFFUSION-BASED MOTION REFINEMENT. The effectiveness of the diffusion-based motion refinement is assessed through the following experiments: First, the performance of the diffusion-based motion refinement is compared with GlobalEgoMocap (Chapter 4) by applying the GlobalEgoMocap optimizer on the single-frame body pose estimation results. The result, labeled as "w GlobalEgoMocap" in Table 7.3, indicates that the refinement method outperforms GlobalEgoMocap.

Second, the uncertainty-aware guidance in the motion refinement is removed. Instead, fixed Gaussian denoising steps are used to refine the motion. The result "w/o uncert. guidance" in Table 3, shows that the uncertainty-aware refinement method performs better. The proposed approach relies on the uncertainty values for each joint, using low-uncertainty joints to guide the generation of high-uncertainty joints. This helps reduce errors in joint predictions caused by egocentric self-occlusion, leading to improved results.

Third, the whole-body motion diffusion model is replaced with the separate human body and left/right-hand diffusion models. The accuracy of the refined body and hand motion is shown in "Only body diffusion" and "Only hand diffusion" in Table 7.3. The results show improvements in the accuracy of motion refined by the whole-body diffusion method, proving that learning the whole-body motion prior can help both the refinement of the body and hand motion by learning the correlation between them.

7.5 LIMITATION

Due to serious self-occlusion issues, the proposed method may still predict poses suffering from physical implausibility. This can be solved by introducing the physics-aware motion diffusion models or motion refinement models, such as PhysDiff Yuan et al., 2023 and PhysCap Shimada et al., 2020.

This chapter has demonstrated that FisheyeViT can enhance the performance of egocentric human motion capture using a single fisheye camera. However, the generalization of FisheyeViT remains limited. For instance, the network cannot generalize to fisheye cameras with significantly different model parameters or to those mounted at different locations on the head. This limitation restricts the application of fisheye cameras in many computer vision areas.

Furthermore, the diffusion-based human motion capture shows promising results in refining the initial human pose estimations. However, the inference speed is slow (around 10 seconds per sequence of 196 frames) since multiple diffusion-denoising steps have to be carried out during the inference. To solve this, one possible solution is to use the DDIM (Song et al., 2020) to reduce the diffusion step or use the DDIM inversion to optimize the initial human pose estimations.

7.6 conclusions

This work introduces an innovative approach to capturing egocentric whole-body human motion. The proposed method comprises a single-frame-based whole-body pose estimation process, which includes FisheyeViT and pixel-aligned 3D heatmap representations. To enhance the initial whole-body pose estimates, an uncertainty-aware diffusion-based motion refinement technique is integrated. The experimental results demonstrate that both the single-frame method and the temporal-based method surpass all existing state-of-the-art techniques in terms of both quality and accuracy. Looking ahead, there exists the potential for extending the applications of FisheyeViT to other vision tasks using fisheye cameras. Future work could also involve incorporating facial expressions in whole-body motion capture. In the following chapters, this work will be referred to as **"EgoWholeMocap"**.

The next chapter summarizes all the contributions presented in this thesis, offering several insights and discussing potential future directions.

8 CONCLUSION

This thesis explores human motion capture using a down-facing egocentric fisheye camera mounted in front of the head. It introduces several novel methods. Chapter 4 features an optimization framework that leverages both global and local human motion priors to derive global human motion from the video sequence of a single egocentric camera. Chapter 5 presents a new large-scale in-the-wild dataset, EgoPW, with pseudo-ground truth obtained from an egocentric and an external camera. An adversarial domain adaptation strategy is also employed to bridge the domain gap between synthetic and real data, as well as between egocentric and external views. Chapter 6 presents a novel single-frame egocentric pose estimation method that incorporates scene constraints. Chapter 7 proposes a new method for whole-body egocentric motion capture. It introduces FisheyeViT to address fisheye distortion issues and a diffusion-based refinement strategy to enhance initial human pose estimations.

While each chapter provides its own conclusions, this final chapter draws insights across different chapters and discusses future directions for egocentric motion capture research.

8.1 INSIGHTS

This section discusses insights about motion prior, egocentric datasets and scene-aware egocentric motion capture. These insights are across chapters and go beyond individual contributions.

8.1.1 *Motion Prior*

Chapter 4 and Chapter 7 show that the human motion prior plays an important role in the egocentric human motion capture. First, egocentric motion capture is an ill-posed problem due to severe self-occlusion. By leveraging the human motion prior, the method can reliably infer the plausible location of joints under occlusion. Second, the size of individual egocentric human motion datasets is still far from sufficient. The human motion prior can be built from large-scale human motion datasets and can enable accurate human motion capture in an analysis-by-synthesis manner. Therefore, a human motion prior can deal with the aforementioned challenges and is particularly useful for egocentric motion capture. The application of the motion prior in egocentric motion capture tasks is demonstrated not only in previous chapters of this thesis but also in numerous related works Du et al., 2023; Li et al., 2023; Luo et al., 2024; Van Wouwe et al., 2024; Zhang et al., 2023a.

To apply the motion prior, the common practice is to first capture the motion prior using a generative model. With such a model available, there are multiple options. One option can involve optimizing the initial pose estimations within the latent space of the generative models by considering different constraints, such as 2D poses, scene geometry, and more. Alternatively, a neural network can be applied to project input features onto the latent space of the motion prior.

8.1.2 *Egocentric Dataset*

This thesis presents several egocentric datasets, including synthetic datasets such as EgoGTA in Chapter 6, EgoWholeBody in Chapter 7, and the real-world dataset EgoPW Chapter 5. Several small-scale real-world datasets are also introduced for evaluation in Chapter 4 and Chapter 6. Each type of dataset has its own limitations and advantages. Synthetic datasets are easy to obtain on a large scale but suffer from the synthetic-to-real domain gap. In contrast, real-world datasets are difficult to collect and are constrained by scale and diversity, but they do not suffer from the domain gap between training and evaluation inputs.

To address the issue of different types of datasets and leverage their respective advantages, Chapter 6 and Chapter 7 adopt the strategy of pre-training the egocentric motion capture model on synthetic datasets followed by fine-tuning on real-world datasets using domain adaptation techniques. These chapters demonstrate that employing both synthetic and real-world datasets can combine their strengths and compensate for the disadvantages. This training strategy significantly enhances the generalizability of the egocentric motion capture network.

8.1.3 *Scene-aware Egocentric Motion Capture*

The fisheye camera can capture a large portion of the surrounding scene. The scene information can be utilized to enhance the motion capture performance. Chapter 4 leverages the scene information by applying the SLAM method to obtain the egocentric camera pose in the global space, thereby enhancing the accuracy of egocentric human motion. Chapter 6 employs single-frame depth estimation to obtain the scene geometry and uses the geometry to constrain the egocentric pose, making it plausible under occlusion. These chapters show that the egocentric camera poses and the geometry of the surrounding scene provide valuable information about human movement and interaction with the environment, thereby further enhancing motion capture accuracy.

By analyzing egocentric camera poses, the locomotion of the human body can be inferred and the human poses can be further obtained. For instance, if the human body is moving forward, the movement of legs and feet can be predicted even if they are occluded.

Considering scene context can also significantly alleviate occlusion issues and improve pose estimation accuracy. For example, when a person is sitting at a desk, their feet may be occluded. However, by understanding the semantics of desks and chairs, the sitting pose can be inferred. If the scene geometry is also known, the human pose can be determined accurately.

Utilizing state-of-the-art SLAM methods such as DROID-SLAM (Teed and Deng, 2021), publicly available VR headsets like the Quest3, and scene understanding methods, future research can achieve a precise understanding of human locomotion, scene geometry, and scene semantics. These advancements will further enhance the accuracy of egocentric human motion capture methods.

8.2 challenges and future directions

Humans have long dreamed of the next evolution in mobile technology: wearable computing. In this context, the human body, particularly human motion, plays a pivotal role in the interaction process between humans and wearable devices. This section explores the challenges and future developments in egocentric human motion capture. It also discusses the possible integration with large language models, robotics, the creation of photorealistic avatars, and human-object/scene interactions.

8.2.1 *Egocentric Human Motion Capture*

Though many technical innovations have been presented in this thesis, the egocentric human motion capture remains unsolved. This section demonstrates the challenges and possible future direction of egocentric motion capture.

generalization to a variety of hardware Egocentric motion capture setups can vary significantly, from simple head-mounted cameras to complex multi-sensor rigs. The method that works on one specific setup usually cannot be directly used in other setups. Each configuration may require unique calibration procedures and tailored algorithms to handle specific problems. Consequently, developing a universally applicable solution is exceptionally challenging. In the future, researchers can design flexible frameworks and algorithms that can fit in or be easily customized for different setups. This adaptability is crucial for making egocentric motion capture more practical in real-world scenarios.

FISHEYE CAMERA Though Chapter 7 introduces FisheyeViT to address the distortion issues of fisheye cameras, this problem is still not completely solved. The proposed undistortion methods can not generalize perfectly across different fisheye camera setups. They often require extra finetuning to enhance performance on cameras with different parameters. Other types of parameters, such as the position and orientation of the fisheye camera, also significantly hinder generalizability. Future research could focus on ways to encode fisheye camera parameters into the method to create a camera-agnostic network. Alternatively, developing fast or weakly-supervised fine-tuning methods could enable easier adaptation across different fisheye cameras.

EGOCENTRIC DATASET This thesis presents several synthetic and real-world datasets for the egocentric motion capture task. However, the current datasets are far from sufficient to train and evaluate an egocentric motion capture network. The scale and diversity of the datasets need to be significantly enhanced to meet the demand. For example, the available egocentric motion capture datasets typically contain 100k to 1M images and several subjects (see Chapter 2). In contrast, external-view human motion capture datasets, such as Motion-X (Lin et al., 2024), can include up to 15 million images and feature thousands of different human appearances, as seen in BEDLAM (Black et al., 2023).

Collecting egocentric data in the real world remains challenging since it is still difficult to obtain accurate in-the-wild 3D human motion annotations. Another common method is to synthesize high-quality synthetic datasets, as demonstrated in Chapter 7, UnrealEgo2 (Akada et al., 2024) and BEDLAM (Black et al., 2023). However, these synthetic datasets lack the diversity of human motion and human-scene interactions.

Future work can address the data collection challenges by proposing simple methods to collect 3D in-the-wild ground truth. For the synthetic dataset, generating realistic human-scene interactions in synthetic environment (Li et al., 2024) will also be a promising solution.

SELF-OCCLUSION Self-occlusion, where parts of the body obstruct other parts, poses significant challenges to egocentric motion capture.

Chapter 6 in this thesis aims at tackling this issue by understanding the surrounding scene geometry. However, the resulting human motion still suffers from inconsistencies and physical errors. To mitigate self-occlusion issues as much as possible, future work need to combine human motion priors, efficient temporal-based methods, physics-aware techniques, and simultaneous scene understanding.

REAL-TIME AND ON-DEVICE PROCESSING Real-time processing is a significant challenge in egocentric motion capture. The ability to capture and process motion data in real-time is crucial for applications such as virtual reality and interactive gaming. Additionally, these motion capture algorithms often need to be deployed on VR headsets and AR glasses, which severely limit the computational resources available to the algorithms.

The challenge of real-time and on-device processing often involves a trade-off between accuracy and speed. Future research should focus not only on enhancing pose estimation accuracy but also on exploring parallel processing methods, hardware acceleration techniques, and re-designing algorithms to be compatible with resource-constrained mobile hardware architectures.

8.2.2 *Egocentric Motion Capture + Large Language Models*

The combination of egocentric human motion capture with large language models (LLMs) is promising. By integrating with LLMs, Google's Project Astra and Meta's Ray-Ban glasses are all claimed to enable the "copilot" capability for everyday life by leveraging LLMs in the background. Integrating multi-modal inputs like text, vision, and motion data can exploit LLMs' general reasoning ability to interpret and predict human actions from a first-person perspective. Future research can develop unified models that are capable of handling multiple modalities, including IMU data, text, images, and the context of surrounding scenes.

8.2.3 *Egocentric Photorealistic Avatar*

Combining egocentric human motion capture with photorealistic avatars can create high-quality telepresence with VR/AR devices. Capturing accurate human motion from egocentric devices enables the avatars to exhibit accurate movements and expressions reflecting realistic human motion, which will further enhance immersion in VR applications.

Future research can focus on generating relightable and animatable full-body avatars from a single egocentric video. Recent methods like

Gaussian Codec Avatars (Saito et al., 2024) can drive relightable head avatars from VR headsets. However, full-body avatars from the egocentric perspective are absent, despite being necessary to achieve realistic interactions in the virtual world. Integrating egocentric views into full-body avatars can enable the driving of the human body avatars with natural full-body motion. It can also accurately capture the appearance of the human body, including clothing wrinkles and illumination details.

Another promising research direction is to enable the personalization of avatars based on real-time capture from egocentric devices. This will allow the online reconstruction of clothing based on the egocentric capture.

8.2.4 *Egocentric Motion Capture + Robotics*

One research direction is to drive the motion of robots using human motion captured by egocentric devices. Stereo cameras mounted on the robots can provide input to the screen of a VR headset, while pressure sensors on the robots can deliver human haptic feedback. This allows egocentric wearable devices to be seamlessly applied for robot teleoperation tasks. By utilizing egocentric sensors on VR/AR headsets, researchers can obtain accurate human motion, which can be projected into physics-based simulators or real-world robots for teleoperation and long-range manipulation (Ding et al., 2024; He et al., 2024a). Future research may focus on motion retargeting and collision avoidance. Motion retargeting requires the human motion captured by egocentric devices to be teleported faithfully. Collision avoidance aims at the operation safety concerns. These will enable robots to perform delicate tasks that require high precision and safety, such as manufacturing or surgery.

Egocentric data will enhance the robot's capability of interaction and navigation through the environment. This can be achieved by asking the robots to mimic human activities from large-scale egocentric datasets. By leveraging the egocentric motion capture method, the captures of egocentric sensors can be used to simulate the sensors on the robots and the human motion in the egocentric view serves as a guidance or ground truth.

8.2.5 *Egocentric Motion Capture + Human-Object/Scene Interactions*

The combination of egocentric motion capture and large-scale hand-object/scene interaction datasets opens up a promising research avenue. Projects like the Epic-Kitchens dataset (Damen et al., 2018) and the Ego4D dataset (Grauman et al., 2022) collect extensive egocentric video footage of everyday activities. By leveraging human motion information obtained by egocentric motion capture, future researchers can learn how humans interact with objects on these large-scale datasets.

8.3 FINAL CONCLUSION

Capturing human motion from wearable egocentric devices unlocks numerous applications, including VR/AR, gaming, robotics teleoperation, healthcare, sports, and movie/animation production. Accurate egocentric human motion capture will also set the foundation for realistic virtual humans. These virtual humans, captured by egocentric devices, will be able to interact with 3D scenes, move naturally, communicate with each other in the virtual world, and further pave the way for the metaverse.

In order to achieve accurate and fast egocentric human motion capture, a great number of challenges are waiting to be solved. Among these challenges, this thesis focuses on a specific egocentric motion capture setup: a single downward-facing camera mounted in front of the head. New datasets and novel methods have been introduced, aimed at enhancing the accuracy and the range of applicability of egocentric motion capture. This thesis further discusses future research directions, not only to improve accuracy and efficiency but also for diverse application scenarios like egocentric avatars. With the fast development of VR headsets and AR glasses, the applications leveraging egocentric motion capture technology will be eventually democratized.

A APPENDIX FOR CHAPTER 4

A.1 COMPARISONS ON DIFFERENT TYPES OF MOTIONS

Method	Mo^2Cap^2 +SLAM	Mo^2Cap^2+SLAM +Smooth	Mo^2Cap^2 +Proposed
walking	38.41	37.35	**35.39**
sitting	70.94	64.45	**60.83**
crawling	94.31	87.41	**75.45**
crouching	81.90	69.68	**63.15**
boxing	48.55	45.19	**40.14**
dancing	55.19	54.76	**53.05**
stretching	99.34	90.89	**84.96**
waving	60.92	49.41	**44.10**
total (mm)	61.40	58.25	**52.90**

Table A.1: The BA-MPJPE of different types of motions on the indoor sequence of Mo2Cap2 dataset (Xu et al., 2019). When based on the local poses estimated by Mo^2Cap^2, the proposed approach improves the Mo^2Cap^2 (Xu et al., 2019) results by 13.8% (8.5 mm).

In Table 4.1 of Chapter 4, The quantitative comparison is shown between the proposed method and the state-of-the-art methods: Mo^2Cap^2 and xR-egopose. In order to further compare the performance on different types of motions, the quantitative comparison on Mo^2Cap^2 (Xu et al., 2019) is shown in Table A.1 and Table A.2. The comparisons on different motions of the test dataset is shown in Table A.3 and Table A.4. In these tables, the BA-MPJPE results of the smoothed global pose of Mo^2Cap^2 and xR-egopose are also demonstrated to give a fair comparison. In the aforementioned results, the proposed method outperforms all of the baselines on every type of motion.

A.2 THE STRUCTURE OF RNN-BASED VAES

In the Sec. 4.3.5 of the main paper, the performance of the proposed CNN-based sequential VAE is compared with the MLP-based VAE and RNN-based VAEs in VIBE (Kocabas et al., 2020) and MEVA (Luo et al.,

Method	Mo^2Cap^2 +SLAM	Mo^2Cap^2+SLAM +Smooth	Mo^2Cap^2 +Proposed
walking	39.69	38.68	**33.66**
sitting	63.64	63.20	**60.34**
crawling	64.90	63.84	**62.33**
crouching	61.22	60.49	**55.67**
boxing	47.87	46.53	**44.24**
dancing	58.37	57.20	**51.29**
stretching	84.64	84.19	**82.63**
waving	53.99	52.58	**46.81**
total (mm)	55.43	54.03	**50.52**

Table A.2: The BA-MPJPE of different types of motions on the indoor sequence of Mo2Cap2 dataset (Xu et al., 2019). When based on the local poses estimated by xR-egopose (Tomè et al., 2019), the proposed method improves the xR-egopose results by 8.9% (4.9 mm).

2020). The implementation details of the aforementioned VAEs will be described in this section.

RNN-BASED VAE IN VIBE The VIBE (Kocabas et al., 2020) explored the performance of RNN-based VAE as a loss term in the training of the VIBE network. At time step t, the body pose $\mathcal{P}_t$ with shape (15, 3) is firstly flattened and put in the encoder. The encoder gives the $\mu_t, \sigma_t \in \mathbb{R}^{2048}$ and the latent vector z_t is sampled from them. The latent vector z_t is put into the decoder and reconstructs the body pose $\mathcal{P}_t$ at time step t. The encoder and decoder are two-layer GRU networks with 512 hidden dimensions.

RNN-BASED VAE IN MEVA The structure of RNN-based VAE is shown in MEVA (Luo et al., 2020) and the code is released in `https://github.com/ZhengyiLuo/MEVA`. Their implementation is directly used in the experiment.

Note that the structure of RNN-based VAE in MEVA is different from the VAE in VIBE. The VAE in VIBE gets different μ and σ for each time step and uses the different latent vector z as the decoder input for each time step. In the VAE of MEVA, the latent vector is obtained with a pooling layer and works as the first input of the RNN-based decoder.

Method	Mo^2Cap^2 +SLAM	Mo^2Cap^2+SLAM +Smooth	Mo^2Cap^2 +Proposed
walking	69.68	66.68	**57.30**
running	77.88	74.14	**66.78**
crouching	63.28	60.76	**56.05**
boxing	79.37	75.59	**67.57**
dancing	82.65	76.88	**61.43**
stretching	117.7	114.9	**107.5**
waving	53.14	49.31	**42.77**
playing balls	60.95	57.69	**53.30**
open door	55.88	53.33	**46.27**
play golf	113.8	104.4	**94.17**
talking	53.93	50.65	**48.16**
shooting arrow	67.07	62.82	**57.58**
sitting	83.24	78.70	**50.89**
total (mm)	74.46	70.84	**62.07**

Table A.3: The BA-MPJPE of different types of motions on the test set. When based on Mo^2Cap^2 (Xu et al., 2019), the proposed approach outperforms Mo^2Cap^2 results by 16.6% (12.4 mm).

Method	*x*R-egopose +SLAM	*x*R-egopose +SLAM +Smooth	*x*R-egopose +Proposed
walking	84.20	82.96	**60.72**
running	76.78	74.43	**64.92**
crouching	96.86	96.53	**75.11**
boxing	85.74	83.67	**63.45**
dancing	94.23	92.42	**64.78**
stretching	119.9	119.7	**116.3**
waving	72.66	71.83	**46.38**
playing balls	95.30	93.94	**58.49**
open door	71.70	70.80	**45.86**
play golf	94.41	92.58	**83.25**
talking	78.10	75.84	**46.90**
shooting arrow	76.75	74.82	**57.86**
sitting	69.10	63.89	**55.97**
total (mm)	87.20	84.70	**64.31**

Table A.4: The BA-MPJPE of different types of motions on the test set. When based on *x*R-egopose (Tomè et al., 2019), the proposed method outperforms *x*R-egopose results by 26.2% (22.9 mm).

MLP-BASED VAE The input pose sequence with n frames is firstly reshaped to a vector with length $n \times 15 \times 3$ and fed into the encoder with $n \times 15 \times 3$ input dimensions. The encoder has five 1D fully connected blocks with 512, 512, 1024, 2048 and 2048 output dimensions. Each fully connected block contains one fully connected layer, one batch norm layer and one leaky relu layer with negative slope=0.01. The output of the encoder is sent into two linear layers giving $\mu, \sigma \in \mathbb{R}^{2048}$. The latent vector z is sampled with μ, σ with the reparameterization trick.

For the decoder, the sampled latent vector z is firstly fed into a linear layer with 2048 output dimension, and five 1D fully connected blocks with 2048, 2048, 1024, 512 and 512 output channels. Each block contains one fully connected layer, one batch norm layer and one leaky relu layer with the same hyper-parameters as the encoder. The output vector is obtained from a final fully connected layer. The output vector is eventually reshaped to $(n, 15, 3)$, representing a pose sequence as the input.

B APPENDIX FOR CHAPTER 5

B.1 QUANTITATIVE RESULTS ON DIFFERENT MOTIONS

In Chapter 5, the proposed method is proved to outperform the state-of-the-art methods: Mo^{2}Cap^2 and xR-egopose. To further compare the performance on different types of motions, the quantitative comparisons on the test dataset of Chapter 4 are shown in Table B.1. The results on Mo^{2}Cap^2 dataset (Xu et al., 2019) are shown in Table B.2. The proposed method outperforms all of the baselines on most types of motion in these results. Note that this method is trained on the EgoPW dataset while the focal length and distortion of the fisheye camera in the EgoPW dataset is different from the fisheye camera used in Mo^{2}Cap^2, which affects the performance of the proposed method on the Mo^{2}Cap^2 test dataset.

B.2 QUALITATIVE RESULTS

This section shows more qualitative results for the in-the-wild images from the test sequence of either EgoPW in Figure B.1 or Mo^{2}Cap^2 in Figure B.2. These results show that the proposed method significantly outperforms the state-of-the-art methods especially when the body parts are occluded.

B.3 DETAILS AND COMPARISONS OF EGOPW DATASET

The details of the EgoPW dataset and comparisons between EgoPW and other 3D pose estimation datasets are shown in Table B.3. This dataset contains 97 sequences and 318k frames in total, which is performed by 10 actors in 20 clothing styles. The actions in the EgoPW dataset include *reading magazine/newspaper*, *playing board games*, *doing a presentation*, *walking*, *sitting down*, *using a computer*, *calling on the phone*, *drinking water*, *writing on the paper*, *writing on the whiteboard*, *making tea*, *cutting vegetables*, *stretching*, *running*, *playing table tennis*, *playing baseball*, *climbing floors*, *dancing*, *opening the door*, and *waving hands*.

To synchronize the egocentric and external camera setup, a mobile phone screen is used to play a video of mostly black frames with a single white frame every 10 seconds, observed by both cameras. Recording is started on both cameras, and synchronization is achieved when the white

Method	Mo^{2}Cap^2	xR-egopose	Ours
walking	69.68	84.20	**59.65**
running	77.88	76.78	**63.84**
crouching	**63.28**	96.86	68.87
boxing	79.37	85.74	**72.91**
dancing	82.65	94.23	**65.21**
stretching	117.7	119.9	**108.8**
waving	53.14	72.66	**44.57**
playing balls	60.95	95.30	**56.54**
open door	55.88	71.70	**49.06**
play golf	113.8	94.41	**94.29**
talking	53.93	78.10	**51.82**
shooting arrow	67.07	76.75	**60.71**
sitting	83.24	69.10	**65.06**
total (mm)	74.46	87.20	**64.87**

Table B.1: The BA-MPJPE of different types of motions on the test set of Chapter 4. The proposed approach outperforms Mo^{2}Cap^2 results by 9.59 mm and outperforms xR-egopose results by 22.33 mm.

frame is observed. This white frame is used to temporally synchronize the egocentric and external recordings, and further verification is done through hand-clapping movements. Calibration is performed only once at the start of data recording.

In Table B.3, the EgoPW dataset is further compared with other datasets for external-view 3D pose estimation and egocentric view 3D pose estimation. Mo^{2}Cap^2 Xu et al., 2019 and xR-egopose Tomè et al., 2019 provide large synthetic datasets for training the egocentric pose estimation networks. However, these datasets are synthesized and thus suffer from the domain gap with the real images. Mo^{2}Cap^2, xR-egopose, and Chapter 4 also provide small test sequences with ground truth labels obtained with the mocap system. However, these datasets are not sufficient for training an egocentric pose estimation network. The EgoPW dataset contains a large amount of in-the-wild images with accurate pseudo labels generated with an optimization framework, which facilitates training the pose estimation network with in-the-wild images.

The publicly available large datasets for 3D pose estimation from an external view, like Human 3.6M (Ionescu et al., 2013) and MPI-INF-3DHP (Mehta et al., 2017a), are all collected in the studio with a multi-view mocap system. This capturing method is not able to obtain in-the-wild images and the interactions between the human body and the

Indoor	walking	sitting	crawling	crouching	boxing	dancing
Mo^{2}Cap^2	38.41	70.94	94.31	81.90	48.55	55.19
*x*R-egopose	**37.35**	64.45	87.41	69.68	**45.19**	54.76
Proposed	40.23	**60.22**	**70.88**	**62.40**	49.89	**52.41**
Indoor	stretching	waving	total (mm)			
Mo^{2}Cap^2	99.34	60.92	61.40			
*x*R-egopose	90.89	**49.41**	55.43			
Proposed	**82.48**	59.60	**54.78**			
Outdoor	walking	sitting	crawling	crouching	boxing	dancing
Mo^{2}Cap^2	63.10	**85.48**	96.63	92.88	96.01	68.35
*x*R-egopose	62.01	103.45	86.53	80.43	90.48	66.06
Proposed	**58.06**	94.19	**85.50**	**77.61**	**83.91**	**62.56**
Outdoor	stretching	waving	total (mm)			
Mo^{2}Cap^2	123.56	**61.42**	80.64			
*x*R-egopose	117.55	67.49	78.30			
Proposed	**111.9**	65.37	**74.55**			

Table B.2: The BA-MPJPE of different types of motions on the indoor and outdoor sequence of Mo^{2}Cap^2 dataset (Xu et al., 2019). In the indoor sequence, the proposed method improves the Mo^{2}Cap^2 results by 6.62 mm and *x*R-egopose results by 0.65 mm; In the outdoor sequence, the proposed method improves the Mo^{2}Cap^2 results by 6.09 mm and *x*R-egopose results by 3.75 mm.

environment. 3DPW (Von Marcard et al., 2018) is a dataset collected in the in-the-wild scenes with pseudo labels obtained from a moving camera and an IMU system. This capturing method provides accurate pseudo labels for body pose with various interactions between the human body and the environment. However, this dataset only contains 51k frames, which is less than the frames in the EgoPW dataset. All of the aforementioned datasets do not contain any egocentric images and thus cannot be used for training the egocentric pose estimation networks.

B.4 network architecture

This section describes the architecture of the pose estimation network and the domain classifier network used in the method.

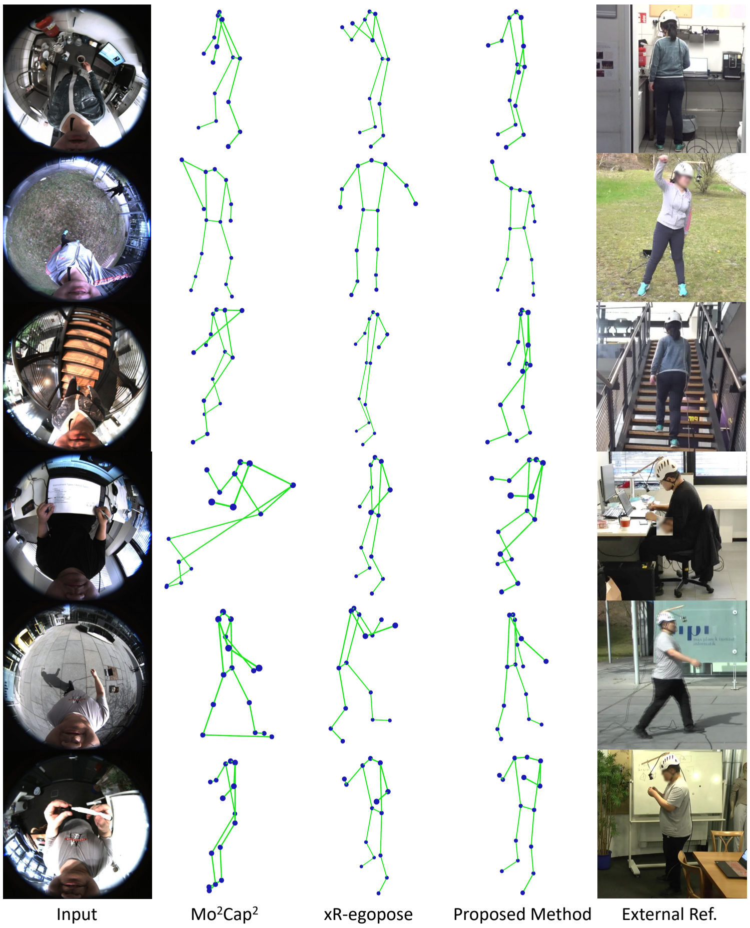


Figure B.1: Qualitative comparison between the proposed method and the state-of-the-art methods on the test images of the EgoPW dataset. From left to right: input image, Mo^{2}Cap^2 result, xR-egopose result, the result of the proposed method, and external image. Note that the external images are only for visualization and they are not used for predicting the pose.

B.4.1 *Pose Estimation Network*

The architecture in Mo^{2}Cap^2 Xu et al., 2019 is used for obtaining the 3D poses and 2D heatmaps. The pose estimation network contains a 2D module for the full-body heatmap, a 2D module for the zoomed-in body heatmap, and a 3D module. The 2D module for full-body pose can be represented as an encoder-decoder network, which first gets the features $\mathcal{F}_{Full2D}$ with a Resnet-50 network He et al., 2016 as the encoder and uses the features $\mathcal{F}_{Full2D}$ to predict the full-body heatmap with convolutional layers. The 2D module for zoomed-in body heatmaps has the same architecture as the former one. It takes the zoomed-in

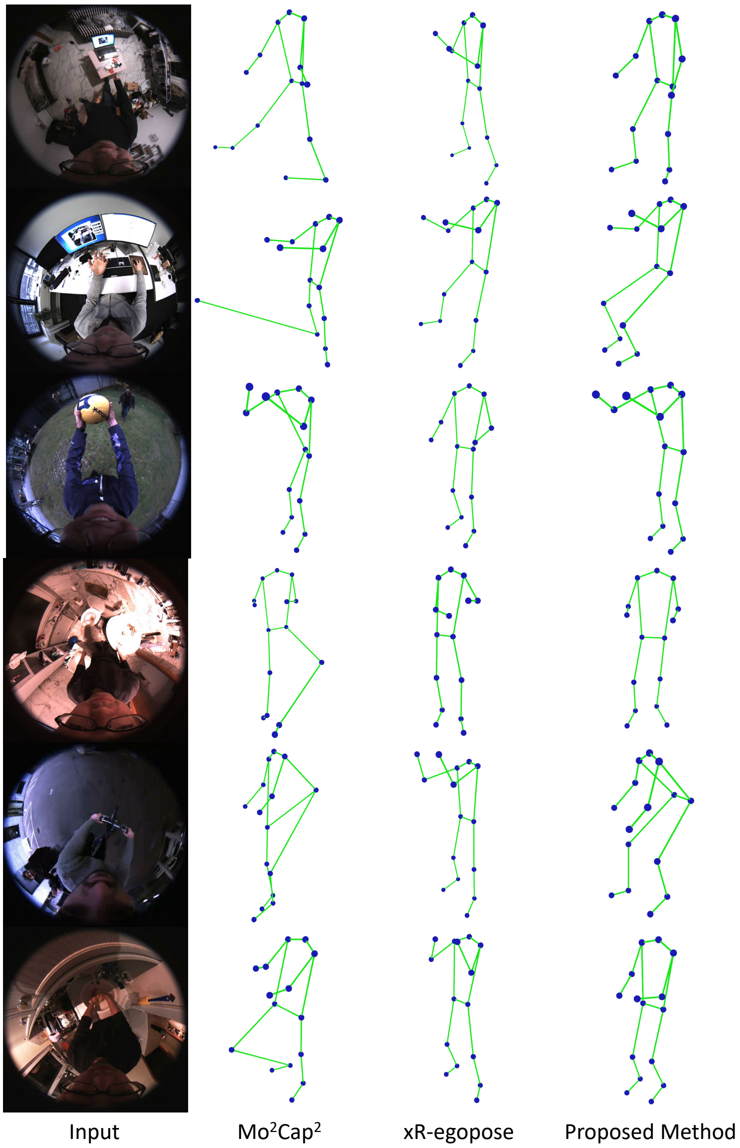


Figure B.2: Qualitative comparison between the proposed method and the state-of-the-art methods on the test images of Mo^{2}Cap^2 work. From left to right: input image, Mo^{2}Cap^2 result, *x*R-egopose result, and the result of the proposed method.

Dataset Name	Frames	Sequences	Subjects	Context	Action Types
Ionescu et al. (2013)	3.6M	1376	11	Studio	17
Mehta et al. (2017a)	1.3M	64	16	Studio	8
Von Marcard et al. (2018)	51k	60	18	In the wild	8
Xu et al. (2019)	530k	-	700	Synthetic	3000
Mo^{2}Cap^2-test	5591	2	2	Studio & in the wild	8
Tomè et al. (2019)	383k	-	-	Synthetic	9
xR-egopose-test	10k	-	3	Studio	6
Chapter 4	47k	19	9	Studio	13
EgoPW	318k	97	10	In the wild	20

Table B.3: Comparison between the EgoPW dataset and publicly available 3D pose estimation datasets.

egocentric images as input and first generates features $\mathcal{F}_{Zoom2D}$ and predicts zoomed-in heatmaps from the intermediate features. The full-body heatmaps and zoomed-in heatmaps are finally averaged to get the final prediction of heatmaps $\hat{\mathcal{H}}$. The distance module takes the features from both the aforementioned 2D modules as input and predicts the distances $\hat{\mathcal{D}}$ between body joints and the camera. More details about the pose estimation network can be found in Mo^{2}Cap^2 Xu et al., 2019.

B.4.2 *Domain Classifier*

The domain classifier takes the intermediate features $\mathcal{F}_{Full2D}$ with shape $2048 \times 8 \times 8$ or $\mathcal{F}_{Zoom2D}$ with shape $2048 \times 8 \times 8$ as input and predicts whether the input feature is from synthetic or real image. The network contains two Resnet "bottleneck" blocks He et al., 2016 with 1024 and 256 output channels and one final classification block. The classification block contains two convolutional blocks and a linear layer for the domain classification task. The first convolutional block contains one 2D convolutional layer (kernel size=4, stride=2, and padding=1), one batch norm layer, and one relu layer. The second convolutional block contains one 2D convolutional layer (kernel size=3, stride=2, and padding=1), one batch norm layer, and one relu layer. The output features of the convolutional blocks are sent to the linear layer giving the domain label prediction.

B.4.3 *Egocentric-external View Classifier*

Similar to the domain classifier for distinguishing synthetic and real images, the egocentric-external view classifier also takes the intermediate features $\mathcal{F}_{Full2D}$ with shape $2048 \times 8 \times 8$ or $\mathcal{F}_{Zoom2D}$ with shape $2048 \times 8 \times 8$ as input and predicts whether the input feature is from the egocentric view or the external view. The network contains two convolutional blocks, one global average pooling layer, and one final classification block. The intermediate features are firstly sent to the convolutional blocks and then the generated features with shape $1024 \times 8 \times 8$. The spatial dimension of the features is eliminated with a global average pooling layer (Zhou et al., 2016a) to generate a feature vector with length 1024. Next, the feature vector is sent to the final classification block to predict whether the input feature is from the egocentric view or the external view. Each of the convolutional blocks consists of one 2D convolutional layer (output channel=1024, kernel size=3, stride=2, and padding=1), one batch norm layer, and one relu layer. The classification block includes one fully connected layer (output dimension=256), one batch norm layer, one relu layer, and one final fully connected layer (output dimension=2) which predicts the labels of egocentric/external views.

B.5 ENERGY FUNCTION

This section describes some of the terms in the objective function (Eq. B.1).

$$\begin{aligned} E(\mathcal{P}^{ego}_{seq}, R_{seq}, t_{seq}) = {} & \lambda^{ego}_R E^{ego}_R + \lambda^{ext}_R E^{ext}_R + \lambda^{ego}_J E^{ego}_J \\ & + \lambda^{ext}_J E^{ext}_J + \lambda_T E_T + \lambda_B E_B \\ & + \lambda_C E_C + \lambda_M E_M \end{aligned} \tag{B.1}$$

In this function, E^{ext}_R, E^{ext}_J, E_C, and E_M are the external reprojection term, external 3D pose regularization term, camera pose consistency term, and camera matrix regularization term respectively which have already been described in the paper. E^{ego}_R, E^{ego}_J,E_T, and E_B are the egocentric reprojection term, egocentric pose regularization term, motion smoothness regularization term and bone length regularization term, which are the same as the corresponding terms in Chapter 4. These terms are also depicted here:

HEATMAP-BASED REPROJECTION: This term maximizes the summed heatmap values at the reprojected 2D joint positions:

$$E_R(\mathcal{P}_{seq}^{ego}) = -\sum_{i=1}^{B} \|\mathrm{HM}_i(\Pi(\mathcal{P}_i^{ego}))\|_2^2 \tag{B.2}$$

where $\mathrm{HM}_i(.)$ returns the value at a pixel on $\mathcal{H}_i^{ego}$, the heatmap of i-th frame. $\Pi(.)$ refers to the projection of a 3D point with the fisheye camera model.

POSE REGULARIZATION: The pose regularizer is defined to constrain the optimized pose $\mathcal{P}_i^{ego}$ to stay close to the initial pose $\widetilde{\mathcal{P}}_i^{ego}$.

$$E_J(\mathcal{P}_{seq}^{ego}, \widetilde{\mathcal{P}}_{seq}^{ego}) = \sum_{i=1}^{B} \left\| \mathcal{P}_i^{ego} - \widetilde{\mathcal{P}}_i^{ego} \right\|_2^2 \tag{B.3}$$

MOTION SMOOTHNESS REGULARIZATION: This term constrains the acceleration of each joint over the whole sequence to improve the temporal stability of the estimated poses:

$$E_T(\mathcal{P}_{seq}^{ego}) = \sum_{i=2}^{B} \|\nabla\mathcal{P}_i^{ego} - \nabla\mathcal{P}_{i-1}^{ego}\|_2^2 \tag{B.4}$$

where $\nabla\mathcal{P}_i^{ego} = \mathcal{P}_i^{ego} - \mathcal{P}_{i-1}^{ego}$.

BONE LENGTH REGULARIZATION: This term calculates the difference between the bone length and the average bone length to enforce the length of each bone to be consistent.

$$E_B(\mathcal{P}_{seq}^{ego}) = \sum_{i=1}^{B} \left\| L_{\mathcal{P}_i^{ego}} - \frac{1}{B}\sum_{j=1}^{B} L_{\mathcal{P}_j^{ego}} \right\|_2^2 \tag{B.5}$$

where the $L_{\mathcal{P}_i^{ego}}$ is the length of each bone of 3D pose $\mathcal{P}_i^{ego}$.

APPENDIX FOR CHAPTER 6

C

C.1 DATASETS

C.1.1 *SceneEgo Test Dataset*

This section introduces the data collection process for the SceneEgo test dataset. All personal data in the SceneEgo test dataset is collected with IRB approval. To estimate accurate egocentric camera poses and further obtain the ground truth human body poses under the egocentric camera perspective, a calibration board is mounted on the *head*, rigidly attached to the egocentric camera, and estimate the pose of the egocentric camera with a multi-view capturing system, as shown in Fig. C.1.

Before the data collection process, the transformation matrix $\mathbf{M}_{\text{head2ego}}$ is first estimated between the calibration board and the fisheye camera with hand-eye calibration (Tsai and Lenz, 1988). A second calibration board is placed on the scene in a place where it can be seen by both the egocentric camera and the studio cameras. Then this method estimates the relative pose $\mathbf{M}_{\text{ego2calib}}$ between the egocentric camera and the external calibration board, the relative pose between the studio cameras and the external calibration board $\mathbf{M}_{\text{ext2calib}}$, and the relative pose between the studio cameras and the head-mounted calibration board $\mathbf{M}_{\text{ext2head}}$. Finally, the transformation matrix $\mathbf{M}_{\text{head2ego}}$ can be obtained with:

$$\mathbf{M}_{\text{head2ego}} = \mathbf{M}_{\text{ext2head}}^{-1}\mathbf{M}_{\text{ext2calib}}\mathbf{M}_{\text{ego2calib}}^{-1} \tag{C.1}$$

During the data collection process, the pose of the calibration board is estimated from every single view, and the estimated calibration board poses are further averaged to get the result $\mathbf{M}_{\text{ext2head}}$ (see Fig. C.1). The egocentric camera pose $\mathbf{M}_{\text{ext2ego}}$ can be obtained with:

$$\mathbf{M}_{\text{ext2ego}} = \mathbf{M}_{\text{ext2head}}\mathbf{M}_{\text{head2ego}} \tag{C.2}$$

With the egocentric camera pose, the ground truth pose under the studio camera coordinate system P_{ext} can be transformed to the egocentric camera coordinate system P_{ego}:

$$P_{\text{ego}} = P_{\text{ext}}\mathbf{M}_{\text{ext2ego}} \tag{C.3}$$

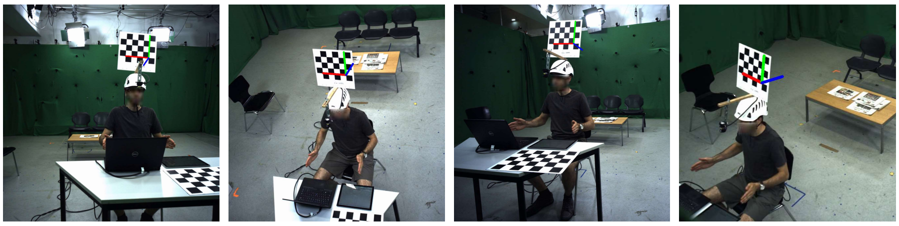


Figure C.1: Visualization of the data collection process for the SceneEgo test dataset. The pose of the head-mounted calibration board, rigidly attached to the egocentric camera, is detected from multiple views in the studio.

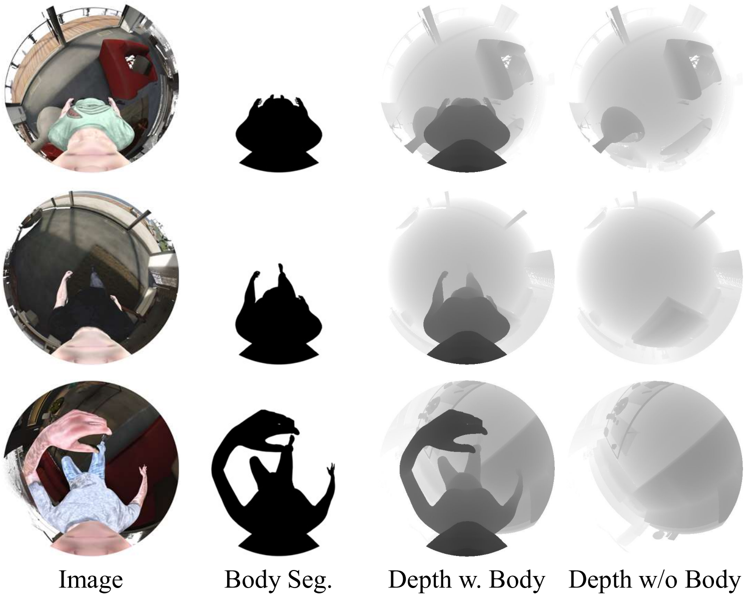


Figure C.2: Example of the EgoGTA dataset.

C.1.2 *EgoGTA Dataset*

Based on the GTA-IM dataset (Cao et al., 2020), the synthetic EgoGTA dataset is generated with ground truth labels for human body segmentation masks, scene depth maps, and human body poses. First, the SMPL-X model is registered on the 3D poses from GTA-IM following HULC (Shimada et al., 2022). Then, the TSDF fusion (Curless and Levoy, 1996) is used to reconstruct the mesh of the scene from the depth map sequences in GTA-IM. Finally, images, semantic labels, and depth maps of the scene with and without the human body are rendered using Blender (*Blender* n.d.). More examples of the EgoGTA dataset are shown in Fig. C.2

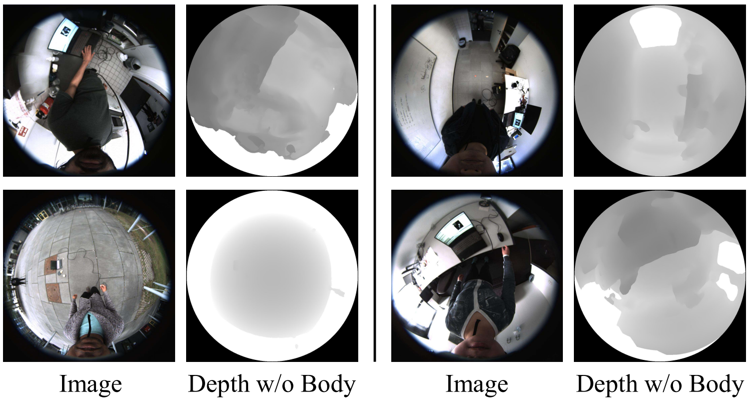


Figure C.3: Example of the EgoPW-Scene dataset.

C.1.3 *EgoPW-Scene Dataset*

The EgoPW-Scene dataset is generated by rendering the scene depth map for each image in the EgoPW dataset (Chapter 5). Since the scan of the background scene is not available for the EgoPW dataset, the mesh of the scene is generated from the EgoPW image sequences with SfM. More examples of the EgoPW-Scene dataset are shown in Fig. C.3.

C.2 IMPLEMENTATION DETAILS

This section describes the implementation details of the scene-aware egocentric pose estimation framework, including the network architectures and training procedure. Details of the scene depth estimator are provided in Sec. C.2.1, which includes a human body segmentation network (Sec. C.2.1.2), a depth estimation network (Sec. C.2.1.1) and a depth inpainting network (Sec. C.2.1.3). The details of the scene-aware egocentric pose estimator are shown in Sec. C.2.2.

C.2.1 *Scene Depth Estimator*

C.2.1.1 *Depth Estimation Network with Human Body*

The same network architecture from Hu et al. (2019)'s work is used as the depth estimation network $\mathcal{D}$. The network $\mathcal{D}$ is trained on the NYU-Depth V2 dataset (Nathan Silberman and Fergus, 2012) following the training procedure from Hu et al. (2019). Next, the network is finetuned on the EgoGTA dataset using the Adam optimizer (Kingma and Ba, 2014) for 40K iterations with the learning rate set to $1 \cdot 10^{-4}$, the weight decay set to $1 \cdot 10^{-4}$, the image size as 256×256, and the batch size as 16.

C.2.1.2 *Human Body Segmentation Network*

The HRNetV2-W48 network from Yuan et al. (2020b)'s work is adopted as the human body segmentation network $\mathcal{S}$. The network $\mathcal{S}$ is trained on the LIP dataset (Gong et al., 2017) following the procedure from Yuan et al. (2020b). Next, the network is finetuned on the EgoGTA dataset for 2000 steps with the weight decay as $1 \cdot 10^{-3}$, the image size as 473×473, and the batch size as 32. During the finetuning step, the Adadelta (Zeiler, 2012) optimizer is used. The learning rate of the first 3 stages in HRNet is set to $1 \cdot 10^{-6}$ and the learning rate of the fourth stage is set to 0.001.

C.2.1.3 *Depth Inpainting Network*

The depth inpainting network $\mathcal{G}$ takes the segmented depth map $\hat{D}^M$ with shape 256×256 and the human body segmentation mask $\hat{S}$ with shape 256×256 as the input and predicts the scene depth map without human body $\hat{D}^S$. The UNet (Ronneberger et al., 2015) is adopted for the depth inpainting task. The encoder of the UNet contains one input convolutional layer with 64 output channels and 4 downsampling layers, each with 128, 256, 512, 512 output channels. Each downsampling layer consists of one 2D-maxpooling layer (kernel size 2) and two convolutional blocks. The decoder contains 4 upsampling layers, each with 256, 128, 64, 64 output channels, and one output convolutional layer with 1 output channel. Each upsampling layer consists of one 2D-bilinear interpolation layer and two convolutional blocks. Each aforementioned convolutional block contains one 2D convolutional layer (kernel size 3, stride 1, and padding 1), one batch norm layer, and one relu layer. The 1st, 2nd, 3rd and 4th input of the downsampling layers is also fed into the 4th, 3rd, 2nd and 1st input of the upsampling layer to form the skip connections in UNet.

The depth inpainting network is trained on the EgoGTA and the EgoPW-Scene datasets simultaneously using the Adam optimizer (Kingma and Ba, 2014) for 28K iterations with the learning rate as $1 \cdot 10^{-4}$, the weight decay as $1 \cdot 10^{-4}$ and batch size as 16.

C.2.2 *V2V Network*

The V2V network has the same architecture as the network in Moon et al. (2018)'s work. During training, the input image is converted to 2D body pose features and further projected into a 3D volumetric space V_{body} with 32 channels. Next, the network concatenates the volumetric body feature, the volumetric representation of ground truth scene geometry V_{scene}, and their intersection V_{inter} and feeds them to the V2V network. The network

Method	PA-MPJPE	BA-MPJPE
EgoPW+Optimizer	79.06	63.56
EgoPW+Depth	78.41	63.75
xR-egopose+Depth	109.7	85.74
w/o Depth	81.04	64.18
+Depth with Body	82.98	65.09
+Depth w/o Body	78.95	64.83
+Depth w/o Inpainting	82.39	66.43
Proposed Method	**76.50**	**61.92**

Table C.1: Results from the proposed method compared to different baselines.

is trained using the Adam optimizer for 24K iterations with the learning rate as $1 \cdot 10^{-3}$ and the batch size as 64.

C.3 EVALUATION METRICS

This section gives a detailed explanation of the evaluation metrics used in the proposed method. MPJPE is the mean of Euclidean distances for each joint in the predicted and ground truth poses. For PA-MPJPE, the estimated pose is rigidly aligned to the ground truth pose using Procrustes analysis before calculating MPJPE. For BA-MPJPE, the bone lengths of predicted poses and ground truth poses are first resized to match the bone length of a standard human skeleton. Then, the PA-MPJPE is calculated between the two resulting poses.

C.4 ABLATION STUDY ON TEST DATASET IN CHAPTER 4

In this section, the ablation study in Sec. 6.3.4 in the main paper is re-evaluated on the test dataset in Chapter 4. The results are shown in Table C.1. Since this dataset does not provide the ground truth scene geometry and ground truth pose annotations in the egocentric camera coordinate system, the experiments only evaluate PA-MPJPE and BA-MPJPE metrics.

The ablation study shows similar results on Wang *et al.*'s dataset as on the SceneEgo test dataset, which demonstrates similar conclusions in Sec. 4.4 in the main paper.

D APPENDIX FOR CHAPTER 7

D.1 FULL COMPARISON WITH EXISTING EGOCENTRIC POSE ESTIMATION METHODS

The comparison results between the proposed method and all previous methods (Liu et al., 2023b; Park et al., 2023; Tomè et al., 2019; Xu et al., 2019) and Chapters 4, 5, 6 are shown in Table D.1, Table D.2, Table D.3 and Table D.4. "*" indicates that the methods are re-trained with the EgoWholeBody training dataset. Since the GlobalEgoMocap (Chapter 4) can be applied to refine the egocentric human body motion predicted from any egocentric pose estimation method, this experiment bases the method on Mo^{2}Cap^2 (Xu et al., 2019) following the original setting in GlobalEgoMocap (Chapter 4). The GlobalEgoMocap results in Mo^{2}Cap^2 test dataset (Xu et al., 2019) are also not shown since it does not provide egocentric camera poses for all of the sequences. Note that the EgoWholeBody dataset does not contain ground truth scene geometry annotations. Therefore, Therefore, the weights of the depth estimation module in SceneEgo (Chapter 6) are frozen, and only the human pose estimation part is trained.

The results in Table D.1, Table D.2, and Table D.3 show the single-frame method and the refinement method consistently outperforms all of the previous methods, even if they are trained on the new EgoWholeBody dataset, which further strengthens the claim in the experiment section (Sec. 7.4.2).

D.2 IMPLEMENTATION DETAILS

In this section, the implementation details of the methods are described. NVIDIA RTX8000 GPUs are used for all experiments.

D.2.1 *FisheyeViT and Pose Regressor with Pixel-Aligned 3D Heatmap*

D.2.1.1 *Network Structure*

FISHEYEVIT In FisheyeViT, the image patches are first undistorted with the method in Sec. 7.2.1.1, then the patches are fed into a ViT transformer. In the ViT transformer, the embedding dimension is 768, the network

Method	MPJPE	PA-MPJPE
SceneEgo test dataset (Chapter 6)		
Mo^{2}Cap^2 (Xu et al., 2019)	200.3	121.2
GlobalEgoMocap†	183.0	106.2
xR-egopose (Tomè et al., 2019)	241.3	133.9
EgoPW	189.6	105.3
SceneEgo	118.5	92.75
Mo^{2}Cap^2* (Xu et al., 2019)	92.20	66.01
GlobalEgoMocap*†	89.35	63.03
xR-egopose* (Tomè et al., 2019)	121.5	98.84
EgoPW*	90.96	64.33
SceneEgo*	89.06	70.10
Proposed Method-Single	64.19	50.06
Proposed Method-Refined†	**57.59**	**46.55**

Table D.1: Performance of the proposed method on SceneEgo test datasets (Chapter 6). The proposed method outperforms all previous state-of-the-art methods. $*$ denotes the method trained with the datasets in Sec. 7.4.1. † denotes the temporal-based methods.

depth is 12, the attention head number is 12, the expansion ratio of the MLP module is 4, and the drop path rate is 0.3. The output sequence from the transformer (with a length equal to 256) is reshaped to a 2D feature map with size 16×16.

POSE REGRESSOR WITH PIXEL-ALIGNED 3D HEATMAP The regressor of the pixel-aligned heatmap first uses two deconvolutional modules to up-sample the feature map from the FisheyeViT. The first deconv module contains one deconv layer with 768 input channels and 1024 output channels, one batch normalization layer, and one ReLU activation function. The deconv layer's kernel size is 4, the stride is 2, the padding is 1, and the output padding is 0. The second deconv module contains one deconv layer with 1024 input channels and 15×64 output channels, one batch normalization layer, and one ReLU activation function. The hyper-parameters of the deconv layer in the second module are the same as that in the first one.

These deconvolutional modules converts the features from shape $(C \times N \times N) = (768 \times 16 \times 16)$ to shape $(J \times D_h \times H_h \times W_h) = (15 \times 64 \times 64 \times 64)$. Then the soft-argmax function and fisheye reprojection function are applied to get the final body pose prediction.

Method	PA-MPJPE	BA-MPJPE
GlobalEgoMocap test dataset (Chapter 4)		
Mo^{2}Cap^2 (Xu et al., 2019)	102.3	74.46
xR-egopose (Tomè et al., 2019)	112.0	87.20
GlobalEgoMocap†	82.06	62.07
EgoPW	81.71	64.87
EgoHMR (Liu et al., 2023b)	85.80	–
SceneEgo	76.50	61.92
Mo^{2}Cap^{2*} (Xu et al., 2019)	78.39	63.48
GlobalEgoMocap*†	75.62	61.06
xR-egopose* (Tomè et al., 2019)	106.3	79.56
EgoPW*	77.95	62.36
SceneEgo*	76.51	61.74
Proposed Method-Single	68.59	55.92
Proposed Method-Refined†	**65.83**	**53.47**

Table D.2: Performance of the proposed method on GlobalEgoMocap test dataset (Chapter 4). The proposed method outperforms all previous state-of-the-art methods. $*$ denotes the method trained with the datasets in Sec. 7.4.1. † denotes the temporal-based methods.

Mo^{2}Cap^2 test dataset (Xu et al., 2019)		
Mo^{2}Cap^2 (Xu et al., 2019)	91.16	70.75
xR-egopose (Tomè et al., 2019)	86.85	66.54
EgoPW	83.17	64.33
Ego-STAN† (Park et al., 2023)	102.4	–
SceneEgo	79.65	62.82
Mo^{2}Cap^{2*} (Xu et al., 2019)	79.76	63.53
xR-egopose* (Tomè et al., 2019)	84.92	65.39
EgoPW*	78.01	62.37
SceneEgo*	79.32	62.77
Proposed Method-Single	74.66	59.26
Proposed Method-Refined†	**72.63**	**57.12**

Table D.3: Performance of the proposed method on Mo^{2}Cap^2 test dataset (Xu et al., 2019). The proposed method outperforms all previous state-of-the-art methods. $*$ denotes the method trained with the datasets in Sec. 7.4.1. † denotes the temporal-based methods.

Method	MPJPE	PA-MPJPE
Mo^{2}Cap^2* (Xu et al., 2019)	89.75	74.32
GlobalEgoMocap*†	86.44	66.76
xR-egopose* (Tomè et al., 2019)	118.2	94.33
EgoPW*	84.21	63.02
SceneEgo*	87.57	69.46
Proposed Method-Single	66.28	43.14
Proposed Method-Refined	**60.32**	**40.35**

Table D.4: Performance of the proposed method on the EgoWholeBody test datasets. This method outperforms all previous state-of-the-art methods. $*$ denotes the method trained with the datasets in Sec. 7.4.1. † denotes the temporal-based methods.

D.2.1.2 *Training Details*

This section introduces the training of the single-frame human body pose estimation network, i. e.the FisheyeViT and pose regressor with pixel-aligned 3D heatmap. The ViT network in FisheyeViT is initialized with the training weight from ViTPose (Xu et al., 2022) and the pose regressor is initialized with normal distribution, whose mean is 0 and standard deviation is 1. The network is trained on the combination dataset of EgoWholeBody and EgoPW. The ratio between the EgoWholeBody and EgoPW datasets is 9:1. The network is trained for 10 epochs with a batch size of 128, a learning rate of $1e^{-4}$ with the Adam optimizer.

D.2.2 *Hand Detection Network*

As described in Sec. 7.2.1.3, the EgoWholeBody dataset is used for training the ViTPose network to regress the heatmap of 2D hand joints. Based on the 2D hand joint predictions, the center $\mathbf{C}_{lh}$, $\mathbf{C}_{rh}$, and the size d_{lh}, d_{rh} of the square hand bounding boxes can be obtained. The ViTPose network is used for simplicity of implementation. Other detection methods can also be used for training the hand detection network. Taking the left hand as an example, the bounding center $\mathbf{C}_{lh}$ is used as the image patch center in **Step 1** of FisheyeViT (Sec. 7.2.1.1) and use the half of the bounding box size $d_{lh}/2$ as the offset d in **Step 2**. After obtaining the projected points of bounding box center $\mathbf{P}^c_{lh}$ and the bounding box edge $\mathbf{P}^x_{lh}$ on the tangent plane $\mathbf{T}_{lh}$, the l in **Step 3** is setted as two times of the Euclidean distance between $\mathbf{P}^x_{lh}$ and $\mathbf{P}^c_{lh}$. Following **Step 4**, the undistorted hand image crop of the left hand $\mathbf{I}_{lh}$ can be finally obtained.

The hand detection network is trained for ten epochs with a batch size of 128 and a learning rate of $1e^{-4}$ with the Adam optimizer.

D.2.3 *Hand Pose Estimation Network*

As described in Sec. 7.2.1.3, the hand-only Pose2Pose network in Hand4Whole method (Moon et al., 2022) is trained with EgoWholeBody training dataset to regress the 3D hand pose from hand image crops. During training, the ground truth 3D hand joint positions are only used as supervision to fine-tune the Pose2Pose network that has been pretrained on the FreiHAND dataset (Zimmermann et al., 2019). The hand pose estimation network is fine-tuned for ten epochs with a batch size of 128 and an initial learning rate of $1e^{-5}$ with the Adam optimizer.

D.2.4 *Diffusion-Based Motion Refinement*

In Sec. 7.2.2, the transformer decoder in EDGE (Tseng et al., 2023) is used as the diffusion denoising network. The music condition in EDGE (Tseng et al., 2023) is disabled by replacing the music features with a learnable feature vector that is agnostic to input. The following describes the training and refinement details of the diffusion model

D.2.4.1 *Training Details*

This section describes the details of training the DDPM model (Ho et al., 2020) for learning the whole-body motion prior. Given a whole-body motion sequence with 196 frames from training datasets (Sec. 7.4.1) represented with joint locations of the human body (with shape 15×3) and hands (with shape 21×3), all poses are transformed to the pelvis-related coordinate system and align them to make the human body poses facing forward, obtaining the aligned whole-body motion sequence $\mathbf{x}$. The motion sequence $\mathbf{x}$ is normalized and sent to the DDPM model for training. During training, a diffusion step $t \in \{0, 1, ..., T-1\}$ is randomly sampled, and the diffusion forward process is applied to generate the noisy motion $\mathbf{x}_t$. Here the T is the maximal diffusion step and T is setted as 1000. Finally, the denoising network is run to obtain the original motion $\hat{\mathbf{x}}$. The reconstructed human motion $\hat{\mathbf{x}}$ and the original human motion $\mathbf{x}_t$ are compared with Eq. 7.4. The network is trained for thirty epochs with a batch size of 256 and an initial learning rate of $2e^{-4}$ with the Adam optimizer.

D.2.4.2 *Refinement Details*

After obtaining the trained diffusion model, Sec. 7.2.2.2 is followed to refine the input whole-body motion. The following describes how to obtain the uncertainty values for each joint in the human body and hands. The 3D heatmap predictions are smoothed using Gaussian smoothing with a standard deviation of 1. The 3D heatmap values **HM** at the predicted joint locations are then obtained using bilinear interpolation. The heatmap values **HM** are firstly normalized to range $[0, 1]$ by making the maximal value of **HM** equal to 1. The uncertainty values **u** is obtained with:

$$\mathbf{u} = 0.05 \times (1 - \mathbf{HM}) \tag{D.1}$$

In this case, the maximal uncertainty value is 0.05. This value is empirically defined to limit the effect of the stochastic diffusion process in motion refinement.

D.3 SYNTHETIC DATASET COMPARISONS

Compared to other egocentric motion capture training datasets, the EgoWholeBody dataset offers several notable advantages (also see Table D.5):
Larger Amount of Frames: EgoWholeBody contains a substantially larger quantity of frames, providing an extensive and diverse dataset for training.
Inclusion of Hand Poses: Unlike other datasets, EgoWholeBody includes hand motion data, making it suitable for egocentric whole-body motion capture.
High Diversity in Motions and Backgrounds: The dataset captures a wide range of human motions and diverse background settings, reflecting real-world scenarios.
Publicly Available Models, Motions, and Backgrounds: The models, motions, and backgrounds are all publicly available. Additionally, the data generation pipeline will be made public, enabling researchers to reproduce or modify the dataset for various different tasks.

These advantages position EgoWholeBody as a valuable resource for advancing research in egocentric whole-body motion capture.

To show the quality of the synthetic dataset, some examples of the synthetic EgoWholeMocap dataset are also visualized in Fig. D.1.

D.4 DETAILS OF EVALUATION METRICS

This section gives a detailed explanation of the evaluation metrics used in this chapter. Mean Per Joint Position Error (MPJPE) is the mean of

Training Dataset	Motion Diversity	Frame Numbers	Motion Type	Image Quality	Annotation Type
EgoPW	low	318 k	body motion	real-world	pseudo ground truth
ECHP (Liu et al., 2023a)	low	75 k	body motion	real-world	pseudo ground truth
Mo^{2}Cap^2 (Xu et al., 2019)	middle	530 k	body motion	low	ground truth
*x*R-EgoPose (Tomè et al., 2019)	middle	380 k	body motion	**realistic**	ground truth
EgoGTA	low	320 k	body motion	low	ground truth
EgoWholeBody	**high**	**870 k**	**body + hands motion**	**realistic**	ground truth

Table D.5: Comparison between different training datasets for egocentric body pose estimation.

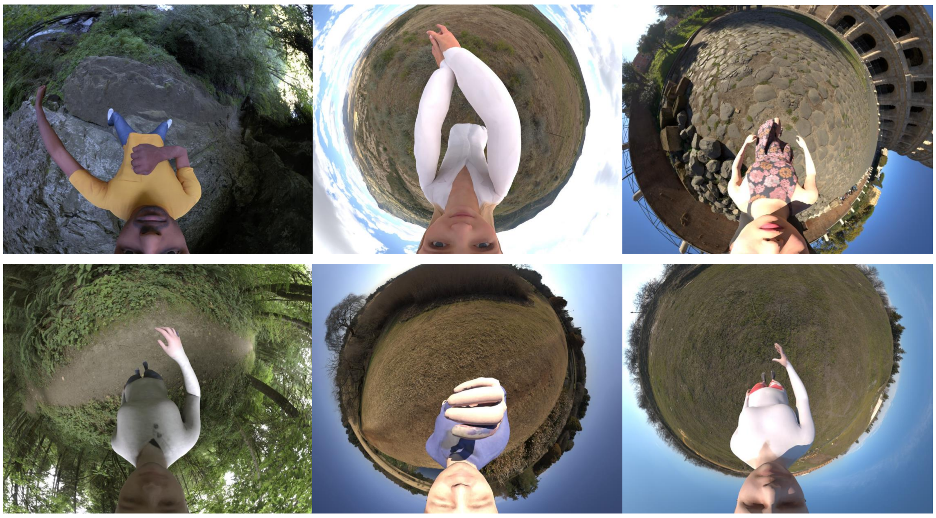


Figure D.1: Examples of the synthetic dataset EgoWholeMocap. The upper row shows the data rendered with Renderpeople models (*RenderPeople* n.d.), and the lower row shows the data rendered with SMPL-X models (Pavlakos et al., 2019).

Euclidean distances for each joint in the predicted and ground truth poses.

For the Mean Per Joint Position Error with Procrustes Analysis (PA-MPJPE), the estimated poses are rigidly aligned to the ground truth poses with Procrustes analysis (Kendall, 1989) and then calculate MPJPE.

The BA-MPJPE, i.e., the MPJPE with aligned bone length, is also evaluated. For BA-MPJPE, the bone lengths of the predicted poses and ground truth poses are first resized to match the bone length of a standard human skeleton. Then, the PA-MPJPE between the two resulting poses is calculated.

D.5 DETAILS OF EVALUATION DATASETS

Experiments in Sec. 7.4.2 use three evaluation datasets including SceneEgo test dataset (Chapter 6), GlobalEgoMocap test dataset (Chapter 4) and Mo^{2}Cap^2 test dataset (Xu et al., 2019). Here these three datasets are introduced in detail.

The SceneEgo test dataset contains around 28K frames of 2 persons performing various motions such as sitting, walking, exercising, reading a newspaper, and using a computer. This dataset provides ground truth egocentric camera pose, allowing for the evaluation of MPJPE. It is evenly divided into training and testing splits. The proposed method is fine-tuned on the training split before evaluation.

The GlobalEgoMocap test dataset contains 12K frames of two people captured in the studio. The Mo^{2}Cap^2 test dataset (Xu et al., 2019) contains 2.7K frames of two people captured in indoor and outdoor scenes. These two datasets do not provide ground truth egocentric camera poses. Therefore, the predicted body poses and ground truth body poses are first rigidly aligned, and then PA-MPJPE and BA-MPJPE are evaluated.

D.6 THE STANDARD DEVIATION OF REFINEMENT METHOD

As described in Sec. 7.4.2, five samples are generated and the mean and standard deviations of the MPJPE values are calculated. The results are shown in Table D.6. The results show that the standard deviations of the results are all around 0.003 mm, which is quite small. The standard deviations of the results are assumed to be small for two reasons:

First, the diffusion process is guided by the low-uncertainty joints. The low-uncertainty joints are more likely to follow the initial motion estimations $\mathbf{x}_e$ and guide the diffusion denoising process of other joints to obtain similar values.

Dataset	MPJPE	PA-MPJPE
SceneEgo-Body	57.59±0.003	46.55±0.003
SceneEgo-Hands	19.37±0.002	9.05±0.002
Dataset	PA-MPJPE	BA-MPJPE
GlobalEgoMocap	65.83±0.003	53.47±0.002
Mo^{2}Cap^2	72.63±0.003	57.12±0.003

Table D.6: The mean and standard deviations of the proposed refinement method. “SceneEgo-Body” and “SceneEgo-Hands” show the body and hand results on the SceneEgo dataset. “GlobalEgoMocap” and “Mo^{2}Cap^2” shows the human body results on the GlobalEgoMocap and Mo^{2}Cap^2 datasets.

Second, according to Eq. D.1, the maximal uncertainty value is 0.05 (the actual uncertainty value can be even smaller), which means that when $k = 0.1$ in Eq. 7.6, the $\mathbf{w} \sim 1$ when $t = 100$ for all joints:

$$\mathbf{w} = 1/\left(1 + e^{-0.1(100-1000\times0.05)}\right) = 0.9933 \tag{D.2}$$

This shows that when t is large enough, the denoising process is always initialized by the estimated motion $\mathbf{x}_e$ and the refinement starts when $t < 100$. When $t < 100$, the Gaussian noise added in Eq. 7.5 is relatively small. This also means that starting from the diffusion step $t = 200$ can accelerate the diffusion refinement process.

D.7 DIFFERENT PARAMETERS IN WEIGHT FUNCTION

This section analyzes the effectiveness of parameter k in the weight function Eq. 7.6. If the uncertainty value of one specific joint is 0.02, the $\mathbf{w}$-t figure is drawn in Fig. D.2. It can be observed that when $t \to 0$, the weight $\mathbf{w}$ is still large when $k = 0.01$. In this case, the initial pose predictions $\mathbf{x}_e$ will significantly affect the final refinement result. When the $k = 1$, the weight $\mathbf{w} \sim 0$ when $t < 15$, which makes the diffusion model generate freely without any guidance of the initial joint estimations. This will make the refined motion largely deviate from the initial joint estimations. In the proposed method, a moderate $k = 0.1$ is chosen so that the diffusion refinement process can be initially guided by the whole-body pose estimations $\mathbf{x}_e$ and finally refined through the generation of diffusion denoising process.

The results under different k values are also shown in Table D.7. The results show that the accuracy of human body poses is the best when $k = 0.1$. It is also observed that the standard deviations become larger when k is larger. This also demonstrates the above analysis.

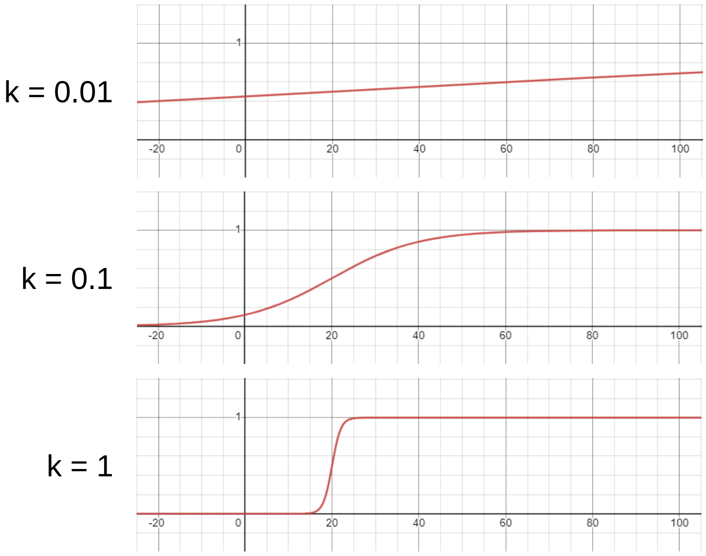


Figure D.2: The weight function with different hyper-parameters k. The x-axis is the diffusion time step t and the y-axis is the weight $\mathbf{w}$.

Method	MPJPE	PA-MPJPE
k=0.01	58.41±0.001	46.92±0.001
k=0.1	**57.59±0.003**	**46.55±0.003**
k=1	59.90±0.006	48.57±0.006

Table D.7: Comparison with different k values.

D.8 MORE VISUALIZATION RESULTS

This section shows more results of the proposed methods in Fig. D.3 and Fig. D.4.

D.9 COMPARISION WITH NETWORKS FOR PANORAMA IMAGES

Recent studies Coors et al., 2018; Li et al., 2022a; Yang et al., 2023a; Yang et al., 2023b; Yu et al., 2023 have adopted various approaches to address fisheye image distortion within deep learning frameworks. Yet, these strategies are tailored to tasks distinctly different from 3D human pose estimation, such as object detection Coors et al., 2018 and depth estimation Li et al., 2022a.

Nevertheless, the FisheyeViT network is compared with two other methods dealing with camera distortions, the SphereNet Coors et al., 2018 and the OmniFusion Li et al., 2022a. In this experiment, the FisheyeViT is

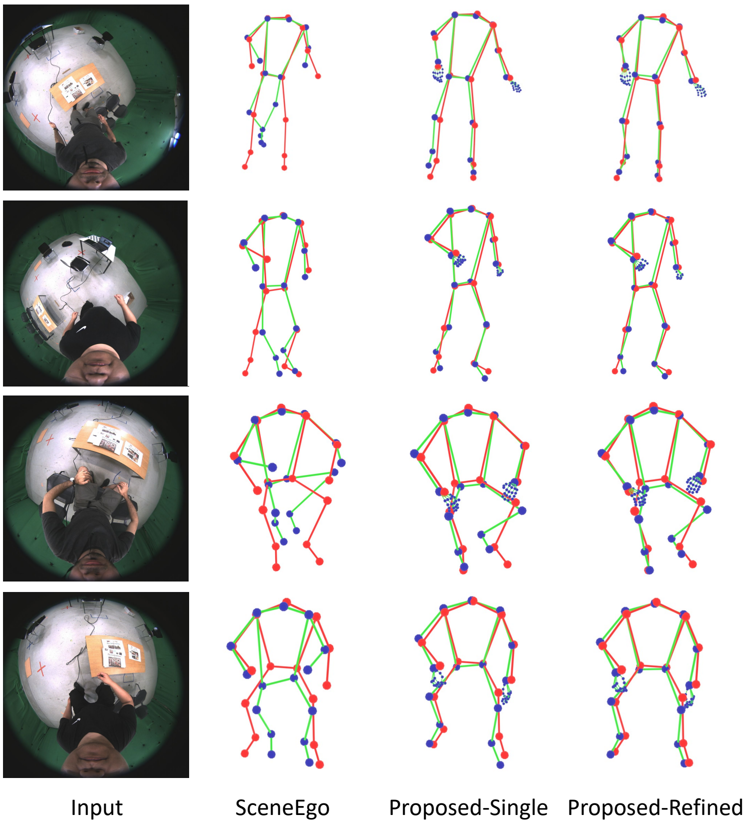


Figure D.3: Qualitative comparison on human body pose estimations between the proposed method and the state-of-the-art SceneEgo (Chapter 6) method. The red skeleton is the ground truth while the green skeleton is the predicted pose. The proposed methods predict more accurate body poses.

replaced with the SphereNet and OmniFusion networks. In SphereNet, the sampling range is limited to the semi-sphere. In OmniFusion, the output of the transformer network is used as the image features, which are then fed into the pose regressor. The accuracy of the estimated human body pose is evaluated on the SceneEgo dataset. The results are shown in Table D.8, which demonstrates that the FisheyeViT performs better than the previous methods for the distorted images. This might caused by the different patch sampling strategies: the proposed method samples the image patches on the fisheye image uv space, while previous methods sample the patches on the $r\theta\phi$ sphere coordinate system. The proposed method can generate patches that align well with the layout of egocentric fisheye images and match the design of the pixel-aligned 3D heatmap as mentioned in the introduction: "the voxels in the 3D heatmap directly correspond to pixels in 2D features, subsequently linking to image patches in FisheyeViT". However, sampling in the $r\theta\phi$ sphere

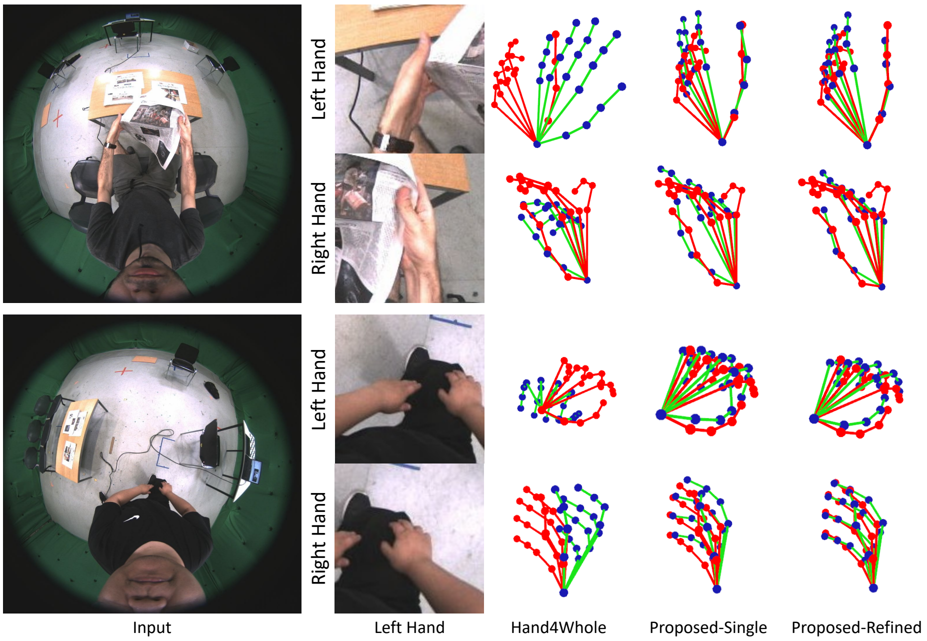


Figure D.4: Qualitative comparison on hand pose estimation results. The proposed single-view and refined hand poses are more accurate than the poses from the Hand4Whole (Moon et al., 2022) method. The red skeleton is the ground truth while the green skeleton is the predicted pose.

Method	MPJPE	PA-MPJPE
SphereNet (Coors et al., 2018)	90.72	75.07
OmniFusion (Li et al., 2022a)	86.58	70.69
Ours-Single	**64.19**	**50.06**

Table D.8: Comparison with Spherenet and Panoformer.

coordinate system will cause discontinuity due to the *coordinate singularity* of the sphere coordinate system. For example, the neighboring pixels on the fisheye image can be assigned to two patches far away from each other.

D.10 REPLACING THE PIXEL-ALIGNED 3D HEATMAP TO MLP

In this section, the pose regressor is replaced with the pixel-aligned 3D heatmap with a simple MLP network. The features extracted with FisheyeViT, with shape $(768 \times 16 \times 16)$ are firstly flattened and two MLP layers are further adopted to regress the 3D human body poses. The first layer contains one fully connected layer with an output dimension of

1024, one batch normalization layer, and one ReLU activation layer. The second layer contains one fully connected layer with an output dimension of 15×3. The MPJPE and the PA-MPJPE on the SceneEgo dataset are 130.7 mm and 73.91 mm respectively. This demonstrates the effectiveness of the egocentric pose regressor with pixel-aligned 3D heatmap.

D.11 COMPARE WITH GAUSSIAN SMOOTH

In this section, the diffusion-based motion refinement method is compared with simple Gaussian smoothing. The MPJPE and the PA-MPJPE on the SceneEgo dataset are 62.68 mm and 48.87 mm respectively. This demonstrates that the refinement method performs better than the Gaussian smooth approach. This shows that the proposed method relies on motion priors to guide the refinement of human motion, making it more effective than the simple smoothing techniques.

BIBLIOGRAPHY

Ahuja, Karan, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D Wilson (2021). "Coolmoves: User motion accentuation in virtual reality." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.2, pp. 1–23.

Akada, Hiroyasu, Jian Wang, Vladislav Golyanik, and Christian Theobalt (2024). "3D Human Pose Perception from Egocentric Stereo Videos." In: *Computer Vision and Pattern Recognition (CVPR)*.

Akada, Hiroyasu, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik (2022). "UnrealEgo: A new dataset for robust egocentric 3d human motion capture." In: *European Conference on Computer Vision*. Springer, pp. 1–17.

Armani, Rayan, Changlin Qian, Jiaxi Jiang, and Christian Holz (2024). "Ultra Inertial Poser: Scalable Motion Capture and Tracking from Sparse Inertial Sensors and Ultra-Wideband Ranging." In: *arXiv preprint arXiv:2404.19541*.

Arnab, Anurag, Carl Doersch, and Andrew Zisserman (2019). "Exploiting temporal context for 3D human pose estimation in the wild." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3395–3404.

Baradel, Fabien, Romain Brégier, Thibault Groueix, Philippe Weinzaepfel, Yannis Kalantidis, and Grégory Rogez (2022). "Posebert: A generic transformer module for temporal 3d human modeling." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.11, pp. 12798–12815.

Bhatnagar, Bharat Lal, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll (2022). "Behave: Dataset and method for tracking human object interactions." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15935–15946.

Black, Michael J, Priyanka Patel, Joachim Tesch, and Jinlong Yang (2023). "Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8726–8737.

Blender (n.d.). `http://www.blender.org`.

Bogo, Federica, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black (2016). "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image." In: *Computer*

Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer, pp. 561–578.

CMU mocap dataset (2008). http://mocap.cs.cmu.edu/.

Cai, Yujun, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann (2019). "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks." In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2272–2281.

Cai, Zhongang, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. (2023). "SMPLer-X: Scaling Up Expressive Human Pose and Shape Estimation." In: *arXiv preprint arXiv:2309.17448*.

Cao, Zhe, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik (2020). "Long-term human motion prediction with scene context." In: *European Conference on Computer Vision (ECCV)*.

Cao, Zhe, Tomas Simon, Shih-En Wei, and Yaser Sheikh (2017). "Realtime multi-person 2d pose estimation using part affinity fields." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299.

Cha, Young-Woon, True Price, Zhen Wei, Xinran Lu, Nicholas Rewkowski, Rohan Chabra, Zihe Qin, Hyounghun Kim, Zhaoqi Su, Yebin Liu, Adrian Ilie, Andrei State, Zhenlin Xu, Jan-Michael Frahm, and Henry Fuchs (2018). "Towards Fully Mobile 3D Face, Body, and Environment Capture Using Only Head-worn Cameras." In: *IEEE Transactions on Visualization and Computer Graphics* 24.11, pp. 2993–3004.

Chen, Ching-Hang and Deva Ramanan (2017). "3d human pose estimation= 2d pose estimation+ matching." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7035–7043.

Chen, Ching-Hang, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg (2019a). "Unsupervised 3d pose estimation with geometric self-supervision." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5714–5724.

Chen, Xipeng, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin (2019b). "Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10895–10904.

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

Cho, Junhyeong, Kim Youwang, and Tae-Hyun Oh (2022). "Cross-attention of disentangled modalities for 3d human mesh recovery with transformers." In: *European Conference on Computer Vision*. Springer, pp. 342–359.

Choi, Hongsuk, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee (2021). "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*, pp. 1964–1973.

Choi, Jeongjun, Dongseok Shim, and H Jin Kim (2022). "Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model." In: *arXiv preprint arXiv:2212.02796*.

Choutas, Vasileios, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black (2020). "Monocular expressive body regression through body-driven attention." In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, pp. 20–40.

Chu, Xiao, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang (2017). "Multi-context attention for human pose estimation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1831–1840.

Ci, Hai, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang (2023). "Gfpose: Learning 3d human pose prior with gradient fields." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4800–4810.

Coors, Benjamin, Alexandru Paul Condurache, and Andreas Geiger (2018). "Spherenet: Learning spherical representations for detection and classification in omnidirectional images." In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 518–533.

Curless, Brian and Marc Levoy (1996). "A volumetric method for building complex models from range images." In: *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 303–312.

Dabral, Rishabh, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt (2023). "Mofusion: A framework for denoising-diffusion-based motion synthesis." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9760–9770.

Dabral, Rishabh, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik (2021). "Gravity-aware monocular 3d human-object reconstruction." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12365–12374.

Dai, Peng, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, and Zeming Li (2024). "HMD-Poser: On-Device Real-time Human Motion Tracking from Scalable Sparse Observations." In: *arXiv preprint arXiv:2403.03561*.

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

Dhariwal, Prafulla and Alexander Nichol (2021). "Diffusion models beat gans on image synthesis." In: *Advances in neural information processing systems* 34, pp. 8780–8794.

Ding, Runyu, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang (2024). "Bunny-VisionPro: Real-Time Bimanual Dexterous Teleoperation for Imitation Learning." In: *arXiv preprint arXiv:2407.03162*.

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

Doersch, Carl and Andrew Zisserman (2019). "Sim2real transfer learning for 3d human pose estimation: motion to the rescue." In: *Advances in Neural Information Processing Systems* 32.

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

Drover, Dylan, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh (2018). "Can 3d pose be learned from 2d projections alone?" In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0.

Du, Yu, Yongkang Wong, Yonghao Liu, Feilin Han, Yilin Gui, Zhen Wang, Mohan Kankanhalli, and Weidong Geng (2016). "Marker-less 3D human motion capture with monocular image sequence and height-maps." In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, pp. 20–36.

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

Elgharib, Mohamed, Mallikarjun BR, Ayush Tewari, Hyeongwoo Kim, Wentao Liu, Hans-Peter Seidel, and Christian Theobalt (2019). *EgoFace: Egocentric Face Performance Capture and Videorealistic Reenactment*. arXiv: 1905.10822 [cs.CV].

Elgharib, Mohamed, Mohit Mendiratta, Justus Thies, Matthias Nießner, Hans-Peter Seidel, Ayush Tewari, Vladislav Golyanik, and Christian Theobalt (Dec. 2020). "Egocentric Videoconferencing." In: *ACM Transactions on Graphics* 39.6.

Feng, Yao, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black (2021). "Collaborative regression of expressive bodies using moderation." In: *2021 International Conference on 3D Vision (3DV)*. IEEE, pp. 792–804.

Foo, Lin Geng, Jia Gong, Hossein Rahmani, and Jun Liu (2023). "Distribution-aligned diffusion for human mesh recovery." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9221–9232.

Fu, Huan, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao (2018). "Deep ordinal regression network for monocular depth estimation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2002–2011.

Gao, X., Xiaorong Hou, Jianliang Tang, and H. Cheng (2003). "Complete Solution Classification for the Perspective-Three-Point Problem." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 25, pp. 930–943.

Georgakis, Georgios, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu (2020). "Hierarchical kinematic human mesh recovery." In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, pp. 768–784.

Gong, Jia, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu (2023). "Diffpose: Toward more reliable 3d pose estimation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13041–13051.

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

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2020). "Generative adversarial networks." In: *Communications of the ACM* 63.11, pp. 139–144.

Grauman, Kristen, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. (2022). "Ego4d: Around the world in 3,000 hours of egocentric video." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012.

Grauman, Kristen, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. (2024). "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400.

Guan, Peng, Alexander Weiss, Alexandru O Balan, and Michael J Black (2009). "Estimating human shape and pose from a single image." In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE, pp. 1381–1388.

Guler, Riza Alp and Iasonas Kokkinos (2019). "Holopose: Holistic 3d human reconstruction in-the-wild." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10884–10894.

Guzov, Vladimir, Julian Chibane, Riccardo Marin, Yannan He, Torsten Sattler, and Gerard Pons-Moll (2022). "Interaction Replica: Tracking human-object interaction and scene changes from human motion." In: *arXiv preprint arXiv:2205.02830*.

Guzov, Vladimir, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll (2021). "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*, pp. 4318–4329.

Hartley, Richard and Andrew Zisserman (2003). *Multiple view geometry in computer vision*. Cambridge university press.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, Tairan, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi (2024a). "Learning human-to-humanoid real-time whole-body teleoperation." In: *arXiv preprint arXiv:2403.04436*.

He, Yannan, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll (June 2024b). "NRDF: Neural Riemannian Distance Fields for Learning Articulated Pose Priors." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models." In: *Advances in neural information processing systems* 33, pp. 6840–6851.

Holmquist, Karl and Bastian Wandt (2023). "Diffpose: Multi-hypothesis human pose estimation using diffusion models." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15977–15987.

Hossain, Mir Rayat Imtiaz and James J Little (2018). "Exploiting temporal information for 3d human pose estimation." In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 68–84.

Hoyet, Ludovic, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan (2012). "Sleight of hand: perception of finger motion from reduced marker sets." In: *Proceedings of the ACM SIGGRAPH symposium on interactive 3D graphics and games*, pp. 79–86.

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

Hua, Guoliang, Hong Liu, Wenhao Li, Qian Zhang, Runwei Ding, and Xin Xu (2022). "Weakly-supervised 3D human pose estimation with cross-view U-shaped graph convolutional network." In: *IEEE Transactions on Multimedia*.

Huang, Yinghao, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll (2018). "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time." In: *ACM Transactions on Graphics (TOG)* 37.6, pp. 1–15.

Hwang, Dong-Hyun, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike (2020). "Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera." In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 98–111.

Ionescu, Catalin, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu (2013). "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." In: *IEEE transactions on pattern analysis and machine intelligence* 36.7, pp. 1325–1339.

Iqbal, Umar, Pavlo Molchanov, and Jan Kautz (2020). "Weakly-supervised 3d human pose learning via multi-view images in the wild." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5243–5252.

Iskakov, Karim, Egor Burkov, Victor Lempitsky, and Yury Malkov (2019). "Learnable Triangulation of Human Pose." In: *International Conference on Computer Vision (ICCV)*.

Jahangiri, Ehsan and Alan L Yuille (2017). "Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections." In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 805–814.

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

Jiang, Jiaxi, Paul Streli, Manuel Meier, Andreas Fender, and Christian Holz (2023a). “EgoPoser: Robust Real-Time Ego-Body Pose Estimation in Large Scenes.” In: *arXiv preprint arXiv:2308.06493*.

Jiang, Jiaxi, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz (2022a). “Avatarposer: Articulated full-body pose tracking from sparse motion sensing.” In: *European Conference on Computer Vision*. Springer, pp. 443–460.

Jiang, Yifeng, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu (2022b). “Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation.” In: *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9.

Jiang, Zhongyu, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, and Jenq-Neng Hwang (2023b). “Back to Optimization: Diffusion-based Zero-Shot 3D Human Pose Estimation.” In: *arXiv preprint arXiv:2307.03833*.

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

Kanazawa, Angjoo, Michael J Black, David W Jacobs, and Jitendra Malik (2018). “End-to-end recovery of human shape and pose.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7122–7131.

Kanazawa, Angjoo, Jason Y Zhang, Panna Felsen, and Jitendra Malik (2019). “Learning 3d human dynamics from video.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5614–5623.

Kang, Taeho, Kyungjin Lee, Jinrui Zhang, and Youngki Lee (2023). “Ego3dpose: Capturing 3d cues from binocular egocentric views.” In: *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–10.

Kannala, Juho and Sami S Brandt (2006). “A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses.” In: *IEEE transactions on pattern analysis and machine intelligence* 28.8, pp. 1335–1340.

Katircioglu, Isinsu, Bugra Tekin, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua (2018). “Learning latent representations of 3d human pose with deep neural networks.” In: *International Journal of Computer Vision* 126, pp. 1326–1341.

Kendall, David G (1989). “A survey of the statistical theory of shape.” In: *Statistical Science* 4.2, pp. 87–99.

Khirodkar, Rawal, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani (2023a). "Ego-Humans: An Ego-Centric 3D Multi-Human Benchmark." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19807–19819.

– (2023b). "EgoHumans: An Egocentric 3D Multi-Human Benchmark." In: *arXiv preprint arXiv:2305.16487*.

Kim, Meejin and Sukwon Lee (2022). "Fusion poser: 3D human pose estimation using sparse IMUs and head trackers in real time." In: *Sensors* 22.13, p. 4846.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization." In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes." In: *arXiv preprint arXiv:1312.6114*.

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

Kocabas, Muhammed, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black (2021). "PARE: Part attention regressor for 3D human body estimation." In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11127–11137.

Kocabas, Muhammed, Salih Karagoz, and Emre Akbas (2019). "Self-supervised learning of 3d human pose using multi-view geometry." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1077–1086.

Kolotouros, Nikos, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis (2019a). "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*, pp. 2252–2261.

Kolotouros, Nikos, Georgios Pavlakos, and Kostas Daniilidis (2019b). "Convolutional mesh regression for single-image human shape reconstruction." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4501–4510.

Kolotouros, Nikos, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis (2021). "Probabilistic modeling for human mesh recovery." In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11605–11614.

Kundu, Jogendra Nath, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty (2020). "Self-supervised 3d human pose estimation via part guided novel image synthesis." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6152–6162.

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

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

Lee, Jiye and Hanbyul Joo (2024). "Mocap Everyone Everywhere: Lightweight Motion Capture With Smartwatches and a Head-Mounted Camera." In: *CVPR*.

Lee, Sunmin, Sebastian Starke, Yuting Ye, Jungdam Won, and Alexander Winkler (2023). "Questenvsim: Environment-aware simulated motion tracking from sparse sensors." In: *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–9.

Li, Gen, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang (2024). "EgoGen: An Egocentric Synthetic Data Generator." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14497–14509.

Li, Hao, Laura C. Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma (2015). "Facial performance sensing head-mounted display." In: *ACM Trans. Graph.* 34.4, 47:1–47:9. DOI: 10.1145/2766939. URL: https://doi.org/10.1145/2766939.

Li, Jiaman, Karen Liu, and Jiajun Wu (2023). "Ego-Body Pose Estimation via Ego-Head Pose Estimation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17142–17151.

Li, Sijin and Antoni B Chan (2015). "3d human pose estimation from monocular images with deep convolutional neural network." In: *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12*. Springer, pp. 332–347.

Li, Yuyan, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren (2022a). "Omnifusion: 360 monocular depth estimation via geometry-aware fusion." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2801–2810.

Li, Zhi, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik (2022b). "Mocapdeform: Monocular 3d human motion capture in deformable scenes." In: *2022 International Conference on 3D Vision (3DV)*. IEEE, pp. 1–11.

Liang, Han, Yannan He, Chengfeng Zhao, Mutian Li, Jingya Wang, Jingyi Yu, and Lan Xu (2023). "Hybridcap: Inertia-aid monocular capture of challenging human motions." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 2, pp. 1539–1548.

Lin, Jing, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang (2024). "Motion-x: A large-scale 3d expressive whole-body human motion dataset." In: *Advances in Neural Information Processing Systems* 36.

Lin, Jing, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li (2023). "One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21159–21168.

Lin, Kevin, Lijuan Wang, and Zicheng Liu (2021a). "End-to-end human pose and mesh reconstruction with transformers." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1954–1963.

– (2021b). "Mesh graphormer." In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12939–12948.

Lin, Mude, Liang Lin, Xiaodan Liang, Keze Wang, and Hui Cheng (2017). "Recurrent 3d pose sequence machines." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 810–819.

Liu, Yebin, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt (2011). "Markerless motion capture of interacting characters using multi-view image segmentation." In: *CVPR 2011*. Ieee, pp. 1249–1256.

Liu, Yuxuan, Jianxin Yang, Xiao Gu, Yijun Chen, Yao Guo, and Guang-Zhong Yang (2023a). "EgoFish3D: Egocentric 3D Pose Estimation from a Fisheye Camera via Self-Supervised Learning." In: *IEEE Transactions on Multimedia*.

Liu, Yuxuan, Jianxin Yang, Xiao Gu, Yao Guo, and Guang-Zhong Yang (2022). "Ego+ X: An Egocentric Vision System for Global 3D Human Pose Estimation and Social Interaction Characterization." In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 5271–5277.

– (2023b). "EgoHMR: Egocentric Human Mesh Recovery via Hierarchical Latent Diffusion Model." In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 9807–9813.

Loper, Matthew, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black (Oct. 2015). "SMPL: A Skinned Multi-Person Linear Model." In: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6, 248:1–248:16.

Luo, Zhengyi, Jinkun Cao, Rawal Khirodkar, Alexander Winkler, Kris Kitani, and Weipeng Xu (2024). "Real-Time Simulated Avatar from Head-Mounted Sensors." In: *arXiv preprint arXiv:2403.06862*.

Luo, Zhengyi, S Alireza Golestaneh, and Kris M Kitani (2020). "3d human motion estimation via motion compression and refinement." In: *Proceedings of the Asian Conference on Computer Vision*.

Luo, Zhengyi, Ryo Hachiuma, Ye Yuan, and Kris Kitani (2021). "Dynamics-regulated kinematic policy for egocentric pose estimation." In: *Advances in Neural Information Processing Systems* 34, pp. 25019–25032.

Ma, Lingni, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. (2024). "Nymeria: A Massive Collection of Multimodal Egocentric Daily Motion in the Wild." In: *arXiv preprint arXiv:2406.09905*.

Ma, Minghuang, Haoqi Fan, and Kris M. Kitani (2016). "Going Deeper into First-Person Activity Recognition." In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1894–1903. DOI: `10.1109/CVPR.2016.209`. URL: `https://doi.org/10.1109/CVPR.2016.209`.

Mahmood, Naureen, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black (2019). "AMASS: Archive of motion capture as surface shapes." In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451.

Martinez, Julieta, Rayat Hossain, Javier Romero, and James J Little (2017). "A simple yet effective baseline for 3d human pose estimation." In: *Proceedings of the IEEE international conference on computer vision*, pp. 2640–2649.

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

Mehta, Dushyant, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt (2020). "XNect: Real-time multi-person 3D motion capture with a single RGB camera." In: *Acm Transactions On Graphics (TOG)* 39.4, pp. 82–1.

Mehta, Dushyant, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt (2017b). "Vnect: Real-time 3d human pose estimation with a single rgb camera." In: *Acm transactions on graphics (tog)* 36.4, pp. 1–14.

Millerdurai, Christen, Hiroyasu Akada, Jian Wang, Diogo Luvizon, Christian Theobalt, and Vladislav Golyanik (2024). "EventEgo3D: 3D Human Motion Capture from Egocentric Event Streams." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mixamo (n.d.). https://www.mixamo.com.

Mollyn, Vimal, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja (2023). "IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–12.

Moon, Gyeongsik, Ju Yong Chang, and Kyoung Mu Lee (2018). "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map." In: *CVPR*, pp. 5079–5088.

Moon, Gyeongsik, Hongsuk Choi, and Kyoung Mu Lee (2022). "Accurate 3D Hand Pose Estimation for Whole-Body 3D Human Mesh Estimation." In: *Computer Vision and Pattern Recognition Workshop (CVPRW)*.

Mur-Artal, Raul and Juan D Tardós (2017). "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras." In: *IEEE transactions on robotics* 33.5, pp. 1255–1262.

Nathan Silberman Derek Hoiem, Pushmeet Kohli and Rob Fergus (2012). "Indoor Segmentation and Support Inference from RGBD Images." In: *ECCV*.

Ng, Evonne, Donglai Xiang, Hanbyul Joo, and Kristen Grauman (2020). "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*, pp. 9890–9900.

Nocedal, Jorge and Stephen Wright (2006). *Numerical optimization*.

Novotny, David, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi (2019). "C3dpo: Canonical 3d pose networks for non-rigid structure from motion." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7688–7697.

Papamakarios, George, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan (2021). "Normalizing flows for probabilistic modeling and inference." In: *Journal of Machine Learning Research* 22.57, pp. 1–64.

Park, Jinman, Kimathi Kaai, Saad Hossain, Norikatsu Sumi, Sirisha Rambhatla, and Paul Fieguth (2023). "Domain-Guided Spatio-Temporal Self-Attention for Egocentric 3D Pose Estimation." In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1837–1849.

Park, Sang Il and Jessica K Hodgins (2006). "Capturing and animating skin deformation in human motion." In: *ACM Transactions on Graphics (TOG)* 25.3, pp. 881–889.

Pavlakos, Georgios, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black (2019). "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*, pp. 10975–10985.

Pavlakos, Georgios, Jitendra Malik, and Angjoo Kanazawa (2022a). "Human mesh recovery from multiple shots." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1485–1495.

Pavlakos, Georgios, Ethan Weber, Matthew Tancik, and Angjoo Kanazawa (2022b). "The one where they reconstructed 3d humans and environments in tv shows." In: *European Conference on Computer Vision*. Springer, pp. 732–749.

Pavllo, Dario, Christoph Feichtenhofer, David Grangier, and Michael Auli (2019). "3d human pose estimation in video with temporal convolutions and semi-supervised training." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7753–7762.

Pons-Moll, Gerard, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn (2010). "Multisensor-fusion for 3d full-body human motion capture." In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 663–670.

Popa, Alin-Ionut, Mihai Zanfir, and Cristian Sminchisescu (2017). "Deep multitask architecture for integrated 2d and 3d human sensing." In: *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6289–6298.

Project Aria (n.d.). https://www.projectaria.com/.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen (2022). "Hierarchical text-conditional image generation with clip latents." In: *arXiv preprint arXiv:2204.06125* 1.2, p. 3.

Rempe, Davis, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas (2021). "Humor: 3d human motion model for robust pose estimation." In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11488–11499.

RenderPeople (n.d.). https://renderpeople.com.

Reynolds, Douglas A et al. (2009). "Gaussian mixture models." In: *Encyclopedia of biometrics* 741.659-663.

Rhodin, Helge, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt (2016). "Egocap: egocentric marker-less motion capture with two fisheye cameras." In: *ACM Transactions on Graphics (TOG)* 35.6, pp. 1–11.

Rhodin, Helge, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua (2018). "Learning monocular 3d human pose estimation from multi-view images."

In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8437–8446.

Roetenberg, Daniel, Henk Luinge, and Per J. Slycke (2009). "Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors." In.

Rogez, Gregory, Philippe Weinzaepfel, and Cordelia Schmid (2017). "Lcr-net: Localization-classification-regression for human pose." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3433–3441.

Rong, Yu, Takaaki Shiratori, and Hanbyul Joo (2021). "Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1749–1759.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation." In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.

Saito, Shunsuke, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam (2024). "Relightable gaussian codec avatars." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 130–141.

Scaramuzza, Davide and Katsushi Ikeuchi (2014). "Omnidirectional camera." In.

Scaramuzza, Davide, Agostino Martinelli, and Roland Siegwart (2006). "A toolbox for easily calibrating omnidirectional cameras." In: *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 5695–5701.

Schwarz, Loren Arthur, Diana Mateus, and Nassir Navab (2009). "Discriminative human full-body pose estimation from wearable inertial sensor data." In: *Modelling the Physiological Human: 3D Physiological Human Workshop, 3DPH 2009, Zermatt, Switzerland, November 29–December 2, 2009. Proceedings*. Springer, pp. 159–172.

Shan, Wenkang, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao (2023). "Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation." In: *arXiv preprint arXiv:2303.11579*.

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

Shimada, Soshi, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt (2021). "Neural monocular 3d human motion capture with physical awareness." In: *ACM Transactions on Graphics (ToG)* 40.4, pp. 1–15.

Shimada, Soshi, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt (2020). "Physcap: Physically plausible monocular 3d motion capture in real time." In: *ACM Transactions on Graphics (ToG)* 39.6, pp. 1–16.

Shiratori, Takaaki, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K Hodgins (2011). "Motion capture from body-mounted cameras." In: *ACM SIGGRAPH 2011 papers*, pp. 1–10.

Singh, S., C. Arora, and C. V. Jawahar (2016). "First Person Action Recognition Using Deep Learned Descriptors." In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2620–2628.

Singh, Suriya, Chetan Arora, and C. V. Jawahar (2017). "Trajectory aligned features for first person action recognition." In: *Pattern Recognit.* 62, pp. 45–55. DOI: 10.1016/j.patcog.2016.07.031. URL: https://doi.org/10.1016/j.patcog.2016.07.031.

Song, Jiaming, Chenlin Meng, and Stefano Ermon (2020). "Denoising diffusion implicit models." In: *arXiv preprint arXiv:2010.02502*.

Sridhar, Srinath, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt (June 2015). "Fast and Robust Hand Tracking Using Detection-Guided Optimization." In: *IEEE Conference on Computer Vision and Pattern Recognition*. URL: http://handtracker.mpi-inf.mpg.de/projects/FastHandTracker/.

Sun, Xiao, Jiaxiang Shang, Shuang Liang, and Yichen Wei (2017). "Compositional human pose regression." In: *Proceedings of the IEEE international conference on computer vision*, pp. 2602–2611.

Sun, Xiao, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei (2018). "Integral human pose regression." In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 529–545.

Sun, Yu, Tianyu Huang, Qian Bao, Wu Liu, Wenpeng Gao, and Yili Fu (2022). "Learning monocular mesh recovery of multiple body parts via synthesis." In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2669–2673.

Taheri, Omid, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas (2020). "GRAB: A dataset of whole-body human grasping of objects." In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, pp. 581–600.

Teed, Zachary and Jia Deng (2021). "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras." In: *Advances in neural information processing systems* 34, pp. 16558–16569.

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

Tekin, Bugra, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua (2017). "Learning to fuse 2d and 3d image cues for monocular body pose estimation." In: *Proceedings of the IEEE international conference on computer vision*, pp. 3941–3950.

Tevet, Guy, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano (2022). "Human motion diffusion model." In: *arXiv preprint arXiv:2209.14916*.

Tiwari, Garvita, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll (2022). "Pose-ndf: Modeling human pose manifolds with neural distance fields." In: *European Conference on Computer Vision*. Springer, pp. 572–589.

Tomè, Denis, Patrick Peluse, Lourdes Agapito, and Hernán Badino (2019). "xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera." In: *ICCV*, pp. 7727–7737. DOI: 10.1109/ICCV.2019.00782. URL: https://doi.org/10.1109/ICCV.2019.00782.

Tome, Denis, Chris Russell, and Lourdes Agapito (2017). "Lifting from the deep: Convolutional 3d pose estimation from a single image." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2500–2509.

Trumble, Matthew, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John P Collomosse (2017). "Total capture: 3D human pose estimation fusing video and inertial sensors." In: *BMVC*. Vol. 2. 5. London, UK, pp. 1–13.

Tsai, Roger Y and Reimar K Lenz (1988). "Real time versatile robotics hand/eye calibration using 3D machine vision." In: *Proceedings. 1988 IEEE International Conference on Robotics and Automation*. IEEE, pp. 554–561.

Tseng, Jonathan, Rodrigo Castellon, and Karen Liu (2023). "Edge: Editable dance generation from music." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–458.

Tzeng, Eric, Judy Hoffman, Kate Saenko, and Trevor Darrell (2017). "Adversarial discriminative domain adaptation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176.

Van Wouwe, Tom, Seunghwan Lee, Antoine Falisse, Scott Delp, and C Karen Liu (2024). "DiffusionPoser: Real-time Human Motion Reconstruction From Arbitrary Sparse Sensors Using Autoregressive Diffusion." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2513–2523.

Vlasic, Daniel, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović (2007). "Practical motion capture in everyday surroundings." In: *ACM transactions on graphics (TOG)* 26.3, 35–es.

Von Marcard, Timo, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll (2018). “Recovering accurate 3d human pose in the wild using imus and a moving camera.” In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 601–617.

Von Marcard, Timo, Gerard Pons-Moll, and Bodo Rosenhahn (2016). “Human pose estimation from video and imus.” In: *IEEE transactions on pattern analysis and machine intelligence* 38.8, pp. 1533–1547.

Von Marcard, Timo, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll (2017). “Sparse inertial poser: Automatic 3d human pose estimation from sparse imus.” In: *Computer graphics forum*. Vol. 36. 2. Wiley Online Library, pp. 349–360.

Wan, Ziniu, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li (2021). “Encoder-decoder with multi-level attention for 3d human shape and pose estimation.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13033–13042.

Wandt, Bastian and Bodo Rosenhahn (2019). “Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7782–7791.

Wandt, Bastian, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn (2021). “CanonPose: Self-supervised monocular 3D human pose estimation in the wild.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304.

Wang, Chaoyang, Chen Kong, and Simon Lucey (2019). “Distill knowledge from nrsfm for weakly supervised 3d pose learning.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 743–752.

Wang, Jian, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt (2024). “Egocentric Whole-Body Motion Capture with FisheyeViT and Diffusion-Based Motion Refinement.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, Jian, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt (2022). “Estimating egocentric 3d human pose in the wild with external weak supervision.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13157–13166.

Wang, Jian, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt (2021). “Estimating egocentric 3d human pose in global space.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11500–11509.

Wang, Jian, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt (2023). "Scene-aware Egocentric 3D Human Pose Estimation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13031–13040.

Wang, Jingbo, Sijie Yan, Yuanjun Xiong, and Dahua Lin (2020a). "Motion guided 3d pose estimation from videos." In: *European conference on computer vision*. Springer, pp. 764–780.

Wang, Jingdong, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. (2020b). "Deep high-resolution representation learning for visual recognition." In: *IEEE transactions on pattern analysis and machine intelligence* 43.10, pp. 3349–3364.

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

Winkler, Alexander, Jungdam Won, and Yuting Ye (2022). "Questsim: Human motion tracking from sparse sensors with simulated avatars." In: *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–8.

Xiang, Donglai, Hanbyul Joo, and Yaser Sheikh (2019). "Monocular total capture: Posing face, body, and hands in the wild." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10974.

Xiao, Bin, Haiping Wu, and Yichen Wei (2018). "Simple Baselines for Human Pose Estimation and Tracking." In: *European Conference on Computer Vision (ECCV)*.

Xu, Weipeng, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt (2019). "Mo^{2}Cap^2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera." In: *IEEE Trans. Vis. Comput. Graph.* 25.5, pp. 2093–2101. DOI: 10.1109/TVCG.2019.2898650. URL: https://doi.org/10.1109/TVCG.2019.2898650.

Xu, Yufei, Jing Zhang, Qiming Zhang, and Dacheng Tao (2022). "Vitpose: Simple vision transformer baselines for human pose estimation." In: *Advances in Neural Information Processing Systems* 35, pp. 38571–38584.

Yang, Chenhongyi, Anastasia Tkach, Shreyas Hampali, Linguang Zhang, Elliot J Crowley, and Cem Keskin (2024a). "EgoPoseFormer: A Simple Baseline for Egocentric 3D Human Pose Estimation." In: *arXiv preprint arXiv:2403.18080*.

Yang, Dianyi, Jiadong Tang, Yu Gao, Yi Yang, and Mengyin Fu (2023a). "Sector Patch Embedding: An Embedding Module Conforming to The Distortion Pattern of Fisheye Image." In: *arXiv preprint arXiv:2303.14645*.

Yang, Dongseok, Jiho Kang, Lingni Ma, Joseph Greer, Yuting Ye, and Sung-Hee Lee (2024b). "DivaTrack: Diverse Bodies and Motions from Acceleration-Enhanced Three-Point Trackers." In: *Computer Graphics Forum*. Wiley Online Library, e15057.

Yang, Dongseok, Doyeon Kim, and Sung-Hee Lee (2021). "Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals." In: *Computer Graphics Forum*. Vol. 40. 2. Wiley Online Library, pp. 265–275.

Yang, Shangrong, Chunyu Lin, Kang Liao, and Yao Zhao (2023b). "Dual diffusion architecture for fisheye image rectification: Synthetic-to-real generalization." In: *arXiv preprint arXiv:2301.11785*.

Yang, Wei, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang (2018). "3d human pose estimation in the wild by adversarial learning." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5255–5264.

Yi, Hongwei, Chun-Hao P Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J Black (2022a). "Human-aware object placement for visual environment reconstruction." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3959–3970.

Yi, Xinyu, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu (2023). "EgoLocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors." In: *ACM Transactions on Graphics (TOG)* 42.4, pp. 1–17.

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

Yi, Xinyu, Yuxiao Zhou, and Feng Xu (2021). "Transpose: Real-time 3d human translation and pose estimation with six inertial sensors." In: *ACM Transactions on Graphics (TOG)* 40.4, pp. 1–13.

Yu, Fanghua, Xintao Wang, Mingdeng Cao, Gen Li, Ying Shan, and Chao Dong (2023). "Osrt: Omnidirectional image super-resolution with distortion-aware transformer." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13283–13292.

Yu, Ri, Hwangpil Park, and Jehee Lee (2021). "Human dynamics from monocular video with dynamic camera movements." In: *ACM Transactions on Graphics (TOG)* 40.6, pp. 1–14.

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

– (2019). “Ego-pose estimation and forecasting as real-time pd control.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10082–10092.

Yuan, Ye, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz (2023). “Physdiff: Physics-guided human motion diffusion model.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16010–16021.

Yuan, Yuhui, Xilin Chen, and Jingdong Wang (2020a). “Object-Contextual Representations for Semantic Segmentation.” In.

– (2020b). “Object-contextual representations for semantic segmentation.” In: *European conference on computer vision*. Springer, pp. 173–190.

Zanfir, Andrei, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu (2020). “Weakly supervised 3d human pose and shape reconstruction with normalizing flows.” In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, pp. 465–481.

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

Zeiler, Matthew D (2012). “Adadelta: an adaptive learning rate method.” In: *arXiv preprint arXiv:1212.5701*.

Zhang, Hongwen, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun (2021a). “Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop.” In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11446–11456.

Zhang, Jason Y, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik (2019). “Predicting 3d human dynamics from video.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7114–7123.

Zhang, Mingyuan, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu (2024a). “Motiondiffuse: Text-driven human motion generation with diffusion model.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Siwei, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo (2024b). “RoHM: Robust Human Motion Reconstruction via Diffusion.” In: *CVPR*.

Zhang, Siwei, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang (2023a). “Probabilistic Human Mesh Recovery in 3D Scenes from Egocentric Views.” In: *arXiv preprint arXiv:2304.06024*.

Zhang, Siwei, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang (2022). "Egobody: Human body shape and motion of interacting people from head-mounted devices." In: *European Conference on Computer Vision*. Springer, pp. 180–200.

Zhang, Yahui, Shaodi You, and Theo Gevers (2021b). "Automatic calibration of the fisheye camera for egocentric 3d human pose estimation from a single image." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1772–1781.

Zhang, Yu, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei (2023b). "Dynamic Inertial Poser (DynaIP): Part-Based Motion Dynamics Learning for Enhanced Human Pose Estimation with Sparse Inertial Sensors." In: *arXiv preprint arXiv:2312.02196*.

Zhao, Dongxu, Zhen Wei, Jisan Mahmud, and Jan-Michael Frahm (2021). "EgoGlass: Egocentric-View Human Pose Estimation From an Eyeglass Frame." In: *International Conference on 3D Vision (3DV)*.

Zheng, Yang, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas (2022). "Gimo: Gaze-informed human motion prediction in context." In: *European Conference on Computer Vision*. Springer, pp. 676–694.

Zhou, B., A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba (2016a). "Learning Deep Features for Discriminative Localization." In: *CVPR*.

Zhou, Xiaowei, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis (2016b). "Sparseness meets deepness: 3d human pose estimation from monocular video." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4966–4975.

Zhou, Xingyi, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei (2017). "Towards 3d human pose estimation in the wild: a weakly-supervised approach." In: *Proceedings of the IEEE international conference on computer vision*, pp. 398–407.

Zhou, Yuxiao, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu (2021). "Monocular real-time full body capture with inter-part correlations." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4811–4822.

Zimmermann, Christian, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox (2019). "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 813–822.