Ego+X: An Egocentric Vision System for Global 3D Human Pose Estimation and Social Interaction Characterization

Yuxuan Liu, Jianxin Yang, Xiao Gu, Yao Guo, *Member*, *IEEE*, Guang-Zhong Yang, *Fellow*, *IEEE*

***Abstract*— Egocentric vision is an emerging topic, which has demonstrated great potential in assistive healthcare scenarios, ranging from human-centric behavior analysis to personal social assistance. Within this field, due to the heterogeneity of visual perception from first-person views, egocentric pose estimation is one of the most significant prerequisites for enabling various downstream applications. However, existing methods for egocentric pose estimation mainly focus on predicting the pose represented in the camera coordinates from a single image, which ignores the latent cues in the temporal domain and results in less accuracy. In this paper, we propose Ego+X, an egocentric vision based system for 3D canonical pose estimation and human-centric social interaction characterization. Our system is composed of two head-mounted egocentric cameras, where one is faced downwards and the other looks outwards. By leveraging the global context provided by visual SLAM, we first propose *Ego-Glo* for spatial-accurate and temporal-consistent egocentric 3D pose estimation in the canonical coordinate system. With the help of an egocentric camera looking outwards, we then propose *Ego-Soc* by extending Ego-Glo to various social interaction tasks, e.g., object detection and human-human interaction. Quantitative and qualitative experiments have been conducted to demonstrate the effectiveness of our proposed Ego+X.**

I. Introduction

Egocentric vision can offer sufficient information about how people perceive the world and interact with the environment from a human-centric perspective, furnishing diverse opportunities for the analysis of human behavior and cognition [1], [2]. In general, an egocentric vision system can be composed of either head-mounted or chest-mounted cameras, capturing the visual data in a free-living environment with increased mobility and flexibility. Recent progress in egocentric vision has been widely penetrated in daily life assistive healthcare, including human behavior analysis [2], human-machine interaction [3], and social interaction characterization and assistance [4], [5].

Human pose estimation is one of the most important topics in egocentric vision, which is the prerequisite of human-object interaction or human-human interaction in social assistance. However, due to extremely different viewpoints between egocentric and third-person-view cameras, conventional 3D pose estimation methods learned from third-person-view images or videos can hardly work on egocentric images. Recent research efforts have been devoted to the egocentric 3D pose estimation [6]–[11] and the practicability of proposed methods have been demonstrated. According to the orientation of the egocentric camera, existing methods can be divided into looking outwards [6], [7] and looking downwards [8]–[11]. Outward cameras focus more on perceiving the interaction with surrounding environments but with less observations on human target him/herself. Differently, egocentric cameras looking downwards, especially using a fisheye lens, can capture full human body within the field of view, which is beneficial for achieving accurate human pose estimation [9]–[11]. However, as shown in Fig. 1, significant distortion introduced by the fisheye lens and severe lower limb occlusions are inherent challenges in previous works, leading to degraded or inaccurate pose estimation. Our previous work proposed EgoFish3D, an egocentric 3D pose estimation method via self-supervised learning [11], which improves the performance on egocentric images. Nevertheless, due to the egocentric vision system keeps moving in practice, the estimated poses of previous works represented in the egocentric camera coordinate system will inevitably limit the usability in real-world applications [12], [13]. Therefore, for egocentric vision systems, how to achieve spatial-accurate and temporal-consistent 3D pose estimation in the world coordinate system is the key issue to be solved.

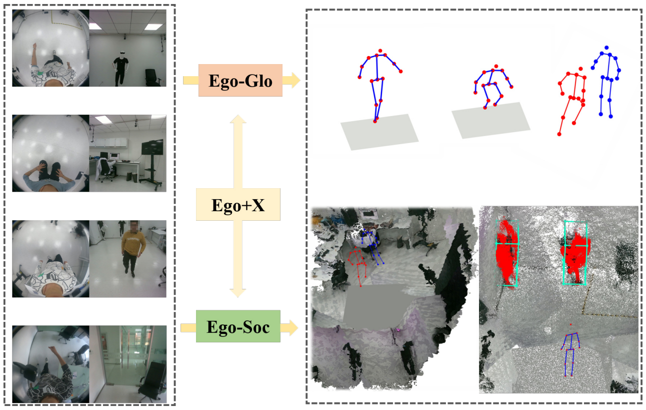

Fig. 1. Illustration of the proposed egocentric vision system for social interaction characterization based on 3D canonical human pose estimation.

Since egocentric vision systems, especially for cameras looking outwards, can provide human-centric perception, increasing popularity has been gained for developing egocentric vision based social interaction characterization, including

This work was supported by the Science and Technology Commission of Shanghai Municipality under Grant 20DZ2220400, and in part by Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University under Grant 21TQ1400203.

Y. Liu, J. Yang, Y. Guo and G.-Z. Yang are with the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China. (20000905lyx@sjtu.edu.cn, jianxinyang@sjtu.edu.cn, yao.guo@sjtu.edu.cn, gzyang@sjtu.edu.cn)

X. Gu is with the Hamlyn Centre, Imperial College London, UK. (xiao.gu17@imperial.ac.uk)

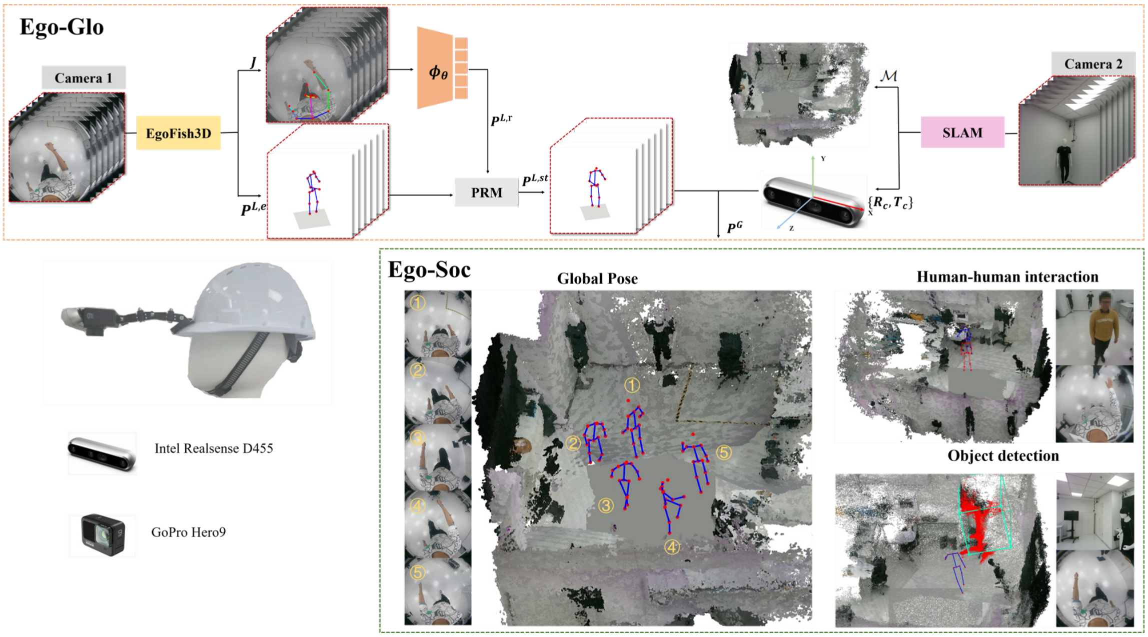

Fig. 2. Overview of our proposed **Ego+X** system, which consists of two egocentric cameras looking downwards and outwards, respectively. The black arrows indicate the direction of information flow. For **Ego-Glo**, egocentric 3D human pose estimation is first performed on each downward egocentric image. We propose a rectify branch and a local pose refinement module (PRM) to improve the performance of local 3D pose estimation. Meanwhile, the outward egocentric camera perceives the surrounding environment and provides the global localization of the system. Thus, the spatial-accurate and temporal-consistent 3D pose represented in the canonical coordinate system can be derived by leveraging global constraints and temporal clues. At last, the **Ego-Soc** enables the downstream social interaction characterization by incorporating advanced object detection, third-person-view human pose estimation, human-human interactions, etc. PRM: Local Pose Refine Module.

human-human or human-robot interaction [14], [15], egocentric object detection [16], [17], egocentric action recognition/anticipation [18]–[20]. It should be pointed out that previous works mainly focus on the interaction or detection within the field of view, ignoring the global context of human target and surrounding environment. Besides, only hand information is leveraged due to outward cameras, where the full-body pose has not been considered. Hence, it is advantageous to represent the 3D human pose and surrounding environment in a canonical coordinate system, which could further facilitate various social interaction tasks and bring opportunities for new research topics.

To address aforementioned challenges, this paper proposes Ego+X, an egocentric vision system consisting of two egocentric cameras looking outwards and downwards, respectively. Our Ego+X contains two parts: **Ego-Glo** for egocentric global pose estimation and **Ego-Soc** for egocentric social behaviour visualization. First, based on our previous EgoFish3D [11], we propose Ego-Glo for the estimation of 3D canonical human pose and the perception of the surrounding environment. Ego-Glo first uses the downward camera to estimate the 3D pose of the human target, in which a dual branch network is proposed to correct the spatial error and the temporal error is reduced by a smoothing method. Next, the global information of the egocentric vision system, as well as the head pose, is determined by the outward camera performing visual SLAM [21]. By fusing the outputs of these two modules, spatial-accurate and temporal-stable 3D pose estimation can be achieved in the canonical coordinate system. In addition, we propose Ego-Soc to perform different egocentric social interaction characterization, such as object detection and human-human interaction. In specific, we take advantage of the global pose predicted by Ego-Glo and use the RGB-D camera looking outwards to collect rich information of two modalities. Along this line, the social activities of the human target can be well visualized and further explored. For egocentric global pose estimation, we evaluate our method on ECHA dataset [11] and on video sequences with ground truth of both camera and human pose collected by a VICON motion capture system. For characterizing social interactions, we demonstrate the qualitative results of egocentric object detection and human-human interaction.

In summary, the main contributions of this paper are:

- An egocentric vision system consisting of two cameras looking downwards and outwards respectively is proposed for social interaction characterization. Its capability in various downstream tasks is demonstrated.
- An effective framework for 3D canonical pose estimation from an egocentric fisheye camera is developed, in which a pose refine module is proposed to improve the estimation in both temporal and spatial domains.

II. Egocentric Vision System: Ego+X

The overview of Ego+X system is shown in Fig. 2. The Ego+X system contains two head-mounted egocentric cameras, one is looking downwards with a fisheye lens to capture RGB images for pose estimation, and the other one is looking outwards to capture RGB-D images for camera localization and other downstream applications. The details of the camera placement can also be found in Fig. 2. In Ego-Glo module, we take the egocentric videos from the downward camera as input to estimate spatial-accurate and temporal-consistent 3D human pose in the canonical coordinate system. In Ego-Soc module, we utilize the global pose predicted by Ego-Glo as self-localization of the human target, and then use the outward camera to perform egocentric object detection and human-human interaction for social characterization.

A. Ego-Glo: Egocentric Global Pose Estimation

In Ego-Glo, we aim to estimate the global 3D body pose sequences from an egocentric video. The proposed method takes T egocentric frames $\mathbf{I} = \{I_1, ..., I_T\}$ as input, and outputs the global 3D human poses $\mathbf{P}^G = \{P_1^G, ..., P_T^G\}$. In this section, we first review the local 2D pose and 3D pose estimation from a single frame by our previous method EgoFish3D [11]. Next, we propose the 3D Pose Refine Module (PRM) to achieve spatial-accurate and temporal-consistent pose estimation. Finally, we extract the camera pose by a visual SLAM system [21] and compose the global pose estimation to get the final output global 3D pose sequences.

1) Local Pose Estimation by EgoFish3D: In our previous work [11], we propose EgoFish3D, which achieves egocentric 2D/3D pose estimation in a self-supervised manner. In [11], we designed three different modules to perform both third-person view and egocentric view pose estimation as well as predicting the transformation between two different viewpoints. The experimental results on the benchmark synthetic datasets [9], [10] and our proposed real-world ECHA dataset demonstrate the effectiveness of our method. Here we directly apply EgoFish3D as the backbone model for local pose estimation, which is denoted as f_θ. Given a sequence of egocentric images $\mathbf{I}$, we have $[\mathbf{J}, \mathbf{P}^{L,e}] = f_\theta(\mathbf{I})$, where $\mathbf{J} = \{J_1, ..., J_T\}$ and $\mathbf{P}^{L,e} = \{P_1^{L,e}, ..., P_T^{L,e}\}$ indicates 2D and 3D local poses estimated by EgoFish3D, respectively. Note that, each pose is consists of 15 body joints and the subscript $\{L, e\}$ indicates the **e**stimated pose represented in the **L**ocal camera coordinates. Since EgoFish3D is trained for pose estimation from a single frame without considering the temporal constraints, the estimated pose sequence $\mathbf{P}^{L,e}$ are prone to transformation error in the spatial domain and unstable jitters in the temporal domain. We refer readers to our previous EgoFish3D [11] for more details of the network structure.

2) Local Pose Refine based on PRM: In this paper, we develop a Pose Refine Module (PRM) to refine the 3D pose estimation in both spatial and temporal domains. In the spatial domain, an efficient rectify branch ϕ_θ is designed to correct the transformation error introduced by the self-supervised EgoFish3D method. Given the camera intrinsic parameters $\mathbf{K}$ and the 2D poses $\mathbf{J}$, we train the rectify branch to estimate the absolute depth maps $\mathbf{Z} = \{z_1, ..., z_T\}$ of 2D joints. Then we can derive the 3D local poses $\mathbf{P}^{L,r}$ by reprojection formula. Pose sequences estimated from the rectify branch can be used to refine the local poses $\mathbf{P}^{L,e}$ mentioned above. The network architecture of the rectify branch is similar to the CNN backbone of the egocentric module in EgoFish3D with only the last several MLPs changed to output the absolute depth of each joint, i.e., $\mathbf{Z} = \phi_\theta(\mathbf{J})$, where θ is the parameters of the rectify branch. As in Eq. (1), the rectify branch is trained with the loss between the estimated pose $\mathbf{P}^{L,r}$ and the triangulated pose $\mathbf{P}^{tri}$ in egocentric camera coordinates and the bone-length loss to force the left and right links $\mathbf{B}$ to be the same. The triangulated pose $\mathbf{P}^{tri}$ is calculated by triangulation method from external cameras as pseudo labels and more details about the triangulated pose can be referred to our previous work [11].

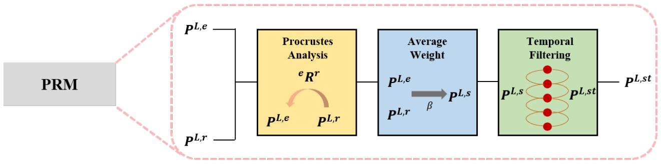

Fig. 3. Overview of our proposed Pose Refine Moduel (PRM), which consists of a rectify branch and a three-step data processing flow. We perform procrustes analysis, average weighting and temporal filtering on the pose sequences to output the spatially accurate and temporally consistent 3D pose estimation.

$$\mathcal{L}_{rec} = \sum_i ||\mathbf{P}_i^{L,r} - \mathbf{P}_i^{tri}||_2 + \frac{1}{\lambda} \sum_j ||\mathbf{B}_j^{left} - \mathbf{B}_j^{right}||_2 \quad (1)$$

From the dual-branch network, we extract two local pose sequences $\mathbf{P}^{L,e}$ and $\mathbf{P}^{L,r}$. We apply Procrustes Analysis (PA) [22] to extract the rotation matrix ${}^r\mathbf{R}_e$ for alignment. After conducting PA, we average the weights of two pose sequences with coefficient β to improve the further refined local 3D pose output, which is denoted as $\mathbf{P}^{L,s}$ in Eq. (2).

$$P_i^{L,s} = \beta P_i^{L,r} + (1-\beta)^r\mathbf{R}_e P_i^{L,e} \quad (2)$$

In temporal domain, for simplicity and effectiveness we directly apply filtering methods, i.e. Gauss filtering $f_g(\cdot)$ and Kalman filtering $f_k(\cdot)$, to make the local pose temporally stable. The final output of the PRM is pose sequences $\mathbf{P}^{L,st} = f_k(f_g(\mathbf{P}^{L,s}))$, where subscript $\{st\}$ indicates **s**patial and **t**emporal refinement, respectively. The detailed structure of PRM is shown in Fig. 3.

3) Egocentric camera localization: For global 3D human pose estimation, we need to obtain the camera pose in the world coordinate system. Here we implement the RGB-D version of ORB-SLAM2 [21] for tracking the 6D pose of the egocentric vision system. We found that when using the images captured by the camera looking downwards, few available ORB features can be detected, especially in cluttered indoor scenes, leading to a failure of tracking the camera trajectory. To solve this, we incorporate another head-mounted RGB-D camera looking outwards in our Ego+X system for performing camera localization. Note that the global position transformation matrix between two cameras is known. Hence, the 6D pose sequences $\mathbf{T}^c = \{(R_i^c, t_i^c)\}$ of the Ego+X system can be extracted by ORB-SLAM. In addition, the Kalman filter is applied to predict the camera pose when SLAM fails to track and smooth the raw data. Similar to [12], we also incorporate the visual SLAM to build the map $\mathcal{M}$ of the environment based on dense point clouds for better visualization shown in Fig. 2.

4) Information fusion for global 3D human pose: With the refined local pose $\mathbf{P}^{L,st}$ and camera pose $\{(R_i^c, t_i^c)\}$, the local pose can be transformed to the global coordinate system as $\mathbf{P}^G$ by Eq. (3).

$$P_i^G = R_i^c P_i^{L,st} + t_i^c \tag{3}$$

B. Ego-Soc: Egocentric Social Interaction Characterization

The global human pose estimated by Ego-Glo can provide both pose information and self-localization of the target human in a canonical coordinate system, thus leading to diverse social interaction characterization and assistance applications. In Ego-Soc, two downstream social characterization tasks are performed based on the global 3D human pose $\mathbf{P}^G$, i.e., human-object interaction (HOI) via egocentric object detection and pose-based human-human interaction (HHI).

1) Egocentric object detection: The detection of 3D objects within the field of view plays a crucial role in modeling HOI, which can be further applied in healthcare, personal assistance and long-term monitoring. Given the global pose $\mathbf{P}^G$ as prior, we first implement the existing 2D object detection method [23] on the RGB images captured from the outward RGB-D camera, extracting the 2D proposal of the object of interest with both bounding box and semantic segmentation. With depth maps, the 3D proposals of the target objects $\mathbf{O}^L$ can be transformed into a 3D point cloud represented in the camera coordinate system. Next, 3D proposals $\mathbf{O}^G$ in the world coordinate system can be derived by using real-time camera pose from visual SLAM, which can be highlighted on the pre-build world map $\mathcal{M}$. Thus, the HOI characterization can be expressed as $\{\mathbf{P}^G, \mathbf{O}^G, \mathcal{M}\}$, which can be used for further applications, such as personal assistance, obstacle avoidance, and navigation.

2) Pose-based Human-human interactions: Pose-based human-human interaction is an important research area in human behavior and cognition analysis. Previous work [14] used a chest-mounted egocentric camera looking outwards to simultaneously estimate the pose of a second-person-view human and predict the pose of the wearer him/herself. However, the outward egocentric camera can only capture a tiny part of the wearer, thus leading to inaccurate and inconsistent pose estimation. To solve this, our proposed Ego+X system contains two cameras that simultaneously capture images of both the interacting person and the wearer. The 3D global pose $\mathbf{P}^G_{self}$ of the wearer him/herself can be estimated by Ego-Glo module. Here, the 3D pose $\mathbf{P}^G_{second}$ of the interacting person can be estimated with the outward RGB-D camera. We first implement OpenPose [24] to extract the 2D pose of the interacting person from RGB images. For the low-confidence and untracking 2D joints, interpolation and temporal filtering are used for smoothing. By using additional depth maps and the 6D camera pose, the global 3D human pose $\mathbf{P}^G_{second}$ of the interacting person can be derived. Consequently, the global 3D poses of both the wearer and the interacting person can be represented in the same canonical coordinate system with pre-build map, noted as $\{\mathbf{P}^G_{self}, \mathbf{P}^G_{second}, \mathcal{M}\}$, which can well describe the characteristics of HHI. In future work, we will use the proposed pose-based HHI characterization for social activity analysis.

III. Experiments

A. Dataset

Following [11], we directly apply the EgoFish3D model trained on ECHA dataset for single-frame 3D pose estimation. To demonstrate the effectiveness of our proposed PRM module on local pose refinement, we conducted several experiments on ECHA test dataset, which consists of 7 different video sequences with 4 subjects performing 10 actions. To evaluate the performance of global pose estimation by proposed Ego-Glo, we capture a new real-world human movement dataset, noted as ECHA-Glo dataset, which contains five video sequences about 7k frames of human walking and other actions with different body textures. The ECHA dataset and ECHA-Glo dataset both provide the ground truth collected by the VICON motion capture system to conduct quantitative results. To evaluate the feasibility of our Ego-Soc system for social interaction characterization, we capture several video sequences for egocentric object detection and pose-based human-human interactions for qualitative results.

B. Implementation Details

The architecture of the rectify branch ϕ_θ is similar to the structure of the egocentric module in EgoFish3D. It consists of an encoder-regressor network which is build upon three CNN layers along with four MLP blocks. We train the rectify branch on the ECHA dataset under the supervision of the absolute joints depth extracted by reprojection methods with loss in Eq. (1) with $\lambda = 20$. The network of the rectify branch is implemented by PyTorch, and we choose Adam for optimization during training for 20 epochs. In PRM, we find that $\beta = 0.5$ generalizes best for pose estimation under different scenarios. Then we choose Gaussian filter with a smoothing window of 10 frames and a Kalman filter with a covariance of 0.5 to conduct temporal smoothing. Two different cameras are well-calibrated with a chessboard to extract the intrinsic and extrinsic parameters.

C. Evaluation Metrics

We evaluate our method with three different metrics to provide quantitative results. One refers to Mean Per Joint Position Error (MPJPE), which calculates the Euclidean distances between ground truth and estimated 3D pose in Eq. (4). Another refers to PA-MPJPE, which calculates the MPJPE after applying Procrustes Analysis on translation, rotation and scale between ground truth and estimated 3D pose. The other refers to Bn-MPJPE, which calculates the MPJPE after applying Procrustes alignment for all the poses in a batch with the certain number n. In our experiments, we report B50-MPJPE and B100-MPJPE, which means the

TABLE I

Comparison MPJPE results of egocentric 3D local pose estimation in millimeters (mm) on ECHA dataset

Approaches	All	squatting	Walking	Dancing	Stretching	Waving	Boxing	Kicking	Touching	Clamping	Knocking
EgoFish3D [11]	107.9	123.8	106.8	110.4	121.4	95.6	111.2	94.6	110.5	101.6	102.7
+PA	89.6	103.2	88.9	98.0	92.4	75.3	88.5	85.8	83.0	96.7	83.7
+PA&Avg	81.8	94.6	81.7	85.4	81.7	66.5	80.6	80.5	76.0	91.2	78.1
+PRM (Ours)	**78.9**	**92.9**	**79.2**	**82.6**	**79.5**	**63.5**	**77.2**	**78.2**	**73.2**	**84.1**	**75.4**
Ablated models	All	squatting	Walking	Dancing	Stretching	Waving	Boxing	Kicking	Touching	Clamping	Knocking
w/o PA&Avg	105.1	121.9	103.8	106.9	118.8	93.2	108.0	92.3	108.2	97.4	100.1
w/o Avg	87.1	101.5	86.6	95.4	90.5	72.8	85.5	83.4	80.7	91.4	81.5
w/o PA	83.9	100.4	84.2	84.1	89.2	70.0	83.6	79.3	79.7	82.7	81.2
w/o Filt	81.8	94.6	81.7	85.4	81.7	66.5	80.6	80.5	76.0	91.2	78.1
Ours	**78.9**	**92.9**	**79.2**	**82.6**	**79.5**	**63.5**	**77.2**	**78.2**	**73.2**	**84.1**	**75.4**

TABLE II

Comparison results in millimeters (mm) on ECHA dataset

Approaches	MPJPE	PA-MPJPE
EgoFish3D [11]	107.9	73.1
EgoFish3D+PRM	**78.9**	**63.3**
Tome [9]	112.4	73.9
Tome+PRM	80.9	64.6
Martinez [25]	118.3	80.0
Martinez+PRM	82.8	66.8

TABLE III

Comparison global 3D pose estimation results in millimeters (mm) on ECHA-Glo dataset

With GT $\mathbf{T}_c$	MPJPE	B50-MPJPE	B100-MPJPE
Ours	**104.0**	**87.5**	**92.1**
w/o Filt	105.6	90.9	95.0
w/o PRM	120.0	103.6	109.2
With EST $\mathbf{T}_c$	PA-MPJPE	B50-MPJPE	B100-MPJPE
Ours	**73.2**	**103.5**	**124.0**
w/o Filt	74.5	107.1	127.4
w/o PRM	83.9	116.0	135.6

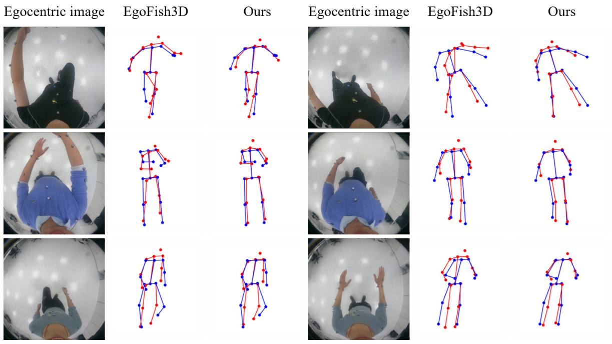

Fig. 4. Visualization results of the improvement of our proposed PRM on ECHA test dataset. The red color is the predicted 3D pose by EgoFish3D or our method. The blue color is the ground truth 3D pose.

corresponding batch number is 50 and 100, respectively.

$$E(\mathbf{P}, \hat{\mathbf{P}}) = \frac{1}{T}\sum_{i=1}^{T}||P_i - \hat{P}_i||_2 \tag{4}$$

D. Comparison methods

On the one hand, to prove the effectiveness of our global pose estimation method Ego-Glo, we conduct several experiments to do comparison and ablation studies. First, we conduct experiments on the ECHA test dataset to evaluate the performance of PRM module. We compare our proposed method with three baseline methods [9], [11], [25]. For ablation studies, we validate the influence of the PRM module step by step by removing or changing each part. We provide quantitative results and qualitative results to give a clear view. Second, we conduct experiments on the new ECHA-Glo dataset to evaluate the global 3D pose estimation with both quantitative and qualitative results. Besides, we present our Ego-Soc method's qualitative results in egocentric object detection and pose-based HHI. The quantitative experiments are defined as follows.

- Baselines: we apply three different models as the baseline method, i.e., Martinez [25], Tome [9], EgoFish3D [11].
- +PA: rotation ${}^{\mathrm{r}}\mathbf{R}_{\mathrm{e}}$ based procrustes analysis by a rectify branch network.
- +PA&Avg: rotation ${}^{\mathrm{r}}\mathbf{R}_{\mathrm{e}}$ based procrustes analysis and average weight of $\mathbf{P}^{L,e}$ and $\mathbf{P}^{L,r}$ for pose refinement.
- +PRM: full local pose refinement module.
- w/o PRM: remove the local pose refine and directly apply the pose $\mathbf{P}^{L,e}$ by baseline method as the final output.
- w/o Filt: remove the temporal smoothing $f_g(\cdot)$ and $f_k(\cdot)$.
- w/o Avg: remove the average weight β.
- w/o PA: remove the rotation based procrustes analysis.
- w/o PA&Avg: only refine the local pose in temporal domain by filtering the pose sequences.

IV. Results and Analysis

A. Quantitative Results

Without further clarification, all the elements indicate the result in millimeters (mm). By incorporating a full-body gait model provided by the VICON motion capture system, the ground truth 3D body joints are from the anatomical level.

1) ECHA dataset: We first evaluate our proposed PRM with three different pose estimation methods, and report the MPJPE and PA-MPJPE for all test data in Table II. EgoFish3D with our PRM module achieves the best performance among other approaches (MPJPE=78.9 and PA-

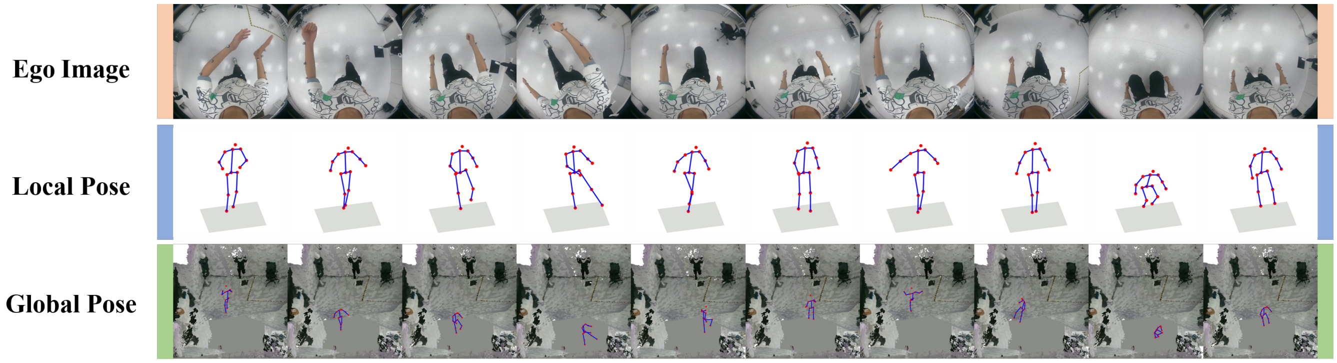

Fig. 5. Visualization of the estimated global 3D human pose. The first row indicates the input egocentric images with different actions. The second row indicates the local pose estimation by our proposed PRM. The third row visualizes the global pose in the pre-build map. The red points are the predicted joint positions and the colorized lines represent the skeleton.

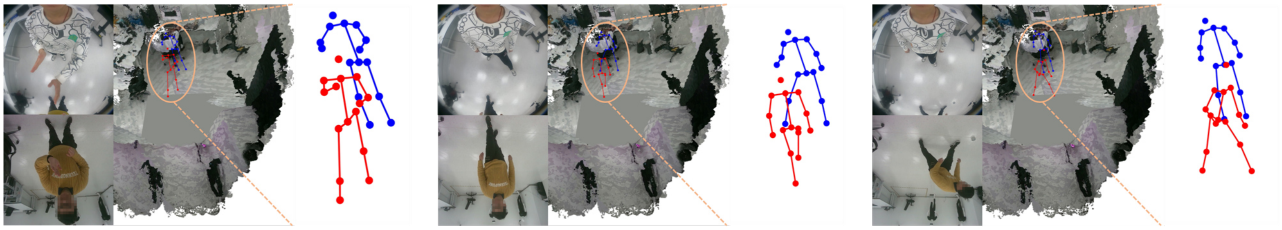

Fig. 6. Visualization results of human-human interaction based on our Ego+X. The blue color is the predicted 3D pose of the human in the looking-downwards camera, and the red color is the predicted 3D pose of the human in the looking-outwards camera.

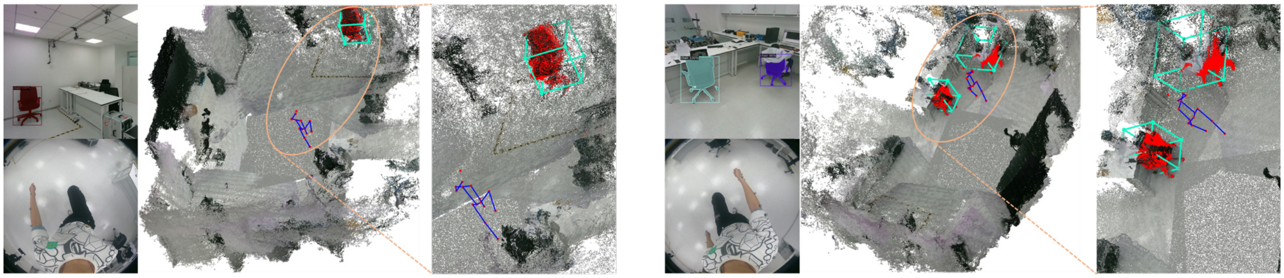

Fig. 7. Visualization results of object detection based on our Ego+X. The red points are the predicted joint positions and the colorized lines represent the skeleton. The highlighted point clouds (red) are the detected objects matched in the global coordinate system.

MPJPE=63.3). It can also be found that after applying our proposed PRM, the local pose improves a lot (nearly 30mm improvement in MPJPE and 10mm improvement in PA-MPJPE), which significantly proves the effectiveness and generalization ability of our proposed local pose refine method. We then conduct experiments on the influence of each part of the PRM module and report MPJPE for all test data and each action in Table I. The upper part of the table shows the improvement of the local pose step by step and the lower part removes some parts of the PRM to do ablation studies. It can be found that three different processes in PRM (noted as PA, Avg and Filt) all play important roles for the local pose refinement.

2) ECHA-Glo dataset: We evaluate the performance of Ego-Glo for global pose estimation on ECHA-Glo dataset, and report three different metrics to conduct the quantitative results in Table III, where GT and EST $\mathbf{T}_c$ indicate the camera poses captured by VICON and estimated by ORB-SLAM2, respectively. The upper part of the table (noted as With GT $\mathbf{T}_c$) refers to that we implement the ground truth camera pose captured by VICON system to compose the global pose for the evaluation. We report MPJPE, B50-MPJPE and B100-MPJPE for the average results, where our method achieves high-accuracy performance (MPJPE=104.0, B50-MPJPE=87.5, B100-MPJPE=92.1). The lower part of the table (noted as With EST $\mathbf{T}_c$) means that we use the camera pose estimated by visual SLAM to compose the global pose. We report PA-MPJPE, B50-MPJPE and B100-MPJPE for evaluation. Since the camera trajectory estimated by visual SLAM has been greatly affected by the surroundings which introduce the scale error compared to ground truth, we report PA-MPJPE instead of MPJPE to remove the influence of the scale ambiguity. Our method with visual SLAM also performs well (PA-MPJPE=73.2, B50-MPJPE=103.5, B100-MPJPE=124.0).

B. *Qualitative Results*

Fig. 4 demonstrates the visualization results of our proposed local pose refine module. The PRM corrects the transformation error introduced by EgoFish3D and can predict more accurate results.

Fig. 5 presents the visualization results of our proposed Ego-Glo for global 3D human pose estimation. The first row represents the input egocentric images, the second row shows the estimated local pose under the camera coordinate system, and the third row visualizes the global 3D human pose represented in the 3D canonical coordinate system with a pre-build 3D map. It can be seen that our Ego-Glo method can predict relatively accurate global 3D human pose with different actions.

Fig. 6 visualizes the social characterization of human-human interaction. We present two human poses under the local camera coordinate system and the global coordinate system with a pre-build map, respectively. We show the interacting scenes like talking, clamping, etc.

Fig. 7 shows the social characterization of human-object interaction based on egocentric object detection. With extracted 2D bounding boxes, we recover the point cloud of the detected object and present combined with our estimated global pose, as well as highlighting the object in the pre-build 3D map. More qualitative results can be found in our attached video.

V. CONCLUSIONS

In this article, we propose Ego+X, an egocentric vision system for global 3D human pose estimation and extend it to human-centric social interaction characterizations, which is achieved by using two head-mounted egocentric cameras looking outwards and downwards, respectively. Specifically, in Ego-Glo, we design a local pose refine module (PRM) to correct the 3D pose from both spatial and temporal domains and also combine visual SLAM to generate the spatial-accurate and temporal-consistent 3D pose in a canonical coordinate system. In Ego-Soc, we extend the global human pose estimation to the applications of egocentric object detection and human-human interactions. The experimental results prove the effectiveness of our method. In the future, we will apply our method to conduct scientific research for human-centric behavior analysis and cognition evaluations.

REFERENCES

[1] I. Rodin, A. Furnari, D. Mavroeidis, and G. M. Farinella, "Predicting the future from first person (egocentric) vision: A survey," *Computer Vision and Image Understanding*, vol. 211, p. 103252, 2021.

[2] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara, "Understanding social relationships in egocentric vision," *Pattern Recognition*, vol. 48, no. 12, pp. 4082–4096, 2015.

[3] D. Kim, B. B. Kang, K. B. Kim, H. Choi, J. Ha, K.-J. Cho, and S. Jo, "Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view," *Science Robotics*, vol. 4, no. 26, p. eaav2949, 2019.

[4] Z. Zuo, L. Yang, Y. Peng, F. Chao, and Y. Qu, "Gaze-informed egocentric action recognition for memory aid systems," *IEEE Access*, vol. 6, pp. 12 894–12 904, 2018.

[5] J. Qiu, F. P.-W. Lo, X. Gu, Y. Sun, S. Jiang, and B. Lo, "Indoor future person localization from an egocentric wearable camera," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8586–8592.

[6] H. Jiang and K. Grauman, "Seeing invisible poses: Estimating 3d body pose from egocentric video," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3501–3509.

[7] Z. Luo, R. Hachiuma, Y. Yuan, and K. Kitani, "Dynamics-regulated kinematic policy for egocentric pose estimation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[8] J. Wang, L. Liu, W. Xu, K. Sarkar, and C. Theobalt, "Estimating egocentric 3d human pose in global space," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 500–11 509.

[9] D. Tome, T. Alldieck, P. Peluse, G. Pons-Moll, L. Agapito, H. Badino, and F. De la Torre, "Selfpose: 3d egocentric pose estimation from a headset mounted camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[10] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt, "Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 5, pp. 2093–2101, 2019.

[11] Y. Liu, J. Yang, X. Gu, Y. Guo, and G.-Z. Yang, "Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning," *https://doi.org/10.36227/techrxiv.18516119.v1*, 2022.

[12] Y. Guo, F. Deligianni, X. Gu, and G.-Z. Yang, "3-d canonical pose estimation and abnormal gait recognition with a single rgb-d camera," *IEEE Robotics and Automation letters*, vol. 4, no. 4, pp. 3617–3624, 2019.

[13] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4318–4329.

[14] E. Ng, D. Xiang, H. Joo, and K. Grauman, "You2me: Inferring body pose in egocentric video via first and second person interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9890–9900.

[15] N. F. Duarte, M. Raković, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor, "Action anticipation: Reading the intentions of humans and robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4132–4139, 2018.

[16] R. Ye, W. Xu, Z. Xue, T. Tang, Y. Wang, and C. Lu, "H2o: A benchmark for visual human-human object handover analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 762–15 771.

[17] K. Lee, A. Shrivastava, and H. Kacorri, "Leveraging hand-object interactions in assistive egocentric vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[18] X. Gu, J. Qiu, Y. Guo, B. Lo, and G.-Z. Yang, "Transaction: Icl-sjtu submission to epic-kitchens action anticipation challenge 2021," *arXiv preprint arXiv:2107.13259*, 2021.

[19] M. Lu, Z.-N. Li, Y. Wang, and G. Pan, "Deep attention network for egocentric action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3703–3713, 2019.

[20] Y. Wu, L. Zhu, X. Wang, Y. Yang, and F. Wu, "Learning to anticipate egocentric actions by imagination," *IEEE Transactions on Image Processing*, vol. 30, pp. 1143–1152, 2020.

[21] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[22] A. Ross, "Procrustes analysis," *Course report, Department of Computer Science and Engineering, University of South Carolina*, vol. 26, 2004.

[23] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[24] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[25] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.