EgoGlass: Egocentric-View Human Pose Estimation From an Eyeglass Frame

Dongxu Zhao Zhen Wei Jisan Mahmud Jan-Michael Frahm
University of North Carolina at Chapel Hill
{dongxuz1,zhenni,jisan,jmf}@cs.unc.edu

Abstract

We present a new approach, EgoGlass, towards egocentric motion-capture and human pose estimation. EgoGlass is a lightweight eyeglass frame with two cameras mounted on it. Our first contribution is a new egocentric motion-capture device that adds next to no extra burden on the user and a dataset of real people doing a diverse set of actions captured by EgoGlass. Second, we propose to utilize body part information for human pose detection - to help tackle the problems of limited body coverage and self-occlusions caused by the egocentric viewpoint and cameras' proximity to the human body. We also propose a concept of pseudo-limb mask as an alternative for segmentation mask when ground truth segmentation mask is absent for egocentric images with real subject. We demonstrate that our method achieves better results than the counterpart method without body part information on our dataset. We also test our method on two existing egocentric datasets: xR-EgoPose and EgoCap. Our method achieves state-of-the-art results on xR-EgoPose and is on par with existing method for EgoCap without requiring temporal information or personalization for each individual user.

1. Introduction

For the head-worn AR/VR devices that can capture human motion, heavy headset or cameras stretching out from the wearer add inconvenience and restriction to both the wearer and the environment. We envision that future devices need to be lightweight and the user will be able to wear it in daily activities, *e.g*. Google smart glasses. To this end, we propose a novel prototype, **EgoGlass**, which is eyeglasses augmented with light cameras and barely adds any extra burden to the wearer. It facilitates flexible data acquisition and human pose detection using a lightweight wearable data capture device.

Understanding human pose from EgoGlass requires egocentric-view pose estimation. However, existing human pose estimation methods are usually from a third-person point-of-view [7, 10, 14, 19, 22, 25, 26, 36] - also called *outside-in* methods. These methods need the cameras to be placed around the scene, which adversely restricts the recording volume to a very limited size. Moreover, they often fail when the subject is occluded by other people (*e.g*., close social interaction) or objects (*e.g*., furniture) in the environment.

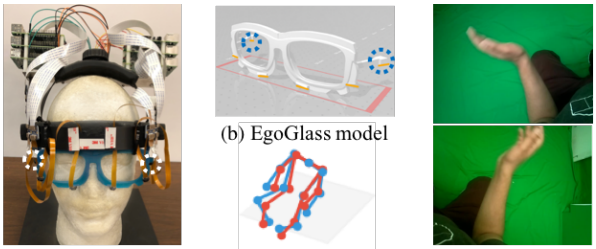


Figure 1. Overview of the egocentric human pose estimation pipeline we proposed. (a) the capture headset we built with EgoGlass attached on it (the headset is only needed when capturing training data); (b) digital model of EgoGlass. Note that for experimental purpose we installed six cameras on it but we finalized to use the two cameras circled in the image; (c) an example visualization of 3D body pose; (d) an example frame captured by EgoGlass consisting of two views.

Due to the aforementioned limitations of the *outside-in* methods, human pose estimation from the egocentric view is of great interest recently, especially with the development of xR technologies (such as AR, VR and MR). Egocentric human pose estimation overcomes the constraint on recording volume imposed by the placement of external cameras. Instead, using wearable capture devices, it allows the users to participate in activities indoor or outdoor without any size constraint. It also handles the occluded scenarios by nature. Moreover, in xR applications, it enables the users to better perceptually immerse themselves into a virtual environment thanks to its special, egocentric viewpoint. While there have been valuable efforts on this task [8, 11, 17, 20, 29, 30, 33], egocentric human pose estimation is yet not thoroughly solved. Some of the previous work [8, 11, 17] rely on sensors worn on human body, while the cumbersome instru-

mentation and extra weight make them not suitable for everyday activities. Others [20, 29, 30, 33] infer the 3D body pose from images captured by head-mounted cameras.

In this paper, we introduce **EgoGlass** (Fig. 1): a novel prototype with two cameras mounted for egocentric motion-capture and human pose estimation. We also create a dataset of real people wearing EgoGlass - towards the task of human pose estimation. Two-step approaches [4, 5, 14, 16, 31, 37] have been popular for human pose detection - consisting of 2D heatmap prediction and 3D lifting. However, these approaches suffer from the low-visibility of some joints. Our key insight is that adding body part information can help the network to reason about human skeleton configuration. Especially in the egocentric view, while having the 2D heatmap information means having the 2D joint location plus uncertainty, body part information will account for the bone configuration even when the joint is not visible.

We validate our EgoGlass design and the essence of multiple cameras by experimenting with a state-of-the-art single-view egocentric method [30]. Our experiments show that multiple camera views allow pose detection with improved accuracy. Our experiments also show that utilizing body part information is useful for pose detection especially when many joints are out of camera views. We also test our proposed method on two existing egocentric datasets, xR-EgoPose [30] and EgoCap [20]. On xR-EgoPose, we decrease the 3D joint reconstruction error by 8%. On EgoCap, our method is on par with existing method without requiring any temporal smooth term or any personalization prior to the capture.

Our work is the first method to motion-capture and human pose estimation from eyeglass frames. Our main contributions include:

- A new egocentric motion-capture approach EgoGlass. EgoGlass can serve towards research on lightweight and portable egocentric devices.
- An egocentric dataset captured by the EgoGlass eyeglass frame.
- A learning algorithm which utilizes body part information for egocentric human pose estimation. The proposed algorithm improves the body pose estimation task especially when a large portion of joints are not visible.

2. Related work

We describe related work on human pose estimation in two categories based on the viewpoint of the camera. We first focus on third-person view approaches and follow with egocentric-view ones. For third-person view pose estimation, we focus on monocular approaches since our method uses two views that can cover the user body as much as when using a single front-facing camera and there is barely overlap between two views. For egocentric-view approaches, we describe their capture setup and methods, as well as the difference between theirs and ours.

Monocular human pose estimation from the third-person view: In the regime of 3D human pose estimation, most approaches fall into one of the two categories: (i) inferring 3D body pose directly from images [12, 13, 15, 18, 21, 26, 28], and (ii) predicting 2D joint positions first and then lifting 2D predictions to 3D predictions [4, 5, 14, 16, 31, 37]. The availability of large datasets with 3D ground truth and advancements of deep neural networks have significantly improved the accuracy of inferring 3D pose directly from images. Among the methods, generating 3D heatmap first as a keypoint localization problem in a discretized 3D space [18, 26], similar to 2D pose estimation tasks, is proved to have advantage over direct regression of 3D joint coordinates [12]. On the other hand, decoupling the step of predicting 2D heatmaps and predicting 3D pose from 2D pose enables us to exploit existing success on 2D pose estimation task (including both methods and datasets) and helps the neural network to explore prior skeletal knowledge when predicting 3D pose.

Egocentric-view motion-capture and pose estimation: Various solutions for egocentric motion-capture have been proposed in recent years. Earlier works are mostly suit-based with the help of non-visual sensors, such as foot pressure-sensor pad [34] or IMU sensors [27]. These suits usually require long setup time and a cumbersome calibration step. Jiang and Grauman [11] proposed to use a single chest-mounted camera looking outwards to infer the full 3D body pose with cues from the surrounding scenes. Ng *et al*. [17] further proposed to leverage cues from interactions with another person as there is often an inherent synchronization between interacting individuals. Hwang *et al*. [8] followed this design but used an ultra-wide fisheye lens which can capture part of the user's body, instead of only leveraging environmental cues. However, the chest-mounted setup would cause inconvenience and discomfort when worn, such as affecting the outfit look.

Rhodin *et al*. [20] proposed to estimate wearer's pose from a pair of fisheye cameras mounted on a V-shape rig on a headgear. They combined a 3D generative body model and a 2D body part detector which was trained using a dataset of multiple people. However, the 3D body model needs personalization of shape and skeleton bone lengths and their two cameras stretch out of the user's body for about 25cm, which severely restrict its use cases. Xu *et al*. presented a capture device based on a single cap-mounted fisheye camera in Mo2Cap2 [33], which significantly decreased the distance between the camera and the user's face to around 8cm. Tome *et al*. [29, 30] further decreased the number to 2cm by installing a fisheye camera on the rim of a

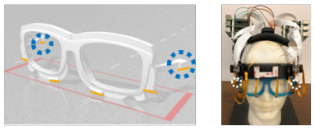
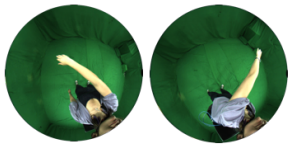
	EgoCap[20]	Mo2Cap2[33]	xR-EgoPose[29,30]	Ours
Device				
Example Data				
Distance to user's face	~25cm from the head	~8cm from the head	~1cm from the head	<1cm from the head

Figure 2. Illustration of the main different aspects among our setup and existing solutions. The images on the first row and distance values on the third row show the capture devices and their distances to the user's head. The second row presents one example frame for each dataset. The first image in our *Device* cell is the prototype of our eyeglass frame with two blue circles showing the locations and angles of the cameras; the second image shows the real eyeglass frame (other parts of the headset are only for training data capture). Given other solutions use fisheye cameras for larger field of views, their views can cover a larger portion of the user's body than ours. However, our device is the smallest and closest to user's head. We had to use cameras small enough to fit the eyeglass frame and cover as much body part as possible at the same time. To this end, we attached two *Raspberry Pi spy cameras* on it.

head mounted virtual reality device. They followed the two-step approach for 3D human pose estimation and proposed a multi-branch architecture to reconstruct 2D heatmaps and joint rotations together with 3D joint positions [29]. See Fig.2 for visual comparison of capture devices and example frames.

In contrast to existing methods, EgoGlass proposes a smaller and lighter capture device. By integrating two *Raspberry Pi spy cameras* on an eyeglass frame, we successfully make the device suitable for almost all daily activities. Our camera choice also avoids the strong distortion from fish-eye cameras. However, we still face the challenge of self-occlusions as other egocentric devices. Moreover, due to the limited field of view of our cameras, even with the two cameras, there is still a larger portion of joints that are out of the views compared to other egocentric datasets. Our proposed pose detection method explicitly utilizes body part information for pose detection, unlike the existing methods, which allows it to improve on the pose detection task especially in occlusion scenarios.

3. EgoGlass dataset

In this section, we introduce a new dataset for egocentric human pose estimation. The images were captured by real human wearing our EgoGlass, which provides the training corpus for egocentric human pose estimation from portable and convenient eyeglass frames. To obtain ground-truth 3D human pose, we made use of an outside-in capture system to estimate pose in world coordinate and projected to our frame-mounted cameras using calibration, which will be detailed in Sec 3.1

3.1. Data acquisition and processing

We designed a capture system that consists of two components: (i) the EgoGlass helmet to capture the user from the egocentric view and (ii) six external cameras for outside-in motion capture. Note that the large helmet is only needed when capturing training data.

Fig. 2 presents the prototype of our EgoGlass frame and the capture helmet. Our goal is to design a lightweight motion-capture setup consisting of small and lightweight cameras. While all existing egocentric datasets use fish-eye cameras, for our setup where camera fields could be easily occluded by nose or cheek, fisheye cameras may not always provide good results. Hence we chose *Raspberry Pi spy cameras* which are tiny and weigh only about 9 grams. While having a very light footprint, the angle of view of a *Raspberry Pi spy camera* is only about 64×48 degrees. In order to compensate for this body part visibility limitation, we mounted two of these cameras with different angles on the eyeglass frame, which we call *bodycams*.

Inspired by Rhodin *et al*. [20] and building upon Cha *et al*. [3], we used six cameras placed around the scene for outside-in motion-capture and applied the state-of-the-art

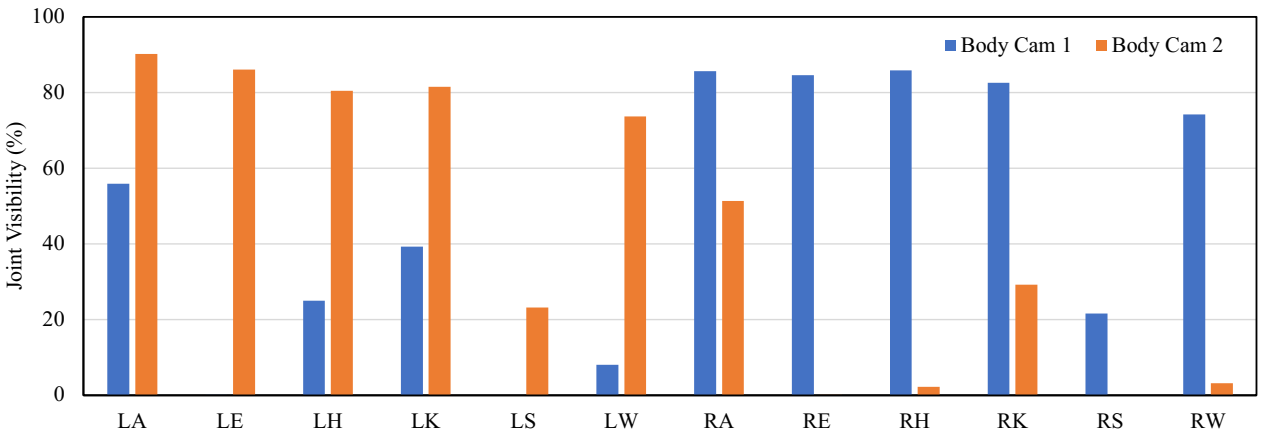


Figure 3. Percentage of coverage of joints from each view aggregated among all subjects and actions. The numbers are computed in the way that dividing the number of frames that a joint is visible in a view by the number of total valid frames. Notation: Ankle (A), Elbow (E), Hip (H), Knee (K), Shoulder (S), and Wrist (W). L/R: Left/Right

human pose estimation method from the third-person view, OpenPose [1, 2, 24, 32], to generate ground truth of 3D pose. Then we projected the ground truth to the coordinate system of each *bodycam*.

3.2. Dataset overview

Our dataset is diverse in terms of both subject appearance characteristics and actions. In the EgoGlass dataset, we captured 10 subjects in total, 5 males and 5 females, with a diversity of heights, body shapes, skin tones and clothing. Each subject did six categories of actions, which are *greeting*, *introducing*, *pointing*, *waiting*, *thinking* and *waving hands*, and they were free to choose the specific poses they would perform within each category. Each subject also repeated their poses three times - while they were *sitting*, *standing* and *walking*. Because of the differences in how a user wears the eyeglasses, *e.g.*, the height of nose is different, a small variation of the body parts within each view exists in the dataset. The variations of views and poses presented by different subjects help the dataset to better generalize to unseen subjects and poses.

The dataset contains 173577 frames in total and each frame has two views (Fig. 1), each of which is captured by one *bodycam* on the eyeglass frame. Table 1 presents the statistics of frames per subject and per action.

3.3. EgoGlass pose estimation challenges

There are two factors making egocentric human pose estimation from our dataset a challenging task. The first one is self-occlusions. Self-occlusion is a common issue shared by all egocentric datasets because of the special camera viewpoints, and the extremely close distance between cameras and the user's face makes it more severe in our dataset than

Subject	Sitting	Standing	Walking
S1	3901	3847	4160
S2	8539	8804	7938
S3	9233	5044	4869
S4	5561	4312	4503
S5	6978	7352	7042
S6	2847	3567	3771
S7	6568	5355	7949
S8	6093	6837	4966
S9	4358	2042	3503
S10	8328	8040	7270

Table 1. Total number of frames per subject per action.

previous works. Thus our method aims at tackling this issue. Fig. 2 provides example frames from existing egocentric datasets. The second challenge is limited body coverage. We attached two cameras on EgoGlass to make them cover a large part of the user's body but occlusions still exist. Due to the cameras' limited field of view and proximity to the human body, when the user is doing stretching poses, the arms or legs may leave the views of all cameras. Fig. 3 provides the statistics of the visibility of each joint within each view. Also see Fig. 2 for a visual comparison of visibility of joints in EgoGlass dataset and other existing egocentric datasets.

4. Method

We adapt the two-step approach for pose detection. Our method consists of two 2D modules to generate 2D heatmaps for two views and a 3D module to generate 3D joint positions using 2D heatmaps as input. The visualiza-

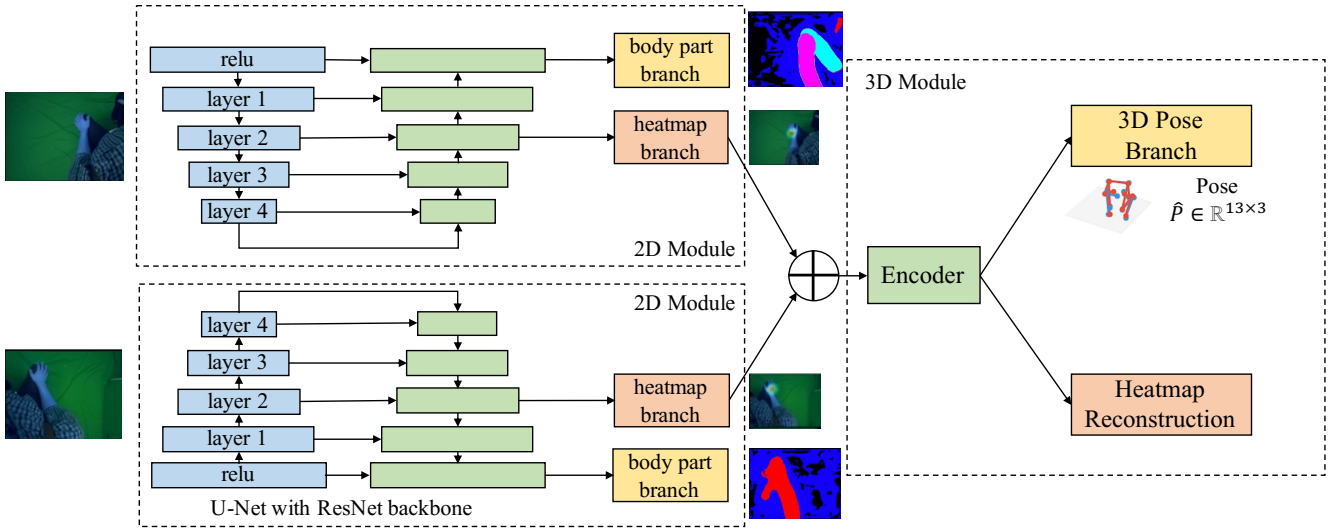


Figure 4. Our architecture consists of two 2D modules and one 3D module.

tion of our network is shown in Fig. 4.

For experimental purpose, we installed six *bodycams* on the eyeglass frame, aiming to enable a larger coverage of human body. However, we find that most of the images captured by the middle four *bodycams* have limited visibility of joints while adversely add large variation among each individual subject.

4.1. 2D heatmap generation

The architecture of our 2D module is based on *U-Net* [23] with *ResNet18* [6] as the backbone for all experiments. The input to the 2D modules is RGB images from two views for our dataset. There are two branches in the upsampling part: one branch for 2D heatmap prediction and the other for body part information prediction. We train one 2D module for one individual view, each of which has the same architecture but does not share weights.

2D heatmap branch: The output of this branch is 2D heatmap, one for each joint except *Neck*. To train this branch, we apply the mean square error (MSE) to the ground truth heatmap HM and the prediction $\hat{HM}$ as loss:

$$L_{2D}(\hat{HM}, HM) = mse(\hat{HM}, HM) \tag{1}$$

body part branch: We propose to explicitly enforce the network to reason about the body part information in the image by adding a body part branch. When a joint is occluded, the image information of body part layout can help the network infer the human skeletal configuration. The intuitive representation for body part information is segmentation masks. For synthetic datasets, segmentation masks can be obtained without difficulty. However, when dealing with images of real people and real scenes, even state-of-the-art segmentation methods cannot provide perfect masks as ground truth given the special egocentric viewpoint. Thus we propose the concept of pseudo-limb mask, which is obtained by connecting the areas between joints on human limbs, including both arms and legs. While lacking real segmentation mask ground truth, this pseudo-limb mask can serve as a rough mask and learning this can improve the accuracy of 3D reconstruction as we demonstrate.

The loss function for this branch is also MSE loss:

$$L_b(\hat{mask}, mask) = mse(\hat{mask}, mask) \tag{2}$$

4.2. 3D module

Our 3D module is an autoencoder architecture, which is inspired by the two-branch decoder model in [30]. It takes the joint heatmaps predicted by the 2D modules as input and pushes them through an encoder to get embedding features. The first branch in the decoder is to generate 3D body pose estimation while the second branch is to enforce the encoding of the uncertainty in 2D prediction by reconstructing the input heatmaps as output. The difference is that we have two sets of heatmaps from both views. In most cases, these two views cover different joints. To fuse these information, we concatenate the heatmaps along the channel dimension.

To train this module, we use the Mean Per Joint Position Error (MPJPE) for the 3D pose branch:

$$L_{reg}(\hat{P}, P) = \frac{1}{N_b}\frac{1}{N_J}\sum_{i=1}^{N_b}\sum_{j=1}^{N_J}||\hat{P}_i^j - P_i^j||_2 \tag{3}$$

where N_b is the batch size and N_J is the number of joints. $\hat{P}$ is the joint position prediction and P is the ground truth. And MSE loss is applied to the heatmap reconstruction branch.

MPJPE(mm)	Lower Body	Upper Body	Full Body
selfPose/*Cam1*	156.5	113.0	134.8
selfPose/*Cam2*	146.4	114.1	130.3
Ours / All	139.0	79.0	**106.7**

Table 2. Quantitative results on our dataset. We apply selfPose [29] to each single egocentric view and apply our method to both views. The results show that two views is minimal number of views and missing either view will downgrade the performance.

We also add the cosine-similarity error as in xR-EgoPose [30]:

$$L_{cos}(\hat{P}, P) = \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{l=1}^{L} \frac{P_i^l \cdot \hat{P}_i^l}{||P_i|| * ||\hat{P}_i||} \tag{4}$$

with $P_i^l \in \mathbb{R}^3$ the l^{th} bone of the pose.

The overall loss function for the whole network becomes

$$\begin{aligned} L_{train} =& \lambda_{2D} L_{2D} + \lambda_b L_b + \lambda_{recon} L_{recon} + \\ & \lambda_{3D}(L_{reg} + \lambda_{cos} L_{cos}) \end{aligned} \tag{5}$$

5. Experiments

We conduct experiments to show the effectiveness of our method on our EgoGlass dataset and compare with two baselines: the first one is a state-of-the-art egocentric human pose estimation method for monocular image proposed by Tome *et al*. [29] and the second one is a method using the same architecture as ours but without the body part branch. Moreover, we test our method on two existing egocentric datasets, EgoCap [20] and xR-EgoPose [30] and show either state-of-the-art or competitive performance with a simpler approach.

5.1. Evaluation metrics

Following the two protocols of Human3.6M [9], we report both the Mean Per Joint Position Error (MPJPE) computed as Equation 3 and the Mean Per Joint Position Error after Procrustes transformation (P-MPJPE). Procrustes transformation is a rigid transformation that aligns the prediction and ground truth in scale, translation, and rotation. It is a meaningful evaluation metric since the main purpose of EgoGlass is to reconstruct the user's body in egocentric view relative to the capture device, regardless of the global joint positions.

5.2. Implementation details

The training set we used is S1, S3, S4, S5, S6, S7, S8 and actions of *sitting* and *standing*; the test set we used is S9 and S10 and actions of *sitting* and *standing*. S2 was used for validation purpose. The resolution of input images and pseudo-limb masks are 128×160, while the 2D heatmaps have a lower resolution of 32×40. We predict 3D positions for 13 joints: *Neck*, *Shoulders*, *Elbows*, *Wrists*, *Hips*, *Knees* and *Ankles*,and the ground truth are relative to the EgoGlass device. The 2D modules and 3D module were trained in an end-to-end manner with a learning rate of 1e-3 and 8 epochs. We used the *Adam* optimizer and *step_lr* scheduler with a step size of 4. Loss weights were set as: $\lambda_{2D} = 1$, $\lambda_b = 1$, $\lambda_{recon} = 0.001$, $\lambda_{reg} = 0.1$, and $\lambda_{cos} = 0.01$.

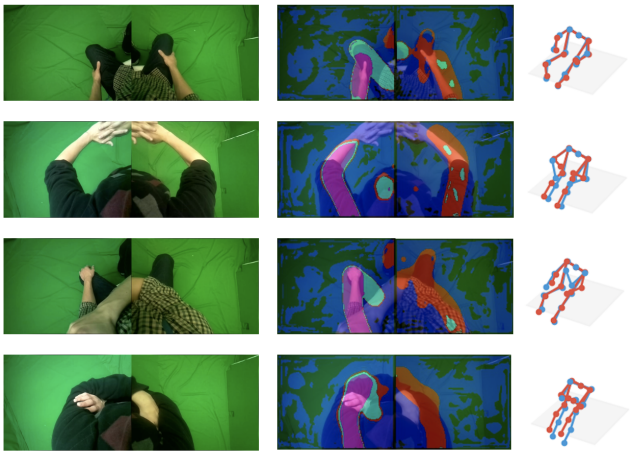


Figure 5. Qualitative results of our method on our EgoGlass dataset. The first column is the input images from two views connected side-by-side; the second column illustrates the predicted pseudo-limb mask from our body part branch for both views, where colors encode individual limbs; the last column is the visualization of 3D joint positions, skeletons in blue are ground truth and skeletons in red are predictions.

5.3. Results on our EgoGlass dataset

The pseudo-limb mask covers the visible part of the user's arms and legs. We connect the areas between joints of *Shoulder*, *Elbow*, and *Wrist* to generate the mask for one arm and the areas between the joints of *Hip*, *Knee* and *Ankle* to generate the mask for one leg.

Baseline 1: The method proposed by Tome *et al*. [29] is the state-of-the-art method for monocular egocentric human pose estimation, and their experiment shows that state-of-the-art methods for front-facing cameras, such as [14], will fail in the egocentric setup. We consider their method as a baseline to validate the necessity of information from multiple views in our setup. Due to the public unavailability of their code, we use our own implementation which achieved 43.0mm MPJPE on the xR-EgoPose dataset (their reported number is 41.0mm). Table 2 shows the quantitative comparison. The results show that we need information from both views and thus selfPose method cannot handle our multi-view setup.

Baseline 2: We use the same architecture as shown in Fig.

MPJPE(mm)	Neck	LShoulder	LElbow	LWrist	RShoulder	RElbow	RWrist	LHip	LKnee	LAnkle	RHip	RKnee	RAnkle	All
w/o body part	0.9	29.4	96.6	181.9	30.2	96.3	198.4	96.5	181.5	238.3	95.3	184.6	221	127
w/ body part	**0.0**	**27.6**	**83.3**	**164.1**	**28.9**	**77.7**	**171.5**	**71.0**	**157.3**	**201.2**	**72.2**	**147.6**	**184.9**	**106.7**
Δ(row1-row2)	0.9	1.8	13.3	17.8	1.3	18.6	26.9	25.5	24.2	37.1	23.1	37	36.1	20.3

Table 3. Quantitative results (MPJPE) and ablation study of our method with body part branch on our EgoGlass dataset. Our method with body part branch outperforms its counterpart without body part branch on all joints and average error.

P-MPJPE(mm)	Neck	LShoulder	LElbow	LWrist	RShoulder	RElbow	RWrist	LHip	LKnee	LAnkle	RHip	RKnee	RAnkle	All
w/o body part	52.8	62	74.3	145.4	53	72.5	161.6	59.9	**75.1**	**72.3**	**53.0**	81.9	**67.3**	79.3
w/ body part	**41.1**	**49.5**	**70.6**	**137.0**	**45.1**	**63.6**	**143.3**	**59.3**	76.4	73.3	57.3	**77.8**	69.2	**74.1**
Δ(row1-row2)	11.7	12.5	3.7	8.4	7.9	8.9	18.3	0.6	-1.3	-1	-4.3	-4.1	-1.9	5.2

Table 4. Quantitative results (P-MPJPE) and ablation study of our method with body part branch on our EgoGlass dataset. Our method with body part branch outperforms its counterpart without body part branch on most joints and average error.

Approach	MPJPE error (mm)
EgoCap [20]	70.0 ± 10.0
Ours	**67.9** ± 13.4

Table 5. Quantitative comparison between our method and the method in the original EgoCap paper [20]. Note that the method in EgoCap paper added temporal smooth constraint while ours does not, thus our standard deviation is expected to be slightly larger than theirs.

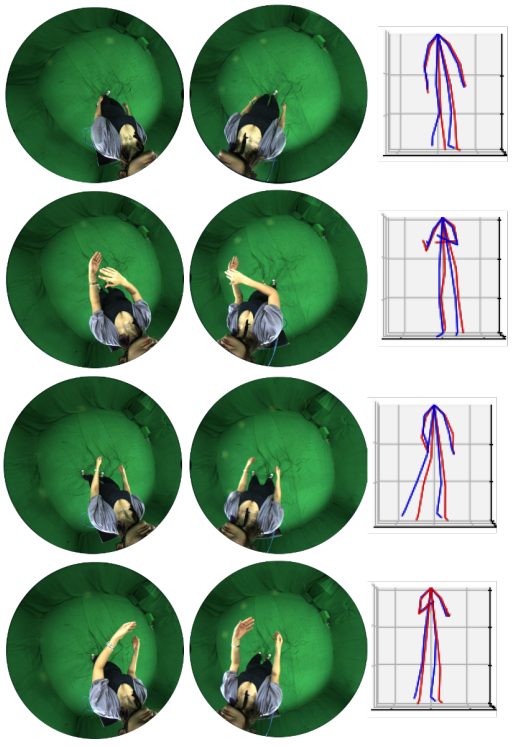


Figure 6. Qualitative results of our method on the EgoCap [20] dataset. Skeletons in blue are ground truth and skeletons in red are prediction.

4 without the body part branch in the 2D modules as our second baseline. We report the mean reconstruction errors for each joint in Table 3. Our method outperforms its counterpart without body part branch on all joints, while the improvements on lower body part are all larger than 20mm. This result aligns with the fact that the lower body part is usually more severely occluded. We also report the reconstruction error in P-MPJPE in Table 4. See Fig. 5 for visualization of the predicted pseudo-limb masks and 3D joints.

5.4. Results on EgoCap Dataset

EgoCap[20] dataset is another egocentric dataset captured by real subjects. We downloaded the training set and validation set from the website of Rhodin *et al*. [20], which contains 35285 frames from 6 subjects in the training set and 1001 frames from 1 subject in the validation set. They provide two sets of training data: one with raw green-screen background and the other augmented by replacing the original background with a random, floor-related image from Flickr. We use the first set as they only provide validation set in raw background. Furthermore, it does not provide segmentation masks so we use the pseudo-limb mask generated by our method. We use the same 17 joints as in their paper.

To the best of our knowledge, there is only one existing method which is proposed in Rhodin *et al*. [20]. Table 5 shows the comparison between our method and their method. Our method achieves better mean accuracy over all joints, while the standard deviation is slightly larger than theirs, which is expected as they imposed temporal smooth constraint but our approach does not use any temporal information. Fig. 6 shows the qualitative results of our method.

5.5. Results on xR-EgoPose dataset

We used the dataset released from xR-EgoPose [29, 30] GitHub page with 10M/10F in training set and 7M/4F in test

Approach	MPJPE error (mm)	Gaming	Gesticulating	Greeting	Lower Stretching	Patting	Reacting	Talking	Upper Stretching	Walking	All
xR-EgoPose [30]	FullBody	56.0	50.2	44.6	51.1	59.4	60.8	43.9	53.9	57.7	58.2
Zhang [35]	FullBody	36.8	34.1	36.7	50.1	57.2	34.4	32.8	54.3	52.6	50.0
SelfPose [29]	FullBody	52.5	49.2	72.0	37.3	53.0	44.4	46.1	**39.3**	**37.2**	41.0
Ours - w/o body part	LowerBody	36.8	31.8	33.3	53.1	59.2	36.8	31.0	58.7	58.0	53.4
	UpperBody	27.6	29.3	31.2	26.6	28.8	25.4	22.2	37.6	28.6	32.7
	FullBody	**32.2**	**30.5**	**32.2**	39.9	**44.0**	**31.1**	**26.6**	48.2	43.3	43.0
Ours - w/ body part	LowerBody	39.6	31.8	33.2	48.3	59.6	39.0	30.1	51.2	50.5	48.1
	UpperBody	26.0	29.2	34.2	22.6	31.9	27.3	24.0	29.0	24.2	27.3
	FullBody	32.8	**30.5**	33.7	**35.5**	45.7	33.2	27.0	40.1	37.4	**37.7**

Table 6. Quantitative comparison with existing methods on the xR-EgoPose dataset[30]. For our methods, we also report the reconstruction errors on upper and lower body. Numbers for existing methods are from respective papers as they did not release their code. Our method with body part information outperforms all other methods.

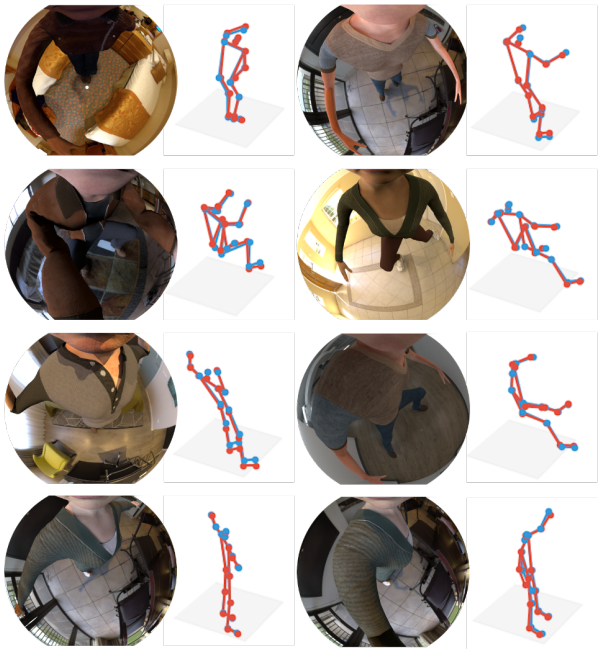


Figure 7. Qualitative results of our method on the xR-EgoPose [30] dataset. Skeletons in blue are ground truth and skeletons in red are prediction.

set and the same 16 joints as in their paper. It is a synthetic dataset and provides the segmentation mask ground truth, so we do not need to generate limb masks. As in selfPose [29], we added a limb-error term in the loss function. None of the previous methods[30, 29, 35] released their code, so we re-implemented the method in selfPose first as they have the state-of-the-art results. Our re-implementation with an added body part branch achieves 37.7mm in comparison to the result of 41.0 mm from Tome *et al.* [29] without our proposed body part branch. See Table 6 for detailed results.

The results confirm that (i) body part masks help with egocentric human pose task regardless of number of views and (ii) the pseudo-limb mask and segmentation masks perform a similar role in egocentric human pose estimation. Segmentation masks can provide accurate body part information while pseudo-limb mask is much easier to generate. Fig. 7 shows the visualization of our results.

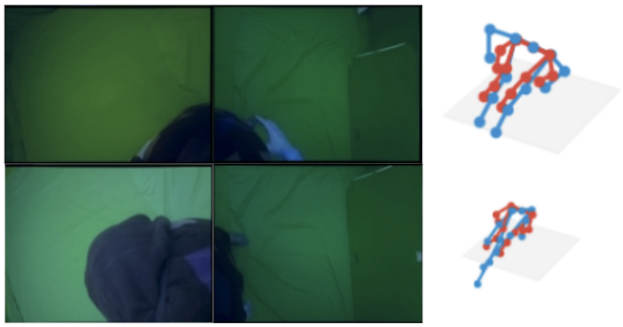


Figure 8. Failure cases when the joints are severely occluded. In the top frame, only right wrist is visible, while no visible joint in the bottom frame.

6. Conclusion and future work

We present EgoGlass, a new solution to egocentric motion-capture with a curated method for egocentric human pose estimation. We hope this eyeglass-frame-based approach can further facilitate the research in egocentric human estimation. Despite the improved accuracy at current stage, the method still suffers from some limitations such as that the estimation for lower body is generally worse than that for upper body. To improve this, utilizing more 3D information by explicitly enforcing multi-view consistency may help. Another failure case lies in the circumstances when the joints are severely occluded, see Fig. 8. Since the subjects were doing continuous motions when captured, adding temporal constraints to take advantage of visible joints in adjacent frames is a possible direction.

References

[1] 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. 4

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 4

[3] Young-Woon Cha, True Price, Zhen Wei, Xinran Lu, Nicholas Rewkowski, Rohan Chabra, Zihe Qin, Hyounghun Kim, Zhaoqi Su, Yebin Liu, et al. Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. *IEEE transactions on visualization and computer graphics*, 24(11):2993–3004, 2018. 3

[4] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7035–7043, 2017. 2

[5] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2262–2271, 2019. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[7] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7779–7788, 2020. 1

[8] Dong-Hyun Hwang, Kohei Aso, and Hideki Koike. Monoeye: Monocular fisheye camera-based 3d human pose estimation. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 988–989. IEEE, 2019. 1, 2

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 6

[10] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. 1

[11] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509. IEEE, 2017. 1, 2

[12] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014. 2

[13] Sijin Li, Weichen Zhang, and Antoni B Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2848–2856, 2015. 2

[14] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 1, 2, 6

[15] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 2

[16] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017. 2

[17] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9890–9900, 2020. 1, 2

[18] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017. 2

[19] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6040–6049, 2020. 1

[20] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 1, 2, 3, 6, 7

[21] Gregory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3108–3116. Curran Associates, 2016. 2

[22] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017. 1

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[24] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 4

[25] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 1

[26] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 1, 2

[27] Jochen Tautges, Arno Zinke, Björn Krüger, Jan Baumann, Andreas Weber, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bernd Eberhardt. Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics (ToG)*, 30(3):1–12, 2011. 2

[28] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. In *British Machine Vision Conference (BMVC)*, 2016. 2

[29] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3, 6, 7, 8

[30] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7728–7738, 2019. 1, 2, 5, 6, 7, 8

[31] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2500–2509, 2017. 2

[32] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 4

[33] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019. 1, 2

[34] KangKang Yin and Dinesh K Pai. Footsee: an interactive animation system. In *Symposium on Computer Animation*, pages 329–338, 2003. 2

[35] Yahui Zhang, Shaodi You, and Theo Gevers. Automatic calibration of the fisheye camera for egocentric 3d human pose estimation from a single image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1772–1781, 2021. 8

[36] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017. 1

[37] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016. 2