Estimating Egocentric 3D Human Pose in the Wild with External Weak Supervision

Jian Wang[1,2] Lingjie Liu[1,2] Weipeng Xu[3] Kripasindhu Sarkar[1,2]
Diogo Luvizon[1,2] Christian Theobalt[1,2]
[1]MPI Informatics [2]Saarland Informatics Campus [3]Facebook Reality Labs
{jianwang,lliu,ksarkar,theobalt}@mpi-inf.mpg.de xuweipeng@fb.com

Abstract

Egocentric 3D human pose estimation with a single fisheye camera has drawn a significant amount of attention recently. However, existing methods struggle with pose estimation from in-the-wild 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, we collect a large-scale in-the-wild egocentric dataset called Egocentric Poses in the Wild (EgoPW). *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. We present a new egocentric pose estimation method, which can be trained on the new dataset with weak external supervision. Specifically, we first generate pseudo labels for the EgoPW dataset 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 the network training, we propose 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 our 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.*

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

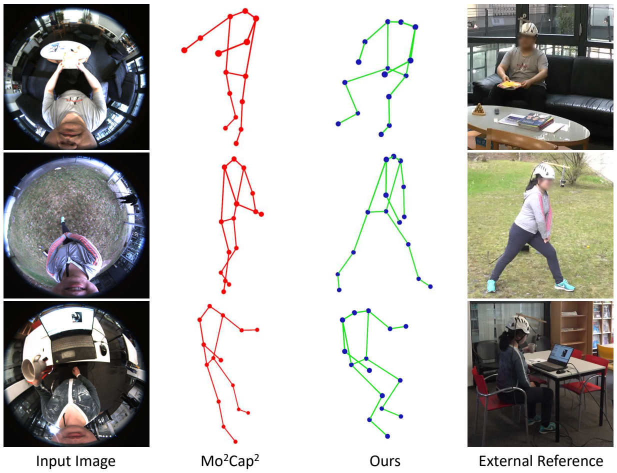

Figure 1. Compared with Mo^{2}Cap^2, our method gets a more accurate egocentric pose from a single in-the-wild image, especially when the body parts are occluded. Note that the external images are only used for visualization, not the inputs to our method.

In this work, we focus on estimating the full 3D body pose from a single head-mounted fisheye camera. The most related works are Mo^{2}Cap^2 [44] and xR-egopose [35]. 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 drop on 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. 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, we capture a large-scale in-the-wild egocentric dataset called *Egocentric Poses in the Wild (EgoPW)*. 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. 3), which is a common recording scenario. Therefore, considering a trade-off between flexibility and 3D accuracy, we use a new device setup consisting of an egocentric camera and a single auxiliary external camera. We demonstrate that 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, we propose a new egocentric pose estimation method for training on the EgoPW dataset in a weakly supervised way. Specifically, we propose a spatio-temporal optimization method 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 [44]. To improve the network performance, we facilitate the training of the egocentric pose estimation network with the extracted features from the external pose estimation network which has been trained on a large in-the-wild body pose dataset. Specifically, we enforce the feature extracted from these two views 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, we also use a synthetic dataset [44] to train the network and adopt a domain adaptation strategy to minimize the domain gap between synthetic and real data.

We evaluate our method on the test data provided by Wang *et al*. [42] and Xu *et al*. [44]. Our method significantly outperforms the state-of-the-art methods both quantitatively and qualitatively. We also show qualitative results on various in-the-wild images, demonstrating that our method can predict accurate 3D poses on very challenging scenes, especially when the body joints are seriously occluded (see our results in Fig. 1). To summarize, our contributions 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. It is publicly available in https://people.mpi-inf.mpg.de/~jianwang/projects/egopw;
- 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.

2. Related Work

Egocentric 3D full body pose estimation. Rhodin *et al*. [30] developed the first method to estimate the full-body pose from a helmet-mounted stereo fisheye camera. Cha *et al*. [4] presented an RNN-based method to estimate body pose with two pinhole cameras mounted on the head. Xu *et al*. [44] introduced a single wide-view fisheye camera setup and proposed a single-frame based egocentric motion capture system. With the same setup, Tome *et al*. [35] captured the egocentric pose with an auto-encoder network which captures the uncertainty in the predicted heatmaps. In order to further mitigate the effect of image distortions, Zhang *et al*. [46] proposed an automatic calibration module. Hwang *et al*. [14] put an ultra-wide fisheye camera on the user's chest and estimate body pose, camera rotation and head pose simultaneously. Jiang *et al*. [16] mounted a front-looking fisheye camera on the user's head and estimated the body and head pose by leveraging the motion of the environment and extremity of the human body. Wang *et al*. [42] proposed an optimization algorithm to obtain temporally stable egocentric poses with motion prior learned from Mocap datasets. However, these methods are all trained on synthetic datasets, thus suffering from the performance drop on the real images due to the domain gap and lack of external supervision. Our method, on the contrary, achieves better performance on the in-the-wild scenes.

Pseudo label generation. The task of pseudo-labeling [20,34,45] is a semi-supervised learning technique that generates pseudo labels for unlabeled data and uses the generated labels to train a new model. This has been applied in the areas of segmentation [8, 22, 47, 48], pose estimation [2,21,23,25] and image classification [1,13,29]. Since the pseudo labels may be inaccurate, some methods have been proposed to filter inaccurate labels to increase the labeling stability. Shi *et al*. [34] set confidence levels on unlabeled samples by measuring the sample density. Chen *et al*. [5] enforced the stability of pseudo labels by adopting an easy-to-hard transfer strategy. Wang and Wu [41] introduced a repetitive re-prediction strategy to update the pseudo labels, while Rizve *et al*. [32] proposed an uncertainty-aware pseudo-label selection framework that selects pseudo labels. Morerio *et al*. [24] used a conditional

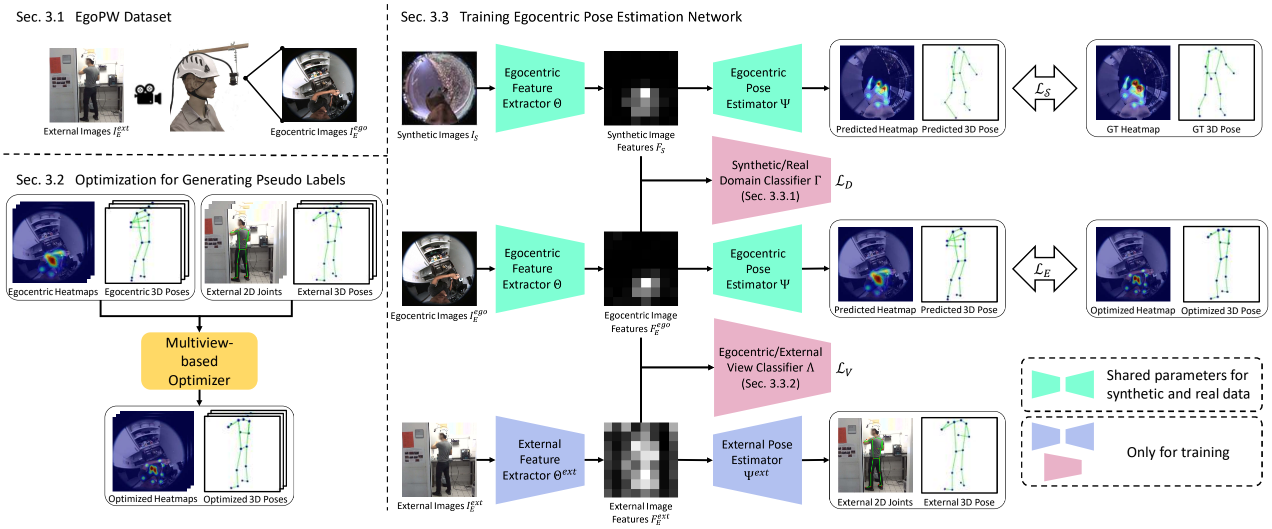
Sec. 3.1 EgoPW Dataset
External Images I_E^{ext}
Egocentric Images I_E^{ego}
Sec. 3.2 Optimization for Generating Pseudo Labels
Egocentric Heatmaps
Egocentric 3D Poses
External 2D Joints
External 3D Poses
Multiview-based Optimizer
Optimized Heatmaps
Optimized 3D Poses
Sec. 3.3 Training Egocentric Pose 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 (Sec. 3.3.1)
$\mathcal{L}_D$
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 (Sec. 3.3.2)
$\mathcal{L}_V$
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 2. Overview of our method. We first collected the new EgoPW dataset (Sec. 3.1), where pseudo labels are generated by a multi-view based optimization method (Sec. 3.2). We then train our model with the proposed framework (Sec. 3.3), where the network is simultaneously trained with EgoPW datasets and synthetic data from Mo^{2}Cap^2. We enforce the egocentric network to learn a better feature representation from the external view (Sec. 3.3.2) and bridge the gap between synthetic and real data with a domain classifier (Sec. 3.3.1).

GAN to filter the noise in the pseudo labels. Different from previous pseudo-labeling works which generate the labels from network predictions or clustering, we design an optimization framework to generate labels with supervision from egocentric and external views simultaneously.

Weakly Supervised 3D Human Pose Estimation. Recently, there is 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 [27, 40] leverages the non-rigid SFM to get 3D joint positions from 2D keypoint annotations in unconstrained images. Some works [6, 7, 10, 28, 38] present an unsupervised learning approach to train the 3D pose estimation network with the supervision from 2D reprojections. The closest to our work are the approaches of [15, 19, 31, 39] in that they leverage the weak supervision from multi-view images for training. Iqbal *et al.* [15] and Rhodin *et al.* [31] supervise the network training process by calculating the differences between Procrustes-aligned 3D poses from different views. Wandt *et al.* [39] 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.* [19] obtain the pseudo labels with epipolar geometry between different views and use the pseudo labels to train the 3D pose lifting network. Different from previous works [15, 19, 31, 39], our method uses spatio-temporal optimization framework that takes egocentric and external view 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.

3. Method

We propose 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 our approach is illustrated in Fig. 2. We first capture a large-scale egocentric in-the-wild dataset, called *EgoPW*, which contains synchronized egocentric and external image sequences (Sec. 3.1). Next, we generate pseudo labels for the EgoPW dataset 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. 3.2). Next, we train the egocentric pose estimation network on the synthetic data from Mo^{2}Cap^2 [44] and on the EgoPW dataset with pseudo labels $\mathcal{P}_{seq}^{ego}$. In the training process, we leverage the feature representation from an on-the-shelf external pose estimation network [43] to enforce our egocentric network to learn a better feature representation in an adversarial way (Sec. 3.3.2). We also use an adversarial domain adaptation strategy to mitigate the domain gap between synthetic and real datasets (Sec. 3.3.1).

3.1. EgoPW Dataset

We first describe 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 an IRB approval. We generate 3D poses as pseudo labels using the egocentric and external images, which will be elaborated later. In terms of size, our EgoPW dataset is larger than existing in-the-wild 3D pose estimation datasets, like 3DPW [37], and has similar scale to the existing synthetic egocentric datasets, including the Mo^{2}Cap^2 [44] and the xR-egopose [35] datasets.

3.2. Optimization for Generating Pseudo Labels

In this section, we present an optimization method based on [42] to generate pseudo labels for EgoPW. Given a sequence, we split it into segments containing B consecutive frames. For the egocentric frames I^{ego}_{seq}, we estimate the 3D poses represented by 15 joint locations in the coordinate system of the egocentric camera (called "egocentric poses") $\widetilde{\mathcal{P}}^{ego}_{seq} = \{\widetilde{\mathcal{P}}^{ego}_1, \dots, \widetilde{\mathcal{P}}^{ego}_B\}$, $\widetilde{\mathcal{P}}^{ego}_i \in \mathbb{R}^{15\times 3}$, and 2D heatmaps $H^{ego}_{seq} = \{H^{ego}_1, \dots, H^{ego}_B\}$ using the Mo^{2}Cap^2 method [44]. Aside from egocentric poses, we also estimate the transformation matrix between the egocentric camera poses of two adjacent frames $[R^{SLAM}_{seq} \mid t^{SLAM}_{seq}] = \{[R^2_1 \mid t^2_1], \dots, [R^B_{B-1} \mid t^B_{B-1}]\}$ using ORB-SLAM2 [26]. For the external frames I^{ext}_{seq}, we estimate the 3D poses (called "external poses") $\mathcal{P}^{ext}_{seq} = \{\mathcal{P}^{ext}_1, \dots, \mathcal{P}^{ext}_B\}$, $\mathcal{P}^{ext}_i \in \mathbb{R}^{15\times 3}$ using VIBE [18] and 2D joints $\mathcal{J}^{ext}_{seq} = \{\mathcal{J}^{ext}_1, \dots, \mathcal{J}^{ext}_B\}$, $\mathcal{J}^{ext}_i \in \mathbb{R}^{15\times 2}$ using openpose [3].

Next, following [42], we learn a latent space to encode an egocentric motion prior with a sequential VAE which consists of a CNN-based encoder f_{enc} and decoder f_{dec}. We then optimize the egocentric pose by finding a latent vector z such that the corresponding pose sequence $P^{ego}_{seq} = f_{dec}(z)$ minimizes the objective function:

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

In this objective function, E^{ego}_R, E^{ego}_J,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 [42]. E^{ext}_R, E^{ext}_J, 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 supplemental material for a detailed definition of each term.

Note that since the relative pose between external camera and egocentric camera is unknown, we also need to optimize the relative egocentric camera pose with respect to the external camera pose 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, we designed the external reprojection term which minimizes the difference between the projected 3D pose with the external 2D joints. The energy term is defined as:

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

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. 2, we first project the egocentric body pose $\mathcal{P}^{ego}_i$ to the 2D body pose in the external view with the egocentric camera pose $[R_i \mid t_i]$ and the intrinsic matrix K, and then compare the projected body pose with the 2D joints estimated by the openpose [3]. Since the relative pose between the external camera and egocentric camera are unknown at the beginning of the optimization, we optimize the egocentric camera pose $[R_i \mid t_i]$ simultaneously while optimizing the egocentric body pose $\mathcal{P}^{ego}_{seq}$. In order to make the optimization process converge faster, we initialize the egocentric camera pose $[R_i \mid t_i]$ with the Perspective-n-Point algorithm [11].

Camera Pose Consistency. We cannot get the accurate 3D pose 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, we introduce the camera consistency term E_C as follows:

$$\begin{aligned} E_C(R_{seq}, t_{seq}) = \sum_{i=1}^{B-1} \Bigg\| & \begin{bmatrix} R_i & t_i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R^{i+1}_i & t^{i+1}_i \\ 0 & 1 \end{bmatrix} \\ & - \begin{bmatrix} R_{i+1} & t_{i+1} \\ 0 & 1 \end{bmatrix} \Bigg\|_2 , \end{aligned} \tag{3}$$

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, we also use the external 3D body poses to supervise the optimization of the egocentric 3D body pose. We define the external 3D pose term which measures the difference between the external and the egocentric body poses after a rigid alignment:

$$E_J(\mathcal{P}^{ego}_{seq}, \mathcal{P}^{ext}_{seq}) = \sum_{i=1}^{B} \left\| \mathcal{P}^{ext}_i - \left[R^{pa}_i \mid t^{pa}_i\right] \mathcal{P}^{ego}_i \right\|^2_2 , \tag{4}$$

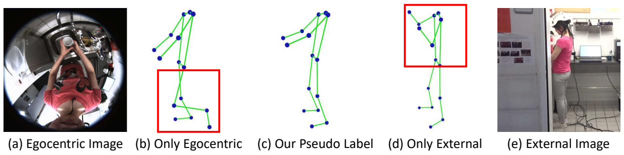

Figure 3. Our pseudo label generation method combines the information from both the egocentric view and external view, therefore leading to more accurate pseudo labels (c). Only with the egocentric camera, the feet cannot be observed and well-tracked (b). Only with the external camera, the hands are occluded and result in the wrong result on the hand part (d).

where $[R_i^{pa} \mid t_i^{pa}]$ 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, we can reconstruct more accurate pseudo labels. As shown in Fig. 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. 3, b), however, the hands can be clearly seen and tracked in the egocentric view (Fig. 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. 3, b), but can be easily viewed and tracked in the external view (Fig. 3, d). By joining the information from both views, we can successfully predict accurate 3D poses as the pseudo labels (Fig. 3, c). We note that the external camera is only used for generating the pseudo labels but at test time, only the egocentric camera is used.

Camera Matrix Regularization. We constrain the camera rotation matrix R_i to be orthogonal:

$$E_J(R_{seq}) = \sum_{i=1}^{B} \left\| R_i^T R_i - I \right\|_2^2. \tag{5}$$

Different from previous single-view pose estimation methods which leverages the weak supervision from multiple views [15, 19, 31, 39], our 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.

3.3. Training Egocentric Pose Estimation Network

Through the optimization framework in Sec. 3.2, we can get accurate 3D pose pseudo labels $\mathcal{P}_{seq}^{ego}$ for each egocentric frame in the EgoPW dataset, which is further processed into the 2D heatmap H_E and the distance between joints and egocentric camera D_E with the fisheye camera model [33] described in supplementary materials.

Afterward, we train a single-image based egocentric pose estimation network on both the synthetic dataset from Mo^{2}Cap^2 and the EgoPW dataset, as shown in the right part of Fig. 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, we note the synthetic dataset $S = \{I_S, H_S, D_S\}$ including synthetic images I_S along with their corresponding heatmaps H_S and distance labels D_S from Mo^{2}Cap^2 dataset, and the EgoPW dataset $E = \{I_E^{ego}, H_E, D_E, I_E^{ext}\}$ 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, we train the egocentric pose estimation network 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:

$$\begin{aligned} L_S &= \text{mse}(\hat{H}_S, H_S) + \text{mse}(\hat{D}_S, D_S) \\ L_E &= \text{mse}(\hat{H}_E, H_E) + \text{mse}(\hat{D}_E, D_E), \end{aligned} \tag{6}$$

where

$$\begin{aligned} \hat{H}_S, \hat{D}_S &= \Psi(F_S), F_S = \Theta(I_S); \\ \hat{H}_E, \hat{D}_E &= \Psi(F_E^{ego}), F_E^{ego} = \Theta(I_E^{ego}). \end{aligned} \tag{7}$$

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.

3.3.1 Adversarial Domain Adaptation

To bridge the domain gap between the synthetic and real data domains, following Tzeng *et al.* [36], we introduce 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:

$$\mathcal{L}_D = -E[\log(\Gamma(F_S))] - E[\log(1 - \Gamma(F_E^{ego}))]. \tag{8}$$

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.

3.3.2 Supervising Egocentric Feature Representation with External View

Although our new 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 further improve the generalizability of our network and prevent overfitting to the training identities, we propose to supervise our egocentric representation by leveraging the high-quality third-person-view features. From a transfer learning perspective, although following Mo^{2}Cap^2 [44], our 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, we use the adversarial training strategy to align the feature representation from egocentric and external networks. Specifically, we use an adversarial discriminator Λ which takes the feature vectors extracted from an egocentric image and the corresponding in-the-wild images and predicts if the feature is from egocentric or external images. The adversarial discriminator Λ is trained with a cross-entropy loss:

$$L_V = -E[\log(\Lambda(F_E^{ego}))] - E[\log(1 - \Lambda(F_E^{ext}))], \quad (9)$$

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.*'s work [43] and keep 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 [9], we use the output feature of the 4th res-block of ResNet-50 network [12] 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, we use an average pooling layer 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 suppl. mat. for further details.

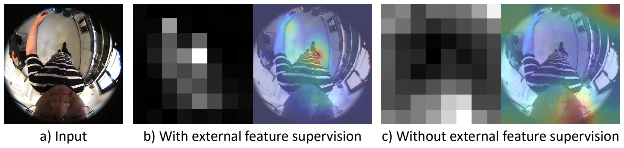
a) Input
b) With external feature supervision
c) Without external feature supervision

Figure 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.

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. 4.

4. Experiments

4.1. Datasets

We quantitatively evaluate our finetuned network on the real-world dataset from Mo^{2}Cap^2 [44] and Wang *et al.* [42]. The real-world dataset in Mo^{2}Cap^2 [44] contains 2.7k frames of two people captured in indoor and outdoor scenes, and that in Wang *et al.* [42] contains 12k frames of two people captured in the studio. To measure the accuracy of our pseudo labels, we evaluate our optimization method (Sec. 3.2) only on the dataset from Wang *et al.* [42] since the Mo^{2}Cap^2 dataset does not include the external view.

To evaluate our method on the in-the-wild data, we also conduct a qualitative evaluation on the test set of the EgoPW dataset. The EgoPW dataset will be made publicly available, and more details and comparisons to other datasets are included in the supplementary materials.

4.2. Evaluation Metrics

We measure the results of our method as well as other baseline methods with two metrics, PA-MPJPE and BA-MPJPE, which estimate the accuracy of a single body pose. For **PA-MPJPE**, we rigidly align the estimated pose $\hat{\mathcal{P}}$ of each frame to the ground truth pose $\mathcal{P}$ using Procrustes analysis [17]. In order to eliminate the influence of the body scale, we also report the **BA-MPJPE** scores. In this metric, we first resize the bone length of each predicted body pose $\hat{\mathcal{P}}$ and ground truth body pose $\mathcal{P}$ to the bone length of a standard skeleton. Then, we calculate the PA-MPJPE between the two resulting poses.

4.3. Pseudo Label Generation

In this paper, we first generate the pseudo labels with the optimization framework (Sec. 3.2) and use them to train our

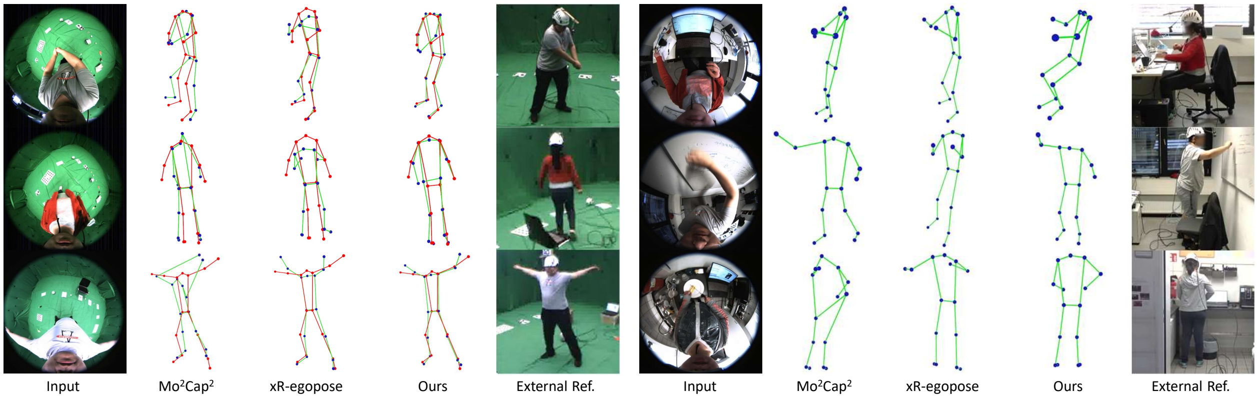

Figure 5. Qualitative comparison between our method and the state-of-the-art methods. From left to right: input image, Mo^{2}Cap^2 result, xR-egopose result, our result, and external image. The ground truth pose is shown in red. Note that the external images are not used during inference. The input images in the left part are from the test dataset in [42], while those in the right part come from EgoPW test sequences.

Method	PA-MPJPE	BA-MPJPE
Mo^{2}Cap^2	102.3	74.46
xR-egopose	112.0	87.20
Wang *et al.* [42]	83.40	63.88
VIBE [18]	68.13	52.99
Our Optimizer	**57.19**	**46.14**

Table 1. The accuracy of pseudo labels on Wang *et al.*'s dataset. Utilizing both egocentric and external view, the body poses from our optimization method (Sec. 3.2) are more accurate and can serve as better pseudo labels.

Method	PA-MPJPE	BA-MPJPE
Wang *et al.*'s test dataset		
Rhodin *et al.* [31]	89.67	73.56
Mo^{2}Cap^2 [44]	102.3	74.46
xR-egopose [35]	112.0	87.20
Ours	**81.71**	**64.87**
Mo^{2}Cap^2 test dataset		
Rhodin *et al.* [31]	97.69	76.92
Mo^{2}Cap^2 [44]	91.16	70.75
xR-egopose [35]	86.85	66.54
Ours	**83.17**	**64.33**

Table 2. Performance of our egocentric pose estimation network (Sec. 3.3) on Wang *et al.*'s test dataset and Mo^{2}Cap^2 test dataset [44]. Our method outperforms the state-of-the-art methods, Mo^{2}Cap^2 [44] and xR-egopose [35], on both metrics.

network (Sec. 3.3). Thus, pseudo labels with higher accuracy generally lead to better network performance. In this experiment, we evaluate the accuracy of pseudo labels on Wang *et al.*'s dataset and show the results in Table 1. This table shows that our method outperforms all the baseline methods by leveraging both the egocentric view and external view during optimization. Note that though compared in Table 1, we cannot use any external-view based pose estimation method, *e.g.* VIBE [18] and 3DPW [37], for training the egocentric pose estimation network. This is because the 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 our optimization approach, the method in [42] performs worse due to the lack of external-view supervision.

4.4. Comparisons on 3D Pose Estimation

In this section, we compare the egocentric pose estimation network trained in Sec. 3.3 with previous single-frame-based methods on the test dataset from [42] under the "Wang *et al.*'s test dataset" in Table 2. Since the code or the predictions of xR-egopose are not publicly available, we use our reimplementation of xR-egopose instead. On this dataset, our method outperforms Mo^{2}Cap^2 by 20.1% and xR-egopose by 27.0% respectively. We also compared with previous methods on the Mo^{2}Cap^2 test dataset and show the results under the "Mo^{2}Cap^2 test dataset" in Table 2. On the Mo^{2}Cap^2 test dataset, our method performs better than Mo^{2}Cap^2 and xR-egopose by 8.8% and 4.2%, respectively.

From the results in Table 2, we can see 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 the supplementary material. For the qualitative comparison, we show the results of our method on the studio dataset and in-the-wild dataset in Fig. 5. Our 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 the supplementary materials for more qualitative results.

We also compared our method with Rhodin *et al.*'s method [31], which uses the weak supervision from multiple views to supervise the training of a single view pose estimation network. In our EgoPW dataset, we only have one egocentric and one external view. Thus, we fix the 3D pose estimation network for the external view and only train the egocentric pose estimation network. Following Rhodin *et al*. [31], we align the prediction from the egocentric and external view with Procrustes analysis and calculate the loss proposed by Rhodin *et al*. Our result in Table 2 shows our method performs better. This is mainly because our spatio-temporal optimization method predicts accurate and temporally stable 3D poses as pseudo labels, while other methods suffer from inaccurate egocentric pose estimations.

4.5. Ablation Study

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 3. The quantitative results of ablation study.

Supervision from the external view. In our work, we introduce the external view 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, we firstly generate the 3D poses as pseudo labels with Wang *et al*.'s method, *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 3. We also show the qualitative results with and without external-view supervision in Fig. 6. Both the qualitative and quantitative results demonstrate that with the external supervision, the performance of our pose estimation network is significantly better especially on occluded cases.

Learning egocentric feature representation and bridging the domain gap with adversarial training. In our work, we train the pose estimation network with two adversarial components in order to learn the feature representation of the egocentric human body (Sec. 3.3.2) and bridge the domain gap between synthetic and real images (Sec. 3.3.1). In order to demonstrate the effectiveness of both modules, we removed the domain classifier Λ in our training process and show the results in the row of "w/o learning representation" in Table 3. We also removed the domain classifier Γ, train the network without L_D and show the quantitative results in the row of "w/o domain adaptation" in Table 3. After moving any of the two components, our method suffers from the performance drop, which demonstrates the effectiveness of both the feature representation learning module and the domain adaptation module.

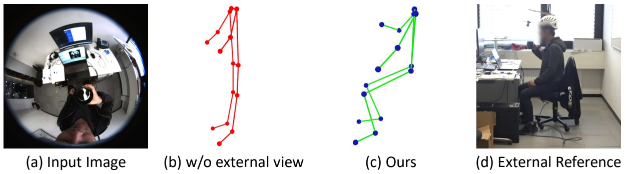

Figure 6. The results of our method with (c) and without external view (b). The network cannot predict accurate poses for the occluded cases without the external view supervision. The external view is only for visualization and not used for predicting the pose.

Comparison with only using unsupervised domain adaptation. In this experiment, we compare our approach with the unsupervised adversarial domain adaptation method [36] which is commonly used for transfer learning tasks. We train the network only with the L_S and L_D in the adversarial domain adaptation module (Sec. 3.3.1) and show the results in the "Unsupervised DA" of the Table 3. Our approach outperforms the unsupervised domain adaptation method due to our high-quality pseudo labels.

5. Conclusions

In this paper, we have proposed a new approach to egocentric human pose estimation with a single head-mounted fisheye camera. We collected a new in-the-wild egocentric dataset (EgoPW) and designed a new optimization method to generate accurate egocentric poses as pseudo labels. Next, we supervise the egocentric pose estimation network with the pseudo labels and the features from the external network. The experiments show that our method outperforms all of the state-of-the-art methods both qualitatively and quantitatively and our method also works well under severe occlusion. As future work, we would like to develop a video-based method for estimating temporally-consistent egocentric poses from an in-the-wild video.

Limitations. The accuracy of pseudo labels in our method is constrained by our in-the-wild capture system, which only contains one egocentric view and one external view, and further constrains the performance of our network. One future solution is to fuse different sensors, including IMUs and depth cameras, for capturing the in-the-wild dataset.

Acknowledgments Jian Wang, Kripasindhu Sarkar and Christian Theobalt have been supported by the ERC Consolidator Grant 4DReply (770784) and Lingjie Liu has been supported by Lise Meitner Postdoctoral Fellowship.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 2

[2] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9498–9507, 2019. 2

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

[4] Y. Cha, T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, A. Ilie, A. State, Z. Xu, J. Frahm, and H. Fuchs. 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. 2

[5] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019. 2

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

[7] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. 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*, pages 10895–10904, 2019. 3

[8] Zhenghao Chen, Rui Zhang, Gang Zhang, Zhenhuan Ma, and Tao Lei. Digging into pseudo label: a low-budget approach for semi-supervised semantic segmentation. *IEEE Access*, 8:41830–41837, 2020. 2

[9] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1831–1840, 2017. 6

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

[11] X. Gao, Xiaorong Hou, Jianliang Tang, and H. Cheng. Complete solution classification for the perspective-three-point problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25:930–943, 2003. 4

[12] 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. 6

[13] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: Similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2021. 2

[14] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera. In *ACM Symposium on User Interface Software & Technology*, page 98–111, 2020. 2

[15] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. 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*, pages 5243–5252, 2020. 3, 5

[16] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. *ICCV*, 2021. 2

[17] David G Kendall. A survey of the statistical theory of shape. *Statistical Science*, pages 87–99, 1989. 6

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5252–5262, 2020. 4, 7

[19] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019. 3, 5

[20] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2

[21] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1491, 2021. 2

[22] Yuexiang Li, Jiawei Chen, Xinpeng Xie, Kai Ma, and Yefeng Zheng. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 614–623. Springer, 2020. 2

[23] Katja Ludwig, Sebastian Scherer, Moritz Einfalt, and Rainer Lienhart. Self-supervised learning for human pose estimation in sports. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. 2

[24] Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, and Vittorio Murino. Generative pseudo-label refinement for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3130–3139, 2020. 2

[25] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12386–12395, 2020. 2

[26] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 4

[27] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7688–7697, 2019. 3

[28] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 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*, pages 7753–7762, 2019. 3

[29] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 2

[30] 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 Trans. Graph.*, 35(6):162:1–162:11, 2016. 2

[31] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018. 3, 5, 7, 8

[32] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 2

[33] Davide Scaramuzza and Katsushi Ikeuchi. Omnidirectional camera. 2014. 5

[34] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 299–315, 2018. 2

[35] Denis Tomè, Patrick Peluse, Lourdes Agapito, and Hernán Badino. xr-egopose: Egocentric 3d human pose from an HMD camera. In *IEEE International Conference on Computer Vision*, pages 7727–7737, 2019. 1, 2, 4, 7

[36] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 5, 8

[37] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 4, 7

[38] Bastian Wandt and Bodo Rosenhahn. 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*, pages 7782–7791, 2019. 3

[39] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2021. 3, 5

[40] Chaoyang Wang, Chen Kong, and Simon Lucey. Distill knowledge from nrsfm for weakly supervised 3d pose learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 743–752, 2019. 3

[41] Guo-Hua Wang and Jianxin Wu. Repetitive reprediction deep decipher for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6170–6177, 2020. 2

[42] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. *ICCV*, 2021. 2, 4, 6, 7

[43] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018. 3, 6

[44] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, 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 Trans. Vis. Comput. Graph.*, 25(5):2093–2101, 2019. 1, 2, 3, 4, 6, 7

[45] Xibei Yang, Shaochen Liang, Hualong Yu, Shang Gao, and Yuhua Qian. Pseudo-label neighborhood rough set: measures and attribute reductions. *International journal of approximate reasoning*, 105:112–129, 2019. 2

[46] 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. 2

[47] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021. 2

[48] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 2