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

Jian Wang\textsuperscript{1,2} Lingjie Liu\textsuperscript{1,2} Weipeng Xu\textsuperscript{3} Kripasindhu Sarkar\textsuperscript{1,2} Diogo Luvizon\textsuperscript{1,2} Christian Theobalt\textsuperscript{1,2} \\
\textsuperscript{1}MPI Informatics \hspace{1cm} \textsuperscript{2}Saarland Informatics Campus \hspace{1cm} \textsuperscript{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 \textsuperscript{x}R.

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\textsuperscript{2}Cap\textsuperscript{2} [44] and \textsuperscript{x}R-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\textsuperscript{2}Cap\textsuperscript{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 \textit{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 \cite{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 \cite{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 \textit{et al.} \cite{42} and Xu \textit{et al.} \cite{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. We will make it publicly available;
- 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

\textbf{Egocentric 3D full body pose estimation.} Rhodin \textit{et al.} \cite{30} developed the first method to estimate the full-body pose from a helmet-mounted stereo fisheye camera. Cha \textit{et al.} \cite{4} presented an RNN-based method to estimate body pose with two pinhole cameras mounted on the head. Xu \textit{et al.} \cite{44} introduced a single wide-view fisheye camera setup and proposed a single-frame based egocentric motion capture system. With the same setup, Tome \textit{et al.} \cite{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 \textit{et al.} \cite{46} proposed an automatic calibration module. Hwang \textit{et al.} \cite{14} put an ultra-wide fisheye camera on the user’s chest and estimate body pose, camera rotation and head pose simultaneously. Jiang \textit{et al.} \cite{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 \textit{et al.} \cite{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.

\textbf{Pseudo label generation.} The task of pseudo-labeling \cite{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 \cite{8,22,47,48}, pose estimation \cite{2,21,23,25} and image classification \cite{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 \textit{et al.} \cite{34} set confidence levels on unlabeled samples by measuring the sample density. Chen \textit{et al.} \cite{5} enforced the stability of pseudo labels by adopting an easy-to-hard transfer strategy. Wang and Wu \cite{41} introduced a repetitive re-prediction strategy to update the pseudo labels, while Rizve \textit{et al.} \cite{32} proposed an uncertainty-aware pseudo-label selection framework that selects pseudo labels. Morerio \textit{et al.} \cite{24} used a conditional 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] leverage 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 $T_{ego}^{seq} = \{T_{ego}^1, \ldots, T_{ego}^B\}$ and external images $T_{ext}^{seq} = \{T_{ext}^1, \ldots, T_{ext}^B\}$ and outputs egocentric 3D poses $P_{ego}^{seq} = \{P_{ego}^1, \ldots, P_{ego}^B\}$ as the pseudo labels (Sec. 3.2). Next, we train the egocentric pose estimation network on the synthetic data from Mo2Cap$^2$ [44] and on the EgoPW dataset with pseudo labels $P_{seq}$. 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, including reading magazine/newspaper, playing board games, doing a presentation, walking, sitting down, using a computer, calling on the phone, drinking water, writing on the paper, writing on the whiteboard, making tea, cutting vegetables, stretching, running, playing table tennis, playing baseball, climbing floors, dancing, opening the door, and waving hands. 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 Mo2Cap2 [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 \( P_{\text{ego}}^{\text{seq}} \), we estimate the 3D poses represented by 15 joint locations in the coordinate system of the egocentric camera (called “egocentric poses”) \( \mathcal{P}_{\text{seq}}^{\text{ego}} = \{ P_1^{\text{ego}}, \ldots, P_B^{\text{ego}} \} \), \( P_i^{\text{ego}} \in \mathbb{R}^{15 \times 3} \), and 2D heatmaps \( H_{\text{seq}}^{\text{ego}} = \{ H_1^{\text{ego}}, \ldots, H_B^{\text{ego}} \} \) using the Mo2Cap2 method [44]. Aside from egocentric poses, we also estimate the transformation matrix between the egocentric camera poses of two adjacent frames \( [ R_{\text{SLAM}}^{\text{seq}} | t_{\text{SLAM}}^{\text{seq}} ] = \{ [ R_1^2 | t_1^2 ] , \ldots , [ R_B^2 | t_B^2 ] \} \) using ORB-SLAM2 [26]. For the external frames \( P_{\text{ext}}^{\text{seq}} \), we estimate the 3D poses (called “external poses”) \( \mathcal{P}_{\text{seq}}^{\text{ext}} = \{ P_1^{\text{ext}}, \ldots, P_B^{\text{ext}} \} \), \( P_i^{\text{ext}} \in \mathbb{R}^{15 \times 3} \) using VIBE [18] and 2D joints \( J_{\text{seq}}^{\text{ext}} = \{ J_1^{\text{ext}}, \ldots, J_B^{\text{ext}} \} \).

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 \( f_{\text{enc}} \) and decoder \( f_{\text{dec}} \). We then optimize the egocentric pose by finding a latent vector \( z \) such that the corresponding pose sequence \( P_{\text{seq}}^{\text{ego}} = f_{\text{dec}}(z) \) minimizes the objective function:

\[
E(P_{\text{seq}}^{\text{ego}}, R_{\text{seq}}, t_{\text{seq}}) = \lambda_R^{\text{ego}} E_R^{\text{ego}} + \lambda^{\text{ext}} E_R^{\text{ext}} + \lambda_J^{\text{ego}} E_J^{\text{ego}} + \lambda_J^{\text{ext}} E_J^{\text{ext}} + \lambda_T E_T + \lambda_B E_B + \lambda_C E_C + \lambda_M E_M.
\]

In this objective function, \( E_R^{\text{ego}}, E_J^{\text{ego}}, E_T, \) and \( E_B \) are egocentric reprojection term, egocentric pose regularization term, motion smoothness regularization term and bone length regularization term, which are the same as those defined in [42]. \( E_R^{\text{ext}}, E_J^{\text{ext}}, E_C, \) and \( E_M \) are the external reprojection term, external 3D body pose regularization term, camera pose consistency term, and camera matrix regularization term, which will be described later. Please see the 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 \( R_{\text{seq}} = R_1, \ldots, R_B \) and translations \( t_{\text{seq}} = t_1, \ldots, t_B \).

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:

\[
E_R^{\text{ext}}(P_{\text{seq}}^{\text{ego}}, R_{\text{seq}}, t_{\text{seq}}) = \sum_{i=1}^{B} || J_i^{\text{ext}} - K [ R_i | t_i ] P_i^{\text{ego}} ||^2_2,
\]

where \( K \) is the intrinsic matrix of the external camera; \( [ R_i | t_i ] \) 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 \( P_i^{\text{ego}} \) to the 2D body pose in the external view with the egocentric camera pose \( [ R_i | t_i ] \) 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 \( [ R_i | t_i ] \) simultaneously while optimizing the egocentric body pose \( P_{\text{seq}}^{\text{ego}} \). In order to make the optimization process converge faster, we initialize the egocentric camera pose \( [ R_i | t_i ] \) 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:

\[
E_C(R_{\text{seq}}, t_{\text{seq}}) = \sum_{i=1}^{B-1} || \begin{bmatrix} R_i & t_i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_{i+1} & t_{i+1} \\ 0 & 1 \end{bmatrix} ||_2^n - \begin{bmatrix} R_{i+1} & t_{i+1} \end{bmatrix} ||_2^n.
\]

It enforces the egocentric camera pose at \( (i + 1) \) th frame \( [ R_{i+1} | t_{i+1} ] \) to be consistent with the pose obtained by transforming the egocentric camera pose at the \( i \) th frame \( [ R_i | t_i ] \) 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
3.3. Training Egocentric Pose Estimation Network

Body poses after a rigid alignment:

\[ E_J(P_\text{ego}^{seq}, P_\text{ext}^{seq}) = \sum_{i=1}^{B} \left\| P_\text{ext}^i - [R_i^{\text{pa}} | t_i^{\text{pa}}] P_\text{ego}^{\text{ego0}} \right\|^2_2, \tag{4} \]

where \([R_i^{\text{pa}} | t_i^{\text{pa}}]\) is the transformation matrix calculated with Procrustes analysis, which rigidly aligns the external 3D pose estimation \(P_\text{ext}^i\) and the egocentric 3D pose \(P_\text{ego}^{\text{ego0}}\).

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 leverage 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.

3.3.1 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 \(\Gamma\) 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 \(\Gamma\) is trained with a cross-entropy loss:

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

Once the discriminator \(\Gamma\) has been trained, the feature extractor \(\Theta\) maps the images from different domains to the same feature space such that the classifier \(\Gamma\) cannot tell if the
features are extracted from synthetic images or real images. Therefore, the pose estimator \( \Psi \) 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 \( F^E_{\Psi} \) and external features \( F^E_{\Theta} \) 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 \( \Lambda \) 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 \( \Lambda \) is trained with a cross-entropy loss:

\[
L_V = -E[\log(\Lambda(F^E_{\Psi}))] - E[\log(1 - \Lambda(F^E_{\Theta}))],
\]

where \( F^E_{\Psi} = \Theta^E_{\Psi}(I^E_{\Psi}) \) and \( \Theta^E_{\Theta}(I^E_{\Theta}) \) 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 \( \Theta^E_{\Psi} \) and the pose estimator \( \Psi^E_{\Theta} \) 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 \( \Lambda \). Furthermore, the spatial position of the joints is quite different in the egocentric view and the external view, which will make the discriminator \( \Lambda \) easily learn the difference between egocentric and external features. To solve this, we use an average pooling layer in the discriminator \( \Lambda \) 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.

During the training process, the egocentric pose estimation network is trained to produce the features \( F^E_{\Psi} \) to fool the domain classifier \( \Lambda \) 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 \( \hat{P} \) of each frame to the ground truth pose \( P \) 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 \( \hat{P} \) and ground truth body pose \( P \) to the bone length of a
standard skeleton. Then, we calculate the PA-MPJPE between the two resulting poses.

### 4.3. Pseudo Label Generation

| Method                  | PA-MPJPE | BA-MPJPE |
|-------------------------|----------|----------|
| Mo²Cap²                 | 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 performance of our optimization framework on Wang et al.’s dataset. Our optimization method outperforms all previous egocentric pose estimation methods and the state-of-the-art pose estimation method of the external view.

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

We compare our approach 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 Mo²Cap² by 20.1% and xR-egopose by 27.0% respectively. We also compared with previous methods on the Mo²Cap² test dataset and show the results under the “Mo²Cap² test dataset” in Table 2. On the Mo²Cap² test dataset, our method performs better than Mo²Cap² 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 Mo²Cap² 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 leverages 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 that the performance of our method is better. This is mainly because our spatio-temporal optimization method predicts accurate and temporally stable 3D poses as pseudo labels, while other methods suffer seriously 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 on the real EgoPW dataset. The external view enables
Figure 5. Qualitative comparison between our method and the state-of-the-art methods. From left to right: input image, Mo2Cap 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.

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.

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 with 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 $\Lambda$ 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 $\Gamma$, 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.

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 method. To solve this, we can fuse different sensors, including IMUs and depth cameras, for capturing the in-the-wild dataset, which can be explored in our follow-up work.

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] 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.

[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] 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.

[5] Chaoqi Chen, Weiping Xie, Wenbiao 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.

[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.

[7] Xijun Cheng, Kwan Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-supervised discovery of geometry-aware representations for 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10895–10904, 2019.

[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.

[9] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan Lu, 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.

[10] Dylan Drover, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuc 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.

[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.

[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.

[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.

[14] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Kois. 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.

[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.

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

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

[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.

[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.

[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.

[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.

[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.

[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.

[24] Pietro Morerio, Riccardo Volpi, Ruggiero Ragonesi, and Vittorio Murino. Generative pseudo-label refinement for
