UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Multi-person 3D Pose estimation from a single image captured by a fisheye camera

Zhang, Y.; You, S.; Karaoglu, S.; Gevers, T.

DOI
10.1016/j.cviu.2022.103505

Publication date
2022

Document Version
Final published version

Published in
Computer Vision and Image Understanding

License
CC BY

Link to publication

Citation for published version (APA):
Zhang, Y., You, S., Karaoglu, S., & Gevers, T. (2022). Multi-person 3D Pose estimation from a single image captured by a fisheye camera. *Computer Vision and Image Understanding*, *222*, Article 103505. https://doi.org/10.1016/j.cviu.2022.103505

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Nov 2026

Computer Vision and Image Understanding 222 (2022) 103505

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Multi-person 3D pose estimation from a single image captured by a fisheye camera

Yahui Zhang [a,*], Shaodi You [a], Sezer Karaoglu [a,b], Theo Gevers [a,b]

[a] University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands
[b] 3DUniversum, Science Park 400, 1098XH Amsterdam, The Netherlands

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
41A05
41A10
65D05
65D17

Keywords:
Multi-person 3D pose estimation
Fisheye camera
Distortion
Dataset

ABSTRACT

Multi-person 3D pose estimation with absolute depths for a fisheye camera is a challenging task but with valuable applications in daily life, especially for video surveillance. However, to the best of our knowledge, such problem has not been explored so far, leaving a gap in practical applications. In this work, we first propose a method for multi-person 3D pose estimation from a single image taken by a fisheye camera. Our method consists of two branches to estimate absolute 3D human poses: (1) a 2D-to-3D lifting module to predict root-relative 3D human poses (HPoseNet); (2) a root regression module to estimate absolute root locations in the camera coordinate (HRootNet). Finally, we propose a fisheye re-projection module without using ground-truth camera parameters to connect two branches, alleviating the impact of image distortions on 3D pose estimation and further regularizing prediction absolute 3D poses. Experimental results demonstrate that our method achieves the state-of-the-art performance on two public multi-person 3D pose datasets with synthetic fisheye images and our newly collected dataset with real fisheye images. The code and new dataset will be made publicly available.

1. Introduction

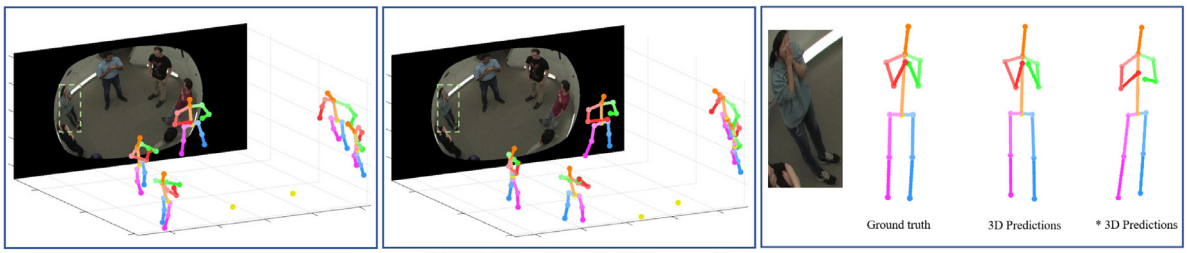
Due to the wide angle, fisheye cameras have been widely used in various practical scenarios such as video surveillance (Kim et al., 2016), virtual reality (Rhodin et al., 2016) and automotive applications (Hughes et al., 2009). Particularly, fisheye cameras will have larger field of view with larger distortion parameters. Many of these applications require the inference of multi-person 3D poses from fisheye images. However, this task has not been studied, and most existing methods focus on 3D pose estimation from images captured by a perspective camera (Moon et al., 2019a; Rogez et al., 2019; Guo et al., 2021; Cheng et al., 2021).

To this end, we aim to compute multi-person 3D poses from a single image taken by a fisheye camera. This is the first approach, to the best of our knowledge, to perform this task. To achieve this, there are three major challenges: ***i***) humans at different distances from the center of images exhibit varying scales and distortions, due to image distortions. Although different methods (Kanazawa et al., 2018; Habibie et al., 2019; Ci et al., 2019; Kolotouros et al., 2019) use a re-projection method to establish a relationship between 2D and 3D poses with predicted scale and translation parameters, the aim is to estimate root-relative 3D human poses, ignoring absolute location information. Pelvises are usually defined as root joints. However, humans at different positions suffer from varying distortion strengths in this task. Therefore, such kind of methods are expected to fail to solve this challenge. ***ii***) This task is complicated that the distance between humans and cameras is not fixed. Recent methods (Tome et al., 2019; Tome et al., 2020; Xu et al., 2019; Wang et al., 2021) predict the egocentric 3D pose from images captured by a fisheye camera installed on a human head/baseball cap. In their settings, the head/neck joints are seen as the root located at the same position on the image. Therefore, the negative impact of image distortions can be avoided by relative joint locations to the root in a learning based manner with one level of image distortions. ***iii***) We intend to predict 3D human joint locations with absolute depths, which is more challenging than root-relative 3D pose estimation because of the inherent depth and scale ambiguity. Recently, some researchers (Li et al., 2020; Lin and Lee, 2020; Moon et al., 2019a; Zhen et al., 2020) focus on the estimation of absolute joint locations from a single image taken by a perspective camera. However, we argue that it is a strong prior to use ground-truth camera parameters for evaluation.

In this paper, we propose a novel top-down approach to multi-person 3D pose estimation from a single image captured by a fisheye camera. The proposed framework consists of two branches, *i.e.,* HPoseNet and HRootNet, to estimate root-relative 3D poses and absolute depths of root joints, respectively. To alleviate the impact of human scales changes caused by unknown distortions, a re-projection

* Corresponding author.
E-mail address: y.zhang5@uva.nl (Y. Zhang).

https://doi.org/10.1016/j.cviu.2022.103505
Received 16 November 2021; Received in revised form 1 July 2022; Accepted 5 July 2022
Available online 9 July 2022
1077-3142/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).



(a) 3D Predictions (Ours). (b) *3D Predictions (Ours without full components). (c) Results on humans with large distortions from another viewpoint.

Fig. 1. 3D pose predictions using our approach. * indicates our method without full components: (1) re-projection module and (2) global and local feature fusion. Given a fisheye image shown in background, our method with full components generates more reasonable 3D poses.

module is proposed to connect the two branches to enforce projected absolute 3D poses consistent with 2D ground truths under image distortions. In this way, our approach takes image distortions into account to estimate multi-person 3D poses and predicted absolute depths are further regularized. Particularly, we adopt a learning based approach to estimate camera parameters circumventing the requirement of ground-truth camera parameters. Fig. 1 shows the predictions generated by our method.

We evaluate the proposed approach on two public datasets including CMU Panoptic (Joo et al., 2015) and Shelf (Belagiannis et al., 2014) datasets. Particularly, we synthetically add different levels of image distortions to public datasets. To test the performance on real fisheye images, we collected a dataset — 3DhUman recorded by two fisheye cameras with 3 persons performing three commonly activities: posing, talking and walking. As ceiling cameras (*e.g.,* video surveillance) are commonly used, we focus on this scenario, *i.e.,* the top-down viewpoint. Our approach outperforms existing methods on both synthesized and real-world datasets.

In summary, the contributions of this work are:

- We propose a top-down method for multi-person 3D pose estimation from a single image taken by a fisheye camera. To the best of our knowledge, this is the first approach to perform this task.
- A re-projection module is proposed to alleviate the effect of image distortion on multi-person 3D pose estimation. Particularly, camera parameters are predicted by our framework instead of using the ground truth.
- Our method significantly outperforms existing state-of-the-art methods on public datasets with synthetic fisheye images and our proposed dataset with real fisheye images.

2. Related work

Multi-person 2D pose estimation. Existing work for multi-person 2D pose estimation can be divided into bottom-up and top-down approaches. Bottom-up approaches (Cao et al., 2017; Hidalgo et al., 2019; Jin et al., 2019; Kocabas et al., 2018; Newell et al., 2017; Nie et al., 2019) simultaneously detect all human joints and then collect them for each person. Top-down approaches (Chen et al., 2018; Fang et al., 2017; Moon et al., 2019b; Papandreou et al., 2017; Sun et al., 2019; Xiao et al., 2018) first employ a detector to predict bounding boxes of humans and then estimate a single 2D human pose from the cropped images.

Multi-person 3D pose estimation. There are many methods (Dabral et al., 2019; Mehta et al., 2018; Rogez et al., 2017, 2019; Zanfir et al., 2018a,b) for multi-person 3D pose estimation. However, most of them require a post-processing step, *i.e.,* an optimization strategy by minimizing the error between projected 3D poses and 2D poses (Dabral et al., 2019; Rogez et al., 2017, 2019) or correspondences between semantic representations (Zanfir et al., 2018a) to obtain absolute joint locations in real spaces. Recently, some methods (Li et al., 2020; Lin and Lee, 2020; Moon et al., 2019a; Zhen et al., 2020) adopt the learning based manner to obtain absolute depths of root joints. Moon et al. (2019a) introduce a novel depth measure combined with a correction factor to obtain the real depth. They rely on the area of the bounding box of humans in image and real spaces. Lin and Lee (2020) consider the depth regression problem as a classification problem to perform depth estimation and localization of root joints. These methods follow the top-down pipeline in which pose estimation is performed from cropped images, and hence ignoring the global information.

Note that recent methods (Li et al., 2020; Lin and Lee, 2020; Moon et al., 2019a; Zhen et al., 2020; Guo et al., 2021; Chen et al., 2022) compute 3D poses according to 2D poses in pixel coordinates and depths in camera coordinates. They assume that intrinsic camera parameters are known both in training and testing procedures. On the other hand, existing methods mainly focus on pose estimation from a perspective camera or multi-view perspective images (Wu et al., 2021; Dong et al., 2021; Lin and Lee, 2021). No research exists on multi-person 3D pose estimation from a single image captured by a fisheye camera.

3D pose estimation from a fisheye camera. There are few works on 3D human pose estimation under fisheye cameras placed on the chest (Jiang and Grauman, 2017; Hwang et al., 2020) or head (Rhodin et al., 2016; Tome et al., 2019; Xu et al., 2019). Recently, Xu et al. (2019) take original and auxiliary images that focus on the lower body as inputs to improve the performance of egocentric pose estimation. Tome et al. (2019) and Tome et al. (2020) propose a method that includes two branches for 2D and 3D pose regression to estimate egocentric 3D poses. Further, Cho et al. (2021) propose an optimization-based method for 3D human pose estimation from a third-person viewpoint to deal with the image distortion problem without camera calibration. However, these methods are based on the root-relative single-person 3D pose estimation where the camera is placed fixedly on the human head for egocentric 3D pose estimation.

3. Multi-person 3D pose estimation from fisheye cameras

The goal of our method is to estimate multi-person 3D joint locations with absolute depths in camera coordinates from a single image captured by a fisheye camera. Here, two issues need to be solved: the negative impact of images distortions and usage of global information.

3.1. Issues on image distortions

Due to the existence of image distortions, persons at different locations on images may cause varying distortion strengths. Therefore, even when persons express different 2D poses, they may be originated from the same 3D pose (please see supplementary material for more analysis). This makes multi-person 3D human pose estimation more challenging when camera parameters are not provided (known). In this paper, we propose a re-projection module based on the fisheye camera model to alleviate the effect of image distortions on multi-person 3D pose estimation.

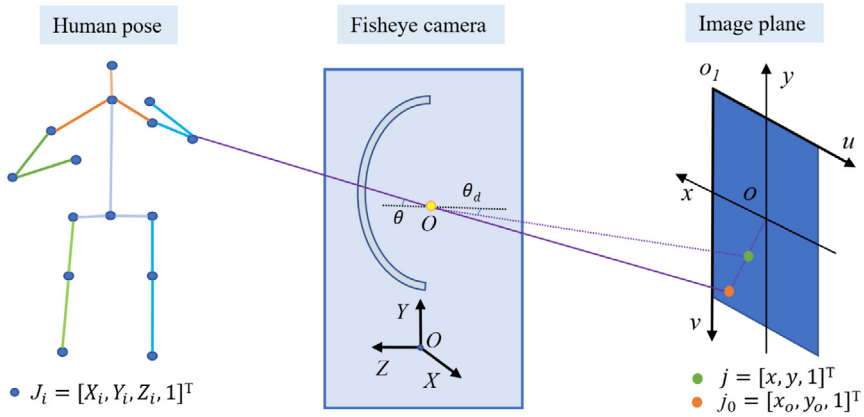


Fig. 2. The process of 3D-to-2D projection for the fisheye camera model. This figure consists of a 3D human pose represented by a set of joints in camera coordinates $OXYZ$, a fisheye camera, and 2D projections on the image plane o_1uv. The angle of refraction θ is decreased to θ_d.

3.1.1. Fisheye camera model

Fig. 2 shows the process of 3D-to-2D projection for the fisheye camera model. Specifically, the 3D human pose is represented by a set of scatter joints, a 4 by n matrix $\boldsymbol{J}_i^{abs} = [X_i^{abs}, Y_i^{abs}, Z_i^{abs}, 1]^T$, in camera coordinates $OXYZ$. After going through the fisheye camera, the angle of refraction θ is decreased to θ_d, and the 2D projections $\boldsymbol{j}_o = [x_o, y_o, 1]^T$ are changed to $\boldsymbol{j} = [x, y, 1]^T$. Particularly, $\boldsymbol{j}_o$ is the 2D projection based on the perspective camera, *i.e.,* without image distortions.

3.1.2. 3D pose estimation from a fisheye camera

To reduce the negative impact of image distortions, we first use a 2D-to-3D lifting module to obtain 3D human joint locations, and then minimize the error between projected 3D predictions and 2D ground truths. This enforces estimated 3D poses to be consistent with corresponding 2D poses under possible distortions. Since the relative depth of human joints is comparable to the distance from humans to cameras, we use perspective projection to calculate 2D projections. Therefore, estimated depths can be regularized.

Let $\mathbf{P}_{3\mathrm{Dabs}} = [\boldsymbol{J}_1, \boldsymbol{J}_2, \ldots, \boldsymbol{J}_n]$ represent human joint locations in camera coordinates $OXYZ$, where n indicates the number of human joints and $\boldsymbol{J}_i = [X_i, Y_i, Z_i, 1]^T$. Particularly, $\mathbf{P}_{3\mathrm{Drel}}$ denotes the root-relative human joint locations. Pelvises are defined as the root joint in this work. 2D projections $\mathbf{p}_{2\mathrm{D}}$ and $\mathbf{p}_{o2\mathrm{D}}$, a 3 by n matrix with $\boldsymbol{j}_i = [x_i, y_i, 1]^T$ and $\boldsymbol{j}_{oi} = [x_{oi}, y_{oi}, 1]^T$, are based on the perspective and fisheye camera model, respectively. With intrinsic and extrinsic camera parameters ($\boldsymbol{K}$, $\boldsymbol{R}$ and $\boldsymbol{T}$), 2D projections $\mathbf{p}_{o2\mathrm{D}}$ under the perspective camera are obtained by:

$$s \cdot \mathbf{p}_{o2\mathrm{D}} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{3\mathrm{Dabs}}, \tag{1}$$

where s is a scale factor and is equal to the value of Z in $\mathbf{P}_{3\mathrm{Dabs}}$. Because $\mathbf{P}_{3\mathrm{Dabs}}$ are in the camera coordinate, the extrinsic camera parameters $\boldsymbol{R}$ and $\boldsymbol{T}$ are the identity matrix.

In terms of fisheye cameras, there are distortion parameters to change the 3D-to-2D projection in Eq. (1). Specifically, Eq. (1) is modified by adding a distortion matrix $\boldsymbol{D}$:

$$s \cdot \mathbf{p}_{2\mathrm{D}} = \boldsymbol{K}\boldsymbol{D}\boldsymbol{I}_{3\times4}\mathbf{P}_{3\mathrm{Dabs}}, \tag{2}$$

where $\boldsymbol{I}_{3\times4}$ is a 3 by 4 identity matrix, and $\boldsymbol{D}$, in this paper, is defined as:

$$\boldsymbol{D} = \begin{bmatrix} \theta_d/l & 0 & 0 \\ 0 & \theta_d/l & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where $l = \frac{\sqrt{X^2+Y^2}}{Z}$. Following previous works (Kannala and Brandt, 2004; Van den Heuvel et al., 2007), the angle of refraction $\theta_d = \theta(1 + k_1\theta^2 + k_2\theta^4 + k_3\theta^6 + \cdots)$, where $\theta = \arctan(l)$, and two of distortion parameters (k_1, k_2) are used for simplification.

3.1.3. Automatic calibration for a fisheye camera

To avoid the need of ground-truth camera parameters, we adopt a learning based approach to estimate camera parameters during training stages. Specifically, five camera parameters are predicted: focal length (f), principal coordinates (c_x, c_y) and distortion parameters (k_1, k_2). To optimize the process of automatic calibration, we minimize the absolute error between absolute 3D joint locations $\mathbf{P}_{3\mathrm{Dabs}}$ and 2D ground truths $\mathbf{p}_{2\mathrm{D}}^{\mathrm{GT}}$.

$$\underset{f,c_x,c_y,k_1,k_2}{\arg\min} \left\| \boldsymbol{K}\boldsymbol{D}\boldsymbol{I}_{3\times4}\mathbf{P}_{3\mathrm{Dabs}} - \mathbf{p}_{2\mathrm{D}}^{\mathrm{GT}} \right\|_1. \tag{4}$$

3.2. Issues on global information

Most existing top-down approaches estimate multi-person 3D poses from a cropped image around humans, ignoring the global relation of each person. We propose to aggregate features from cropped images around humans and the whole image in the latent space to maintain the global information for the estimation of absolute depths and camera parameters.

Inspired by Lin and Lee (2020) and Zhao et al. (2019), features extracted from input images contributes to human pose estimation. However, these features may also contain background, appearance or other useless information to our task. To enhance the role of features contributing to the performance, we employ an attention mechanism to facilitate the process of human pose estimation.

4. Network and training details

We adopt a top-down pipeline to estimate multi-person 3D poses with absolute depths as shown in Fig. 3. Our framework consists of three components including HPoseNet, HRootNet, and a re-projection module. In this section, we provide details of each component and training details.

HPoseNet. HPoseNet is to estimate root-relative joint locations for each person. Following Zhang et al. (2021), the design of HPoseNet is shown in Fig. 4. HPoseNet takes ResNet50 as backbone followed by three deconvolutional layers to estimate 2D poses using heatmap representations. Then, two residual fully connected layers are used to predict root-relative 3D joint locations. To optimize HPoseNet, we minimize the mean square error (MSE) between 1) estimated 2D heatmaps $\mathbf{HM}$ and ground-truth heatmaps $\mathbf{HM}^{\mathrm{GT}}$, which represents the 2D poses in fisheye images; 2) estimated root-relative 3D pose $\mathbf{P}_{3\mathrm{Drel}}$ and ground-truth 3D pose $\mathbf{P}_{3\mathrm{Drel}}^{\mathrm{GT}}$:

$$\begin{aligned} \mathcal{L}_{HM} &= \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbf{HM}_{(i)} - \mathbf{HM}_{(i)}^{\mathrm{GT}} \right\|_2, \\ \mathcal{L}_{3D} &= \frac{1}{n}\sum_{i=1}^{n} \left\| \mathbf{P}_{3\mathrm{Drel}(i)} - \mathbf{P}_{3\mathrm{Drel}(i)}^{\mathrm{GT}} \right\|_2, \end{aligned} \tag{5}$$

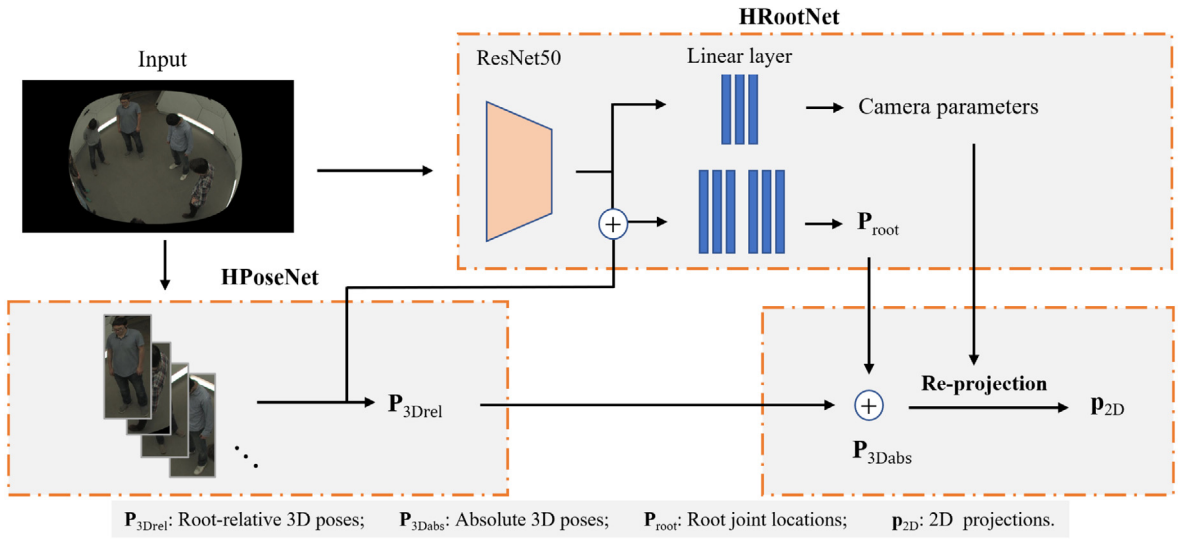


Fig. 3. Overview of the proposed framework. There are two branches, *i.e.,* HPoseNet and HRootNet, to estimate root-relative 3D poses and absolute depths of root joints. Finally, we use a re-projection module to connect the two branches, enforcing the estimated 3D human poses to be consistent with the 2D poses under distortions by minimizing the re-projection error.

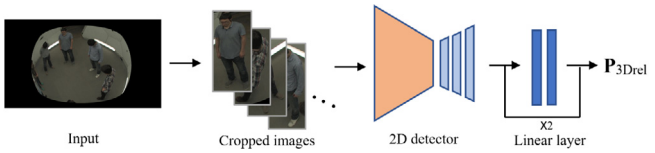


Fig. 4. The design of our HPoseNet. HPoseNet takes images cropped by human bounding boxes as inputs to estimate root-relative 3D human poses.

where i denotes the joint index and n indicates the number of human joints.

HRootNet. We aim to regress absolute root joint locations in camera coordinates and camera parameters. In this branch, ResNet50 is used as backbone to extract latent features from input images. Then, we combine the features from *i*) the entire input image and *ii*) the cropped image around the person to estimate the root joint locations. SENet (Hu et al., 2018) is used to apply the attention mechanism to the extracted features from the cropped images to exploit the meaningful representations in latent space. In addition, we use linear layers to regress camera parameters instead of using the ground truth. To train HRootNet, we optimize the absolute error between estimated root joint locations $\mathbf{P}_{\text{root}}$, and MSE between the ground truth $\mathbf{P}^{\text{GT}}_{\text{root}}$. The loss function is given by:

$$\mathcal{L}_{root} = \left\| \mathbf{P}_{\text{root}} - \mathbf{P}^{\text{GT}}_{\text{root}} \right\|_2. \tag{6}$$

Re-projection module. We propose a re-projection module to connect the two branches. Combining $\mathbf{P}_{\text{3Drel}}$ from HPoseNet and $\mathbf{P}_{\text{root}}$ from HRootNet, absolute 3D joint locations are obtained by $\mathbf{P}_{\text{3Dabs}} = \mathbf{P}_{\text{3Drel}} + \mathbf{P}_{\text{root}}$. To alleviate the negative influence of image distortions and further regularize predicted 3D poses with absolute depths, we propose a re-projection module to project estimated absolute 3D poses onto 2D poses using predicted camera parameters. Then, projected absolute 3D poses $\mathbf{P}_{\text{3Dabs}}$ are forced to be consistent with 2D ground truths $\mathbf{p}^{\text{GT}}_{\text{2D}}$ under distortions. In this way, our approach takes image distortions into account to estimate multi-person 3D poses, reducing the impact of human scales changes caused by unknown distortions. The loss function is as follows:

$$\mathcal{L}_{rep} = \frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{K}\boldsymbol{D}\boldsymbol{I}_{3\times4}\mathbf{P}_{\text{3Dabs}(i)} - \mathbf{p}^{\text{GT}}_{\text{2D}(i)} \right\|_1. \tag{7}$$

4.1. Training

According to Eq. (5)–(7), the overall loss function is given by:

$$\mathcal{L}_{pose} = \lambda_{HM}\mathcal{L}_{HM} + \mathcal{L}_{3Drel} + \lambda_{rep}\mathcal{L}_{rep} + \lambda_{root}\mathcal{L}_{root}, \tag{8}$$

where λ_{HM}, λ_{rep} and λ_{root} are loss weights to obtain a trade-off between each loss.

5. Experiments

5.1. Experimental setup

Current datasets. We use CMU Panoptic (Joo et al., 2015) and Shelf (Belagiannis et al., 2014) datasets for evaluation. Specifically, two views in CMU Panoptic dataset are chosen from HD camera 2 and 19, since these two cameras provide top-down viewpoints which are similar to video surveillance in real world scenarios. For Shelf dataset, we use all views to train and test our method. Since these datasets are created for perspective cameras, we synthetically add image distortions according to Eq. (2). Specifically, distortion parameters k_1 and k_2 are uniformly sampled, where $k_1 \in [-0.9600, -0.7000], k_2 \in [-0.0500, -0.0100]$ in CMU Panoptic dataset and $k_1 \in [-1.500, -1.0000]$, $k_2 \in$ [-0.7000, −0.1000] in Shelf dataset.

Proposed dataset. We collected a new multi-person 3D pose dataset — 3DhUman, captured by two fisheye cameras with grayscale images in an indoor environment. Specifically, images are captured by two fisheye cameras with different camera parameters from two top-down viewpoints. Two lidar cameras are used to capture depth information. The dataset contains 3 participants (2 males and 1 female) performing 3 activities: posing, talking and walking, as shown in Fig. 5. 2D/3D annotations and camera parameters are given by this dataset. Following Joo et al. (2015), 15 joints are included in the annotations.

The dataset consists of 217 fisheye images. Training and testing sets are split by whether the images include the specific participant. Specifically, the images including that participant are taken as the training set, while the remaining images are used as the testing set. For training, we used the (cropped) images containing that participant for root-relative 3D pose estimation, while the entire images are used for the camera parameters and absolute depth estimation of that participant combined with cropped images as inputs. Both training and testing sets include three activities. Since the 3DhUman dataset consists

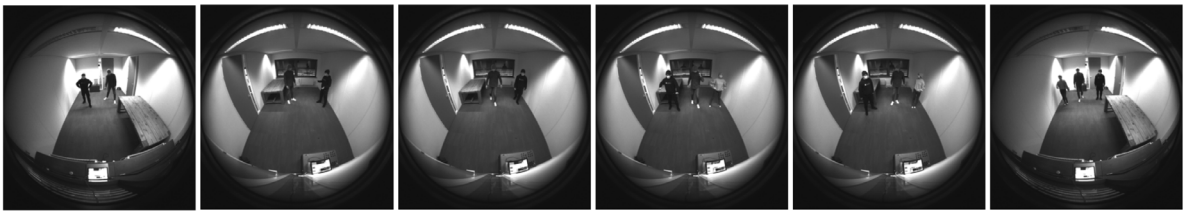


Fig. 5. Example images in 3DhUman dataset. Actions from left to right: posing, talking and walking with 2 persons (first 3 pictures) and 3 persons (last 3 pictures).

Table 1
The MPJPE of root-relative 3D poses and MRPE of absolute root joint locations on modified CMU Panoptic (top) and Shelf (bottom) datasets.

Methods	Haggling	Mafia	Ultimatum	Pizza	MPJPE ↓	MRPE ↓
Moon et al. (2019a) (ICCV'19)	100.26[a]	96.79[a]	99.88[a]	125.09[a]	102.83[a]	783.42
Lin and Lee (2020) (ECCV'20)	100.26[a]	96.79[a]	99.88[a]	125.09[a]	102.83[a]	367.47
Ours	**79.98**	**55.26**	**79.12**	**80.26**	**66.76**	**182.94**

Methods	MPJPE ↓	MRPE ↓
Moon et al.	300.18[a]	696.10
Lin et al.	300.18[a]	793.11
Ours	**132.45**	**589.19**

[a]As Moon et al. (2019a) and Lin and Lee (2020) use the same architecture for root-relative 3D human estimation and Lin and Lee (2020) does not release code for this part, the values of MPJPE are considered to be the same.

of three participants, we employed a 3-fold cross-validation to evaluate the methods.

Metrics. The Mean Per Joint Position Error (MPJPE) is used as the metric for root-relative 3D human poses, while the mean of the root position error (MRPE) (Moon et al., 2019a) is used to evaluate root joint locations.

Implementation details. We first pre-train HPoseNet on the MPII 2D pose dataset, and then the whole network is trained on the 3D pose dataset for 10 epochs with an initial learning rate of 5×10^{-4} with a decay over 8 epochs. Adam is used for optimization. The batch size is set to 32. Loss weights are set to $\lambda_{HM} = 10^7$, $\lambda_{rep} = 1$ and $\lambda_{root} = 0.05$.

Method comparison. To evaluate the proposed method, a comparison is given between two existing methods (Lin and Lee, 2020; Moon et al., 2019a). For a fair comparison, we re-train two models on the modified CMU Panoptic and Shelf datasets following their settings. Since the code has not been released, we will not compare our approach with Tome et al. (2019) and Xu et al. (2019).

Following existing approaches (Dabral et al., 2019; Li et al., 2020; Lin and Lee, 2020; Moon et al., 2019a), we first attempt to use Mask R-CNN (He et al., 2017) to detect each person in the input image. However, it fails to detect accurate bounding boxes for each person. To avoid the influence of the person detector, ground-truth bounding boxes are used for evaluation.

5.2. Results and ablations

5.2.1. Overall performance

Modified CMU Panoptic dataset. We first compare our approach with existing state-of-the art methods (Lin and Lee, 2020; Moon et al., 2019a) on the modified CMU Panoptic dataset. Table 1 (top) lists experimental results including the MPJPE of four activities and MRPE. It is shown that our approach achieves the best performance and obtain an improvement of 35.08% than existing methods with MPJPE. Particularly, our approach performs best over all activities. For MRPE, our approach significantly outperforms compared methods with an improvement of 50.22% than Lin and Lee (2020). Moon et al. (2019a) estimate absolute root joint locations based on the area of bounding boxes around humans in image and real spaces under the perspective camera. However, image distortions in this topic change the scale of each person on the image plane. Therefore, it is expected that Moon et al. (2019a) fail to achieve desirable performance.

Table 2
MPJPE and MRPE on the 3DhUman dataset.

Methods	Posing	Talking	Walking	MPJPE ↓	MRPE ↓
Moon et al.	79.44*	61.57*	70.63*	73.29*	1536.24
Lin et al.	79.44*	61.57*	70.63*	73.29*	1661.02
Ours	**67.87**	**53.56**	**56.95**	**62.14**	**177.95**

Modified Shelf dataset. We then test all approaches on the modified Shelf dataset. Table 1 (bottom) shows that our method outperforms existing methods with an improvement of 55.88% for MPJPE. Further, the proposed method shows best performance on root joint estimation compared to two existing methods. Since the Shelf dataset includes less training data than the CMU Panoptic dataset, the performance of all three methods is degraded.

3DhUman dataset. We conduct experiments on real fisheye images, *i.e.,* 3DhUman dataset. All methods are first trained on modified CMU Panoptic dataset with grayscales and then finetuned on 3DhUman dataset. Table 2 lists experimental results with metrics of MPJPE and MRPE, and our method achieves the best performance on root-relative 3D human pose and absolute root joint estimation. Particularly, it seems that Moon et al. and Lin et al. do not generalize well to real fisheye images for root joint estimation.

5.2.2. Performance on perspective images

We compare our HPoseNet with Moon et al. (2019a) and Lin and Lee (2020) for 3D human pose estimation on perspective images. HPoseNet estimates root-relative 3D human poses. Therefore, Human3.6 m dataset, a large-scale dataset and a commonly used benchmark for 3D human pose estimation, is used to evaluate all methods. Following the same setting as in (Moon et al., 2019a; Lin and Lee, 2020), we select subjects S1, S5, S6, S7, and S8 for training and S9 and S11 for testing. During the training procedure, $\lambda_{rep} = 0$ and $\lambda_{root} = 0$. Table 3 reports the experimental results for MPJPE. Despite not using ground-truth camera parameters, our method still achieves similar performance of root-relative 3D human poses.

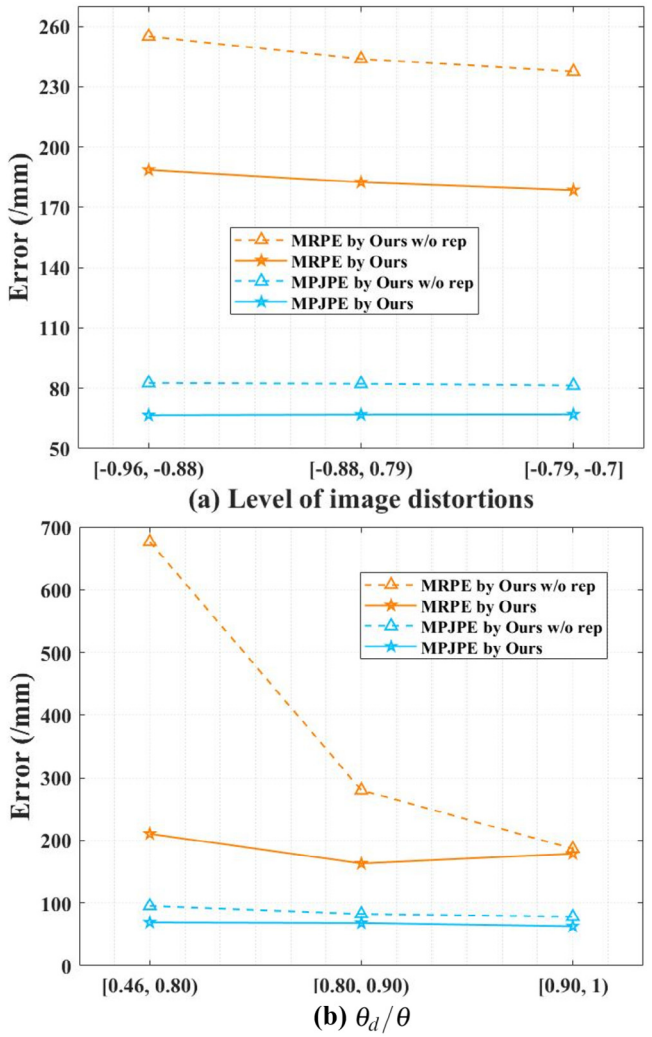


Fig. 6. Analysis of the sensitivity of our method and our method without using proposed re-projection module (Ours w/o rep) on (a) image level defined by image distortions and (b) instance level defined by θ_d/θ for each human appeared in images on the modified CMU Panoptic dataset.

Table 3
MPJPE on the Human3.6 m dataset (Ionescu et al., 2013).

Methods	MPJPE ↓
Moon et al.	53.3*
Lin et al.	53.3*
Ours	**54.1**

5.2.3. Ablation study

We perform an ablative study to validate the effectiveness of proposed contributions: local and global feature fusion and re-projection module on the modified CMU Panoptic dataset. Therefore, we take our method combining the 2D heatmap loss ($\mathcal{L}_{HM}$) and 3D loss ($\mathcal{L}_{3Drel}$ and $\mathcal{L}_{root}$) as the baseline, which is the common setting in single/multi-person 3D pose estimation (Zhou et al., 2017; Habibie et al., 2019; Lin and Lee, 2020). The experimental results are listed in Table 4.

From Table 4, our method with full components achieves the best performance in both metrics of MPJPE and MRPE on modified CMU Panoptic and 3DhUman datasets. For modified CMU Panoptic dataset, the performance achieved by our method without re-projection module drops by 22.86% and 33.98% in the metric of MPJPE and MRPE, respectively. On the other hand, our full method improves the performance on MPJPE and MRPE by 10.53% and 9.56% compared with our method without feature fusion. The results of our method on the 3DhUman dataset show a similar trend. Experimental results demonstrate that both components of our method contribute to the overall improvement.

Table 4
Ablation study on the modified CMU Panoptic and 3DhUman datasets.

Methods	Modified CMU Panoptic		3DhUman	
	MPJPE ↓	MRPE ↓	MPJPE ↓	MRPE ↓
Baseline ($\mathcal{L}_{HM}$ + 3D loss)	84.75	272.23	69.10	230.51
+ Feature Fusion	82.02	245.10	67.61	222.95
+ $\mathcal{L}_{rep}$	74.62	202.28	63.43	185.29
Ours (full components)	**66.76**	**182.94**	**62.14**	**177.95**

Table 5
MPJPE and MRPE on the Pizza group from the modified CMU Panoptic dataset with HD camera 2 and 19, 4, 6, and 13.

Methods	Cam2&19	Cam4	Cam6	Cam13	MPJPE ↓	MRPE ↓
Moon et al.	125.09*	140.26*	145.87*	132.29*	133.72	809.61
Lin et al.	125.09*	140.26*	145.87*	132.29*	133.72	382.50
Ours	**80.26**	**103.43**	**114.81**	**98.59**	**95.47**	233.05

5.2.4. Sensitivity analysis

We conduct experiments to study the sensitivity of our method with/without using our re-projection module (Ours w/o rep) in two dimensions: image and instance level. Experimental results on the modified CMU Panoptic dataset are shown in Fig. 6.

Image level. We first analyze the sensitivity of our method on each image for different levels of image distortions shown in Fig. 6(a). Images in the testing set are divided into three groups based on distortion parameter k_1: [−0.96, −0.88), [−0.88, −0.79), [−0.79, −0.70]. We then compare the relative change of the values of MPJPE and MRPE. Specifically, the absolute relative changes of MPJPE and MRPE are (1) 0.60% and 5.60% for our approach, (2) 1.48% and 7.32% for our approach without using the proposed re-projection module, respectively

Instance level. We analyze the sensitivity of our method on each person with different distortion strengths defined by θ_d/θ. θ_d/θ is categorized into [0.46, 0.8), [0.8, 0.9), [0.9, 1) for all humans appearing in the testing set. As the number of humans suffering from strong distortions is small, θ_d/θ is not uniformly grouped. In this setting, the number of humans in [0.9, 1) is still larger than the number of humans in the other two ranges. For simplification, we use the value of θ/θ_d of the root joint locations to represent the value of the full body. Therefore, the instance still suffers from image distortions even if the value of θ_d/θ is equal to 1. The results are shown in Fig. 6(b).

It is shown in Fig. 6 that the larger the distortion, the larger the value of the two metrics in both dimensions. That is expected as large image distortions cause significant changes of persons on the image plane. Experimental results demonstrate that our re-projection module reduces the negative impact of image distortions on multi-person 3D pose estimation, especially for absolute root joint estimation.

5.3. Discussion

Performance for other fisheye camera settings. In this paper, we synthesize images captured by HD cameras 2 and 19 from the CMU panoptic dataset. To validate the effectiveness of our method on images with different camera settings, images taken by the HD cameras 4, 6, and 13 are synthesized with different levels of distortion parameters. Particularly, the focal lengths and principal points are different. For simplification, we only select images from the Pizza group as the testing set to evaluate the methods, while the training set is the same as the setting in Section 5.1. The results are listed in Table 5. It is shown that:

Table 6
MPJPE results on the 3DhUman dataset: D1 represents the images captured by the first fisheye camera in the training set, while the testing set consists of images taken by the second fisheye camera. D2 is the opposite of D1.

Methods	D1	D2
Moon et al.	68.32*	71.44*
Lin et al.	68.32*	71.44*
Ours	**61.93**	**67.72**

(1) changing the viewpoints and fisheye camera settings degrade the performance of all methods; (2) our method outperforms other methods for the new camera parameters.

Camera settings of the 3DhUman dataset. The 3DhUman dataset includes two sets of camera parameters. To avoid the potential of overfitting on our 3DhUman dataset, we additionally define training and testing sets by whether the image is taken by the same fisheye camera. Table 6 shows the results with the MPJPE metric. Our method provides superior performance in both settings and exhibits the potential to mitigate the distortion problem on real-world scenes.

6. Conclusion

In this paper, we first presented a novel top-down approach for multi-person 3D pose estimation from a single image captured by a fisheye camera. In contrast to existing top-down approaches, our method maintains the global information to estimate absolute root depths and camera parameters. We proposed a re-projection module to enforce projected 3D predictions consistent with 2D ground truths under image distortions by minimizing the re-projection error. In this way, the impact of image distortion has been alleviated, and absolute depths of root joints has been further regularized. Compared with existing work, our method showed the state-of-the-art performance on both synthesized and real-world datasets.

CRediT authorship contribution statement

Yahui Zhang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Shaodi You:** Conceptualization, Writing – review & editing. **Sezer Karaoglu:** Resources, Writing – review & editing. **Theo Gevers:** Resources, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The first author is financially supported by China Scholarship Council [NO. 201806160025].

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2022.103505.

References

Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S., 2014. 3D pictorial structures for multiple human pose estimation. In: CVPR. pp. 1669–1676.
Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299.
Chen, Z., Huang, Y., Yu, H., Wang, L., 2022. Learning a robust part-aware monocular 3D human pose estimator via neural architecture search. Int. J. Comput. Vis. 130 (1), 56–75.
Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., 2018. Cascaded pyramid network for multi-person pose estimation. In: CVPR. pp. 7103–7112.
Cheng, Y., Wang, B., Yang, B., Tan, R.T., 2021. Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks. In: CVPR. pp. 7649–7659.
Cho, H., Cho, Y., Yu, J., Kim, J., 2021. Camera distortion-aware 3D human pose estimation in video with optimization-based meta-learning. In: ICCV. pp. 11169–11178.
Ci, H., Wang, C., Ma, X., Wang, Y., 2019. Optimizing network structure for 3d human pose estimation. In: ICCV. pp. 2262–2271.
Dabral, R., Gundavarapu, N.B., Mitra, R., Sharma, A., Ramakrishnan, G., Jain, A., 2019. Multi-person 3d human pose estimation from monocular images. In: 3DV. IEEE, pp. 405–414.
Dong, Z., Song, J., Chen, X., Guo, C., Hilliges, O., 2021. Shape-aware multi-person pose estimation from multi-view images. In: ICCV. pp. 11158–11168.
Fang, H.-S., Xie, S., Tai, Y.-W., Lu, C., 2017. Rmpe: Regional multi-person pose estimation. In: ICCV. pp. 2334–2343.
Guo, Y., Ma, L., Li, Z., Wang, X., Wang, F., 2021. Monocular 3D multi-person pose estimation via predicting factorized correction factors. Comput. Vis. Image Underst. 213, 103278.
Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C., 2019. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In: CVPR. pp. 10905–10914.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: ICCV. pp. 2961–2969.
Hidalgo, G., Raaj, Y., Idrees, H., Xiang, D., Joo, H., Simon, T., Sheikh, Y., 2019. Single-network whole-body pose estimation. In: ICCV. pp. 6982–6991.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141.
Hughes, C., Glavin, M., Jones, E., Denny, P., 2009. Wide-angle camera technology for automotive applications: a review. IET Intell. Transp. Syst. 3 (1), 19–31.
Hwang, D.-H., Aso, K., Yuan, Y., Kitani, K., Koike, H., 2020. MonoEye: Multimodal human motion capture system using a single ultra-wide fisheye camera. In: Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. pp. 98–111.
Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2013. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI 36 (7), 1325–1339.
Jiang, H., Grauman, K., 2017. Seeing invisible poses: Estimating 3d body pose from egocentric video. In: CVPR. IEEE, pp. 3501–3509.
Jin, S., Liu, W., Ouyang, W., Qian, C., 2019. Multi-person articulated tracking with spatial and temporal embeddings. In: CVPR. pp. 5664–5673.
Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y., 2015. Panoptic studio: A massively multiview system for social motion capture. In: ICCV. pp. 3334–3342.
Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J., 2018. End-to-end recovery of human shape and pose. In: CVPR. pp. 7122–7131.
Kannala, J., Brandt, S., 2004. A generic camera calibration method for fish-eye lenses. In: ICPR 2004, Vol. 1. IEEE, pp. 10–13.
Kim, H., Jung, J., Paik, J., 2016. Fisheye lens camera based surveillance system for wide field of view monitoring. Optik 127 (14), 5636–5646.
Kocabas, M., Karagoz, S., Akbas, E., 2018. Multiposenet: Fast multi-person pose estimation using pose residual network. In: ECCV. pp. 417–433.
Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K., 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: ICCV. pp. 2252–2261.
Li, J., Wang, C., Liu, W., Qian, C., Lu, C., 2020. HMOR: Hierarchical multi-person ordinal relations for monocular multi-person 3D pose estimation. In: ECCV. pp. 242–259.
Lin, J., Lee, G.H., 2020. HDNet: Human depth estimation for multi-person camera-space localization. In: ECCV. pp. 633–648.
Lin, J., Lee, G.H., 2021. Multi-view multi-person 3d pose estimation with plane sweep stereo. In: CVPR. pp. 11886–11895.
Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C., 2018. Single-shot multi-person 3d pose estimation from monocular rgb. In: 3DV. IEEE, pp. 120–130.
Moon, G., Chang, J.Y., Lee, K.M., 2019a. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: ICCV. pp. 10133–10142.
Moon, G., Chang, J.Y., Lee, K.M., 2019b. Posefix: Model-agnostic general human pose refinement network. In: CVPR. pp. 7773–7781.
Newell, A., Huang, Z., Deng, J., 2017. Associative embedding: End-to-end learning for joint detection and grouping. In: NeurIPS. pp. 2277–2287.

Nie, X., Feng, J., Zhang, J., Yan, S., 2019. Single-stage multi-person pose machines. In: ICCV. pp. 6951–6960.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K., 2017. Towards accurate multi-person pose estimation in the wild. In: CVPR. pp. 4903–4911.

Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.-P., Schiele, B., Theobalt, C., 2016. Egocap: egocentric marker-less motion capture with two fisheye cameras. New York, NY, USA, ACM Trans. Graph. New York, NY, USA, 35 (6).1–11,

Rogez, G., Weinzaepfel, P., Schmid, C., 2017. Lcr-net: Localization-classification-regression for human pose. In: CVPR. pp. 3433–3441.

Rogez, G., Weinzaepfel, P., Schmid, C., 2019. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. TPAMI 42 (5), 1146–1161.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703.

Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., De la Torre, F., 2020. Selfpose: 3D egocentric pose estimation from a headset mounted camera. TPAMI.

Tome, D., Peluse, P., Agapito, L., Badino, H., 2019. Xr-egopose: Egocentric 3d human pose from an hmd camera. In: ICCV. pp. 7728–7738.

Van den Heuvel, F., Verwaal, R., Beers, B., 2007. Automated calibration of fish-eye camera systems and the reduction of chromatic aberration. Photogramm. Fernerkundung Geoinformation 2007 (3), 157.

Wang, J., Liu, L., Xu, W., Sarkar, K., Theobalt, C., 2021. Estimating egocentric 3d human pose in global space. In: ICCV. pp. 11500–11509.

Wu, S., Jin, S., Liu, W., Bai, L., Qian, C., Liu, D., Ouyang, W., 2021. Graph-based 3d multi-person pose estimation using multi-view images. In: ICCV. pp. 11148–11157.

Xiao, B., Wu, H., Wei, Y., 2018. Simple baselines for human pose estimation and tracking. In: ECCV. pp. 466–481.

Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.-P., Theobalt, C., 2019. Mo2cap2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. IEEE Trans. Vis. Comput. Graphics 25 (5), 2093–2101.

Zanfir, A., Marinoiu, E., Sminchisescu, C., 2018a. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: CVPR. pp. 2148–2157.

Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.-I., Sminchisescu, C., 2018b. Deep network for the integrated 3d sensing of multiple people in natural images. In: NeurIPS. pp. 8410–8419.

Zhang, Y., You, S., Gevers, T., 2021. Automatic calibration of the fisheye camera for egocentric 3D human pose estimation from a single image. In: WACV. pp. 1772–1781.

Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N., 2019. Semantic graph convolutional networks for 3D human pose regression. In: CVPR. pp. 3425–3435.

Zhen, J., Fang, Q., Sun, J., Liu, W., Jiang, W., Bao, H., Zhou, X., 2020. SMAP: Single-shot multi-person absolute 3D pose estimation. In: ECCV. pp. 550–566.

Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y., 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: ICCV. pp. 398–407.