Geometric modeling for 3D human pose estimation and motion transfer

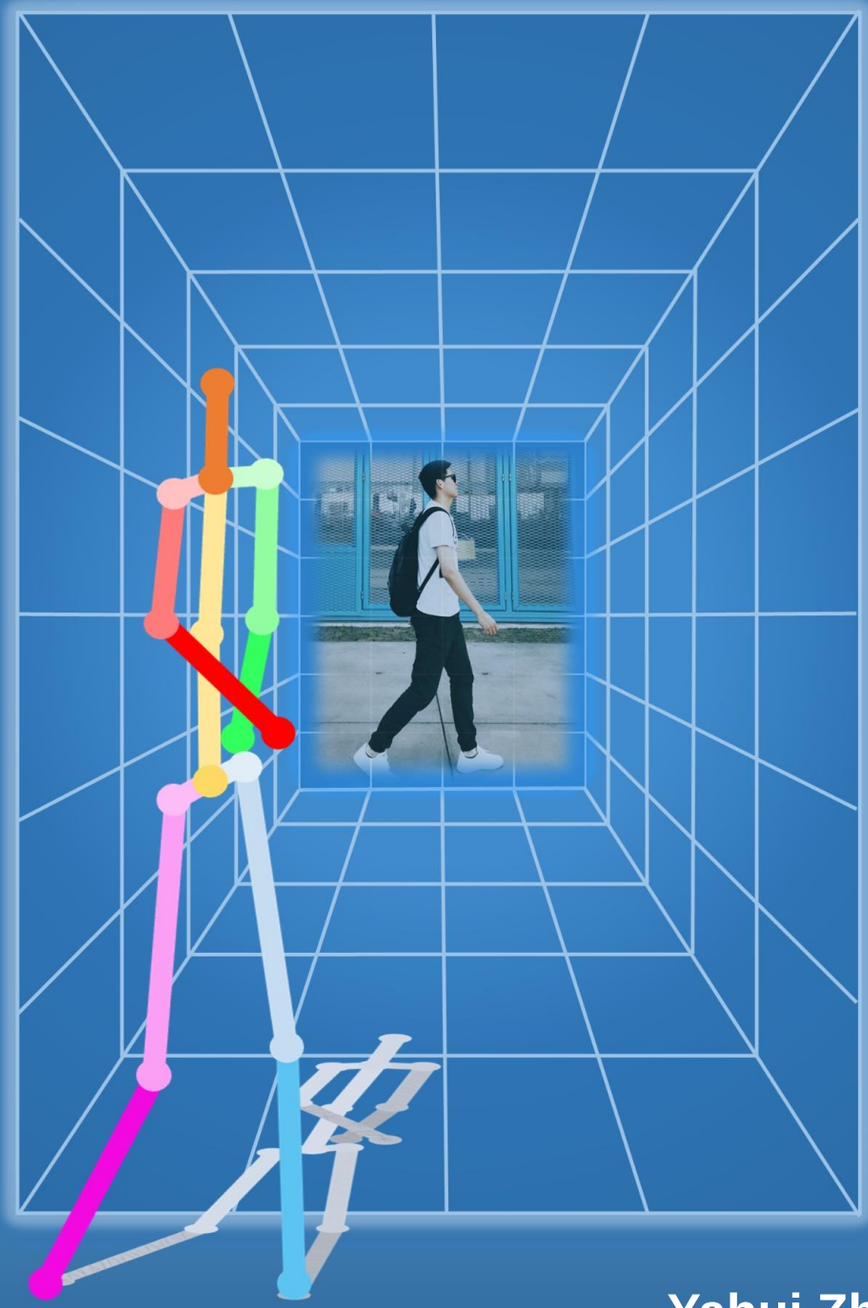
Zhang, Y.

Publication date
2022
Document Version
Final published version

Geometric Modeling for 3D Human Pose Estimation and Motion Transfer

Yahui Zhang

Geometric Modeling for 3D Human Pose Estimation and Motion Transfer



Yahui Zhang

Geometric Modeling for 3D Human Pose Estimation and Motion Transfer

Yahui Zhang

This book was typeset by the author using LaTeX 2_ε.

ISBN 978-94-93278-32-5

Geometric Modeling for 3D Human Pose Estimation and Motion Transfer

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 20 december 2022, te 10.00 uur

door

Yahui Zhang

geboren te Henan

Promotiecommissie

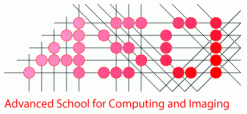
Promotor:	prof. dr. T. Gevers	Universiteit van Amsterdam
Co-promotor:	dr. S. You	Universiteit van Amsterdam
	dr. S. Karaoglu	Universiteit van Amsterdam
Overige leden:	prof. dr. A. A. Salah	Universiteit Utrecht
	prof. dr. ir. B. J. A. Kröse	Universiteit van Amsterdam
	prof. dr. M. Worring	Universiteit van Amsterdam
	dr. H. Dibeklioglu	Bilkent University
	dr. D. Tzionas	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

UNIVERSITEIT VAN AMSTERDAM

The research was supported by China Scholarship Council (CSC) under project number 201806160025.

The work described in this thesis has been carried out within the graduate school ASCI, dissertation number 440, at the Computer Vision Lab of the University of Amsterdam.



CONTENTS

1

INTRODUCTION

Human body movements (poses) convey essential information in non-verbal communication. They show how we feel and convey messages such as happiness, excitement, anger and support what we say. For example, people (*e.g.,* Italians) may use their posture and hands to accentuate what they say. "The body can't lie." said Traci Brown. Body language shows your true emotion and plays an elementary role in our daily life.

According to Charles Robert Darwin, the evolution of the human posture has undergone a great process of change: a specific human posture corresponds to a particular era. Human poses also represent what you are doing or going to do. By analyzing humans poses, we can derive human behaviour and intention. Obviously, human poses are the key in many activities such as diving, figure skating, and dancing. Amateurs or athletes can imitate postures of top performers to perform their exercises, while doctors can analyze humans to perform posture correction for children or young people and injury prevention for athletes. In conclusion, human poses are essential to convey and understand nonverbal signals by all of us.

Today, we can use cameras to record our daily activity and corresponding human body movements (poses). With the recent success of deep learning techniques, the automation of human-related tasks in computer vision has been achieved great improvement. Human pose estimation (HPE) is applied in many applications such as healthcare, virtual reality, and camera surveillance as shown in Figure 1. In the future, with the rise of Metaverse, it will become a regular part of our everyday existence.

HPE can be categorized into 2D HPE and 3D HPE approaches. 2D HPE aims to detect human joint locations in 2D planes. Due to the emergence of deep learning and large-scale 2D pose datasets such as MPII [2], COCO [109], and LSP [70], 2D human pose estimation [12, 95, 136, 196] has achieved superior performance on real-world images. The goal of 3D HPE is to estimate human joint locations in 3D Cartesian space derived from 2D images. In fact, 3D poses can better represent human postures than 2D poses due to the inherent ambiguity in 2D. However, it is a challenge for computer vision algorithms to understand 3D human poses from 2D planes.

A more challenging but active task is 3D HPE from a single image. The problem is depth ambiguity. Furthermore, due to self-occlusion, different 2D poses may correspond to the same 3D pose. Deep learning-based methods to deal with 3D HPE require large-scale datasets to train the models. However, it is labor intensive and extremely difficult to annotate 3D poses, especially for in-the-wild images. Therefore, public datasets of 3D human poses are usually collected for indoor environments. Lack of 3D ground truth for in-the-wild images may result in limited generalization of 3D HPE methods.

Current methods for single-person 3D HPE mainly aim to predict root-relative 3D human joint locations in the camera coordinate system from perspective images. It is

Surveillance Virtual reality Healthcare

Figure 1: Applications of human pose estimation.

common to use the pelvis joint as the root. Due to the large field of view (FoV), fisheye cameras have been widely used in different applications such as augmented/virtual reality, autonomous driving, photography. Generally, there are first-person view (egocentric) and third-person view 3D HPE's. For egocentric 3D HPE, the fisheye camera is usually installed at the head [180] or the baseball cap [205]. However, image distortions and strong perspective effects may negatively influence the results. A more general scenario for third-person view 3D HPE is that images may include several humans, *i.e.,* multi-person 3D human pose estimation for fisheye cameras. The goal is to predict 3D human joint locations with absolute depths in the camera coordinate system from a single image captured by a fisheye camera. Humans in the images usually introduce different distortion strengths, causing this task to be challenging.

3D human poses are useful for human motion transfer to deal with the self-occlusion problem. Human motion transfer aims to animate a human in a source image based on the driving poses of a human in target images/videos. It has many practical applications including movie production, entertainment and education. The challenge is to build a relationship between two humans in source and driving images. Generative adversarial network (GAN)-based methods show promising performance on image generation. They usually focus on global or style transformations but ignore the geometric relations. Instead, most existing methods utilize off-the-shelf models to estimate 2D or 3D human poses followed by optical flow computation. Although superior performance has been achieved, the problem to perform human motion transfer with pose consistency still remains.

1.1 RESEARCH OUTLINE AND QUESTIONS

3D human pose estimation from a single 2D image in the wild is an important computer vision task. Although there are several large-scale datasets for 3D human poses including Human3.6M [64] and MPI-INF-3DHP [127], they are collected for indoor environments. Unlike images taken from indoor and well constrained environments, 2D outdoor (in-the-wild) images are extremely complex because of varying imaging conditions. Furthermore, 2D images usually do not have corresponding 3D ground truth causing supervised approaches to be ill-constrained. Existing methods [51, 185] attempt to regularize the estimated 3D poses from in-the-wild images by minimizing the distance between projected 3D estimations and 2D poses. However, the perspective re-projections based

on a camera model may cause overfitting and can make the training process unstable. Therefore, in this thesis, the first research question is:

How can we improve the generalization of 3D human pose estimation for in-the-wild images?

In Chapter 2, we propose to associate 3D human pose, 2D human pose projection and 2D image appearance through a new orthographic projection based linear regression module. Unlike existing re-projection based approaches, our orthographic projection and regression do not suffer from small angle problems, which usually lead to overfitting in the depth dimension. The proposed orthographic projection based linear regression is used to associate 3D predictions with 2D poses. In this way, the network properly adapts to various datasets without fully retraining on them.

Hence, we propose a deep neural network which adopts 2D poses, 3D pose regression and orthographic projection linear regression modules. The network uses a two-stage scheme, where 2D poses using heatmap representations are detected first and followed by a 2D-to-3D lifting module. The proposed method shows state-of-the-art performance on the Human3.6M dataset and generalizes well to in-the-wild images.

With a large field of view, 3D human pose estimation from egocentric fisheye viewpoints has many valuable applications. The problem of estimating egocentric 3D poses for a fisheye camera is that images may be subject to strong image distortions, *i.e.,* 2D poses on the image plane that pass through the line of sight of the fisheye lens. Recent works [180,205] focus on the self-occlusion problem of the lower-body estimation due to the top-down viewpoint. The effect of image distortions on egocentric 3D pose estimation still remains. Therefore, we pose the second research question:

How can we alleviate the negative influence caused by image distortions for egocentric 3D human pose estimation?

In Chapter 3, we propose a method for egocentric 3D human pose estimation from a single image captured by a fisheye camera. We approach this problem by an automatic calibration module. Given a single image, our network first estimates 3D joint locations of a human in camera coordinates. To deal with the impact of image distortions on 3D human pose estimation, we then use the automatic calibration with self-correction to further regularize 3D predictions.

The proposed calibration module automatically estimates the intrinsic and distortion camera parameters with self-correction instead of using a post-processing step [205] to enforce the 3D predictions to be consistent with the corresponding distorted 2D poses. In this way, the effect of distortions on 3D pose estimation is alleviated.

To assess the effectiveness of the proposed automatic calibration module, we modified the *x*R-EgoPose dataset [180], a recent public dataset for 3D human pose estimation collected by a fisheye camera, by adding different levels of image distortions. Experimental results demonstrate that the proposed method achieves state-of-the-art performance.

Multi-person 3D pose estimation with absolute depths taken by a fisheye camera is a challenging task with many interesting applications such as surveillance and monitoring. However, to the best of our knowledge, such problem has not been explored so far. Compared with 3D HPE from pinhole cameras, humans at different positions may cause

different distortion strengths. In addition, different from egocentric 3D pose estimation, this task is more complicated because the distance between humans and cameras is not fixed. Finally, it is hard to predict 3D human joint locations with absolute depth as it is more challenging than root-relative 3D pose estimation because of the inherent depth and scale ambiguity. Therefore, the third research question is as follows:

How can we deal with the negative influence caused by image distortions for multi-person 3D pose estimation?

In Chapter 4, 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.

3D human pose estimation based on visual information aims to predict 3D poses of humans in images and videos. The aim of human action recognition is to classify what type of action a person takes. Both topics are widely studied in the field of computer vision. A more challenging but valuable task is to apply the above two tasks using fisheye images/videos. However, public datasets are mainly collected by pinhole cameras and ignoring the widespread use of fisheye cameras for 3D human pose estimation and skeleton-based action recognition. Therefore, the fourth research question is:

How can we evaluate models for 3D pose estimation and action recognition on real-world images captured by a fisheye camera?

In Chapter 5, we propose a new dataset for multi-person 3D pose estimation (F-M3DHPE), and skeleton-based HAR (F-HAR) captured by a fisheye camera. A comparison is conducted to analyze the performance of existing methods on the proposed dataset. We also provide a comprehensive survey on the recent advances of 3D human pose estimation and action recognition for both perspective and fisheye cameras.

Human motion transfer aims to animate the pose of a human in a source image driven by the poses of a human in a target video. To warp (transfer) human poses, most of the existing methods are based on optical flow or affine transformations as an intermediate representation followed by a generator module to perform the motion transfer. Existing methods perform well in terms of reconstruction quality. However, the quality of the human pose transfer has received less attention although it is an important part of the motion transfer process. Therefore, the fifth research question is as follows:

How can we perform human motion transfer with pose consistency?

In Chapter 6, we propose a method focusing on both the reconstruction quality as well as pose consistency. In contrast to existing methods, performing warping procedures in 2D- or 3D-space, we introduce a strategy to combine the warped features in both 2D- and 3D-space to alleviate the self-occlusion problem. In this way, our method benefits from 2D (robustness) and 3D (steering) information to guide the generation process. To reduce the pose error caused by inaccurate 3D estimation, a method is proposed to maintain semantic consistency between predictions and target images at arm and leg regions. Experiments conducted on large scale datasets show that the proposed method outperforms existing methods. Ablation studies clarify the benefits of using feature fusion and semantic consistency.

1.2 ORIGINS

This thesis is based on the following publications:

- **Chapter 2** is based on "Orthographic Projection Linear Regression for Single Image 3D Human Pose Estimation", published in *International Conference on Pattern Recognition*, 2021, by Yahui Zhang, Shaodi You, and Theo Gevers [223].

  Contribution of authors

  Yahui Zhang: all aspects,
  Shaodi You: conceptualization, supervision and writing,
  Theo Gevers: conceptualization, supervision, insight and writing.

- **Chapter 3** is based on "Automatic Calibration of the Fisheye Camera for Egocentric 3d Human Pose Estimation from a Single Image", published in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, by Yahui Zhang, Shaodi You, and Theo Gevers [222].

  Contribution of authors

  Yahui Zhang: all aspects,
  Shaodi You: conceptualization, supervision and writing,
  Theo Gevers: conceptualization, supervision, insight and writing.

- **Chapter 4** is based on "Multi-person 3D Pose Estimation from a Single Image Captured by a Fisheye Camera", published in *Computer Vision and Image Understanding*, 2022, by Yahui Zhang, Shaodi You, Sezer Karaoglu, and Theo Gevers [225].

  Contribution of authors

  Yahui Zhang: all aspects,
  Shaodi You: conceptualization, supervision and writing,
  Sezer Karaoglu: conceptualization, supervision and writing,
  Theo Gevers: conceptualization, supervision, insight and writing.

- **Chapter 5** is based on "Monocular 3D Human Pose Estimation and Action Recognition using Fisheye Cameras: A Survey and Benchmark", under review in *IEEE Transactions on Multimedia*, 2022, by Yahui Zhang, Shaodi You, Sezer Karaoglu,

and Theo Gevers [224].

Contribution of authors

Yahui Zhang: all aspects,
Shaodi You: conceptualization, supervision and writing,
Sezer Karaoglu: conceptualization, supervision and writing,
Theo Gevers: conceptualization, supervision, insight and writing.

- **Chapter 6** is based on "Pose Guided Human Motion Transfer by Exploiting 2D and 3D Information", published in *International Conference on 3D Vision*, 2022, by Yahui Zhang, Shaodi You, Sezer Karaoglu, and Theo Gevers [226].

  Contribution of authors

  Yahui Zhang: all aspects,
  Shaodi You: conceptualization, supervision and writing,
  Sezer Karaoglu: conceptualization, supervision and writing,
  Theo Gevers: conceptualization, supervision, insight and writing.

The author has further contributed to the following publication:

- Wei Wang, Shaodi You, Yahui Zhang, Sezer Karaoglu, and Theo Gevers. "Identity Invariant Age Transfer for Kinship Verification of Child-Adult Images", under review in *Computer Vision and Image Understanding*, 2022.

2

SCALED ORTHOGRAPHIC PROJECTION FOR 3D HUMAN POSE ESTIMATION

2.1 INTRODUCTION

Human pose estimation from a single image is an important computer vision task. It enables different applications in motion capture, virtual reality, and human-robot interaction. Noticeable achievements in two-dimensional (2D) human pose estimation have been made recently using Convolutional Neural Networks (CNNs) [84] in a data-driven fashion [2, 70, 109]. Recovering the 3D human pose is a challenging task because of the varying imaging conditions changing the appearance and occluded body parts. Moreover, it is an ill-posed problem because a single 2D image does not contain depth. Early stage methods focused on laboratory settings where 3D ground truth is measured using multi-view geometry or motion capture systems [18, 183]. In recent years, methods are focusing more on realistic and challenging tasks to estimate the 3D human pose in the wild without 3D ground truth. Recent work [51, 185] proposes to use perspective geometric constraints for 3D-2D poses. However, such constraints are point-wise and do not incorporate the context between joints. Also, while it is geometrically correct to constrain the depth from a perspective geometry, it is very unreliable because the angle between the camera center and projection line is usually small generating large errors in depth.

In this chapter, we propose an orthographic projection based linear regression method to constrain the 3D pose, 2D pose and 2D appearance. The advantage is that the orthographic constraint ensures not to generate strong changes in depth and avoids overfitting. The constrained linear regression exploits contextual information to properly constrain the 3D pose, 2D pose and 2D appearance.

We adopt a two-stage scheme to regress 3D human joint locations from a single *RGB* image. The proposed network first estimates the 2D pose using a heatmap representation. Then, the 2D heatmaps are used as inputs to a residual network followed by a series of fully connected layers to regress the 3D human pose in camera coordinates. Finally, the proposed orthographic projection based linear regression is used to associate the 3D predictions with 2D poses. In this way, the network properly adapts to various datasets without fully retraining on them. The network can be used to in-the-wild images without depth ground truth.

Experiments on several datasets are conducted to assess the proposed method. Specifically, we evaluate our method on Human3.6M [64] and MPI-INF-3DHP [127] datasets quantitatively and on in-the-wild MPII [2] and LSP [70] 2D human pose datasets qualitatively.

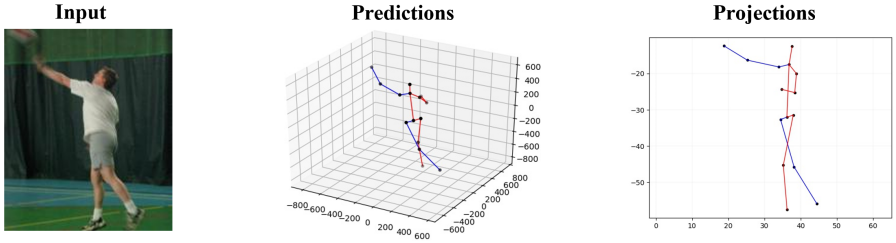


Figure 2: Estimated 3D pose and 2D projection of our proposed method from an example of in-the-wild image in LSP dataset [70].

The contributions of this work are summarized as follows:

- We propose a novel orthographic projection and linear regression to constrain the 3D and 2D poses.
- A network is proposed which is adaptive to various in-the-wild images without retraining the 3D pose.
- Our network achieves state-of-the-art performance on the Human3.6M dataset and generalizes well to in-the-wild datasets.

2.2 RELATED WORK

The proposed method estimates 3D human joint locations from a single *RGB* image. In this section, we focus on recent approaches to 3D human pose estimation. We divide the related work into one-stage methods, two-stage methods, adversarial learning methods and re-projection methods.

One-Stage Methods. One-stage approaches regress 3D coordinates of human joints directly from monocular images. Li and Chan *et al*. [98] apply a deep neural network to obtain 3D human pose estimation from a single image. Tekin *et al*. [176] adopt a pose auto-encoder for structure learning for human pose predictions. Pavlakos *et al*. [143] extend 2D joint heatmaps into a discretized 3D space by embedding a coarse-to-fine mechanism. To make the training process differentiable, Sun *et al*. [174] utilize a soft-argmax operation to estimate the 2D/3D human pose. However, most of above approaches rely on a large number of training data with 3D pose annotations.

Two-Stage Methods. For two-stage approaches, most of them use 2D pose estimators to detect 2D keypoints [136, 196, 203] first and then regress the 3D human joint positions from the estimated 2D keypoints [16, 42, 90, 126, 133, 199] or a combination of 2D pose and some other information [47, 142, 190]. Martinez *et al*. [126] design a simple fully connected residual network to regress 3D poses using estimated 2D keypoints. Moreno-Noguer [133] estimates 3D pose by using a 2D-3D pairwise distance matrix. Pavlakos *et al*. [142] use ordinal depth relations of each joint as auxiliary information to estimate the 3D human joint coordinates. Fang *et al*. [42] design a learning pose grammar to encode relations of human body for 3D human pose estimation. Dabral *et al*. [33] introduce anatomical constraints including bone lengths, joint angle limits and limb interpenetration, to ensure plausible 3D human poses. Zhou *et al*. [233] propose a geometric constraint to supervise depth information from in-the-wild 2D images. However, there is a limitation for these methods using geometric constraints to regularize

the 3D pose: the global scale needs to be known to map the scale between the estimated 3D pose and 3D ground truth during the inference process.

Adversarial Learning Methods. There are a number of adversarial learning methods for human pose estimation [17, 38, 72, 185, 209]. These methods use adversarial learning to distinguish the estimated pose from the real human pose. Yang *et al*. [209] introduce an adversarial network to determine whether the predicted 3D pose generated by a 3D regression network is plausible compared to the ground truth. Wandt *et al*. [185] propose an adversarial re-projection network to relax the constraint of training with 2D-3D correspondences. Similarly, this method also uses adversarial learning to map the distribution of the estimated 3D pose to the domain of the 3D ground truth. Instead of learning human pose priors from 3D ground truth, Chen *et al*. [17] and Drover *et al*. [38] employ adversarial learning to learn 3D priors for human pose based on 2D projections by 3D human joint regression without 3D annotations. However, these methods need augmented 2D projections from a dataset or visual cameras to augment the training set. Besides, the 3D priors can only be learned from the 3D pose datasets from an indoor environment, which means that the variety of 3D poses is limited.

Re-projection Methods. To keep the 3D predictions and intermediate 2D poses consistent, a re-projection method is used to obtain geometric self-consistency to regularize 3D pose regression. In general, there are two types of re-projection methods: projection from 1) 3D human keypoints [8, 51, 185] and 2) 3D human body (usually using a parametric SMPL body model) [6, 72, 139, 144].

Bogo *et al*. [6] provide 3D human pose estimation by minimizing the error between the detected 2D pose and 2D pose projections of the estimated statistical body model. Kanazawa *et al*. [72] combine 2D pose re-projection losses of the parametric SMPL model with several adversarial regularizers to constrain the SMPL model. Pavlakos *et al*. [144] employ a differentiable renderer to project the parametric model to the image and then minimize the detected 2D pose and silhouette error.

Instead of fitting a parametric human body model, recent methods attempt to minimize the detected 2D pose and 2D projections of the estimated 3D human joint positions. Habibie *et al*. [51] first regress root-relative 3D human joint positions using 2D heatmaps and intermediate 3D representations. Then, the predicted camera parameters, *i.e.*, focal length and principal coordinate are used to project root-relative 3D positions into 2D poses. The perspective re-projections based on a camera model may cause overfitting and can make the training process unstable. In this chapter, we employ orthographic projection linear regression to relate the estimated 3D pose in camera coordinates with the 2D pose on the image plane.

2.3 ORTHOGRAPHIC PROJECTION LINEAR REGRESSION

To avoid suffering from small angle problem resulting in overfitting in the depth dimension, we propose an orthographic projection linear regression method to constrain 3D predictions, 2D poses and 2D appearance. In this section, we introduce the orthographic projection as well as the perspective projection. Further, a method to estimate the 3D human skeleton from a single 2D image is presented. Based on it, we introduce our constrained linear regression model.

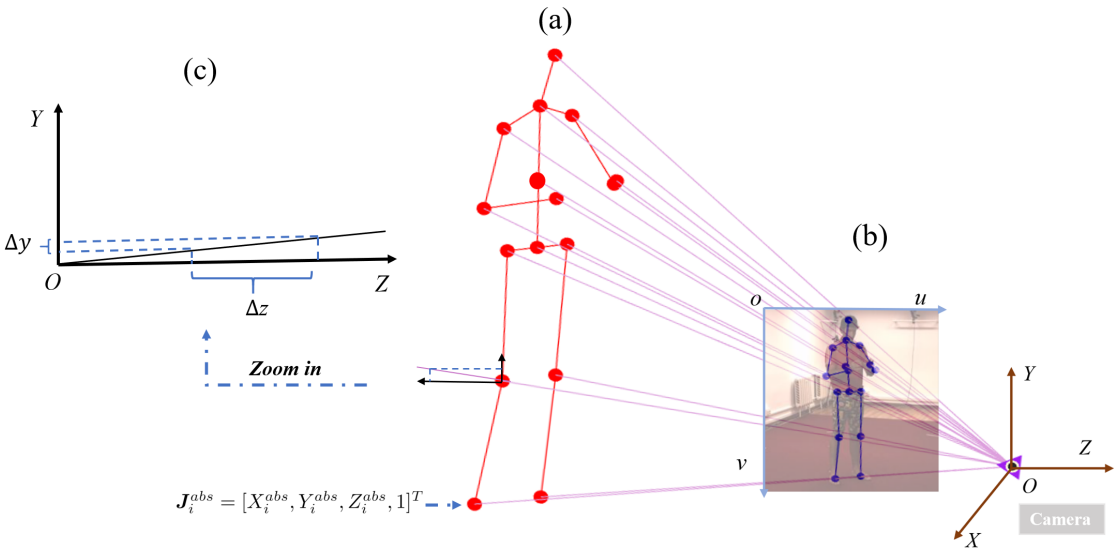


Figure 3: Perspective projection from **(a)** 3D pose in the camera coordinate system (*OXYZ*) to **(b)** 2D pose on image plane (*ouv*). **(c)** And the effect of depth (*i.e.*, *Z* axis) fluctuation on the coordinate on *Y* axis (the same as *X* axis). In small angle case, large depth changes in the camera coordinate will be projected as small changes on *Y* axis of the image plane.

2.3.1 *3D Human Pose Projection from a Single Image*

From 3D joints to 2D joints. As illustrated in Figure 3, human pose in 3D space is presented by a set of joints. Without loss of generality, the representation in 3D Euclidean space with camera coordinates is: a 4 by n matrix $\mathbf{P}_{3D}^{abs} = [\boldsymbol{J}_1^{abs}, \boldsymbol{J}_2^{abs}, ..., \boldsymbol{J}_n^{abs}]$ from a single *RGB* image, where n denotes the number of human joints and $\boldsymbol{J}_i^{abs} = [X_i^{abs}, Y_i^{abs}, Z_i^{abs}, 1]^T$ in homogeneous coordinates which is the coordinate vector for each joint. It is assumed that the camera projection matrix is known and consists of intrinsic ($\boldsymbol{K}$) and extrinsic ($\boldsymbol{R}$ and $\boldsymbol{T}$) parameters. Since 3D pose locations are in camera coordinates, 3D joints $\mathbf{P}_{3D}^{abs}$ are projected into 2D joints $\mathbf{p}_{2D}$ by a 3 by n matrix with $\boldsymbol{j}_i^{abs} = [x_i^{abs}, y_i^{abs}, 1]^T$ on the image plane:

$$\mathbf{p}_{2D} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{3D}^{abs}, \tag{2.1}$$

where $\boldsymbol{R}$ and $\boldsymbol{T}$ are identical matrices. Or simply $\mathbf{p}_{2D} = \boldsymbol{K}\mathbf{P}_{3D}^{abs}$.

In this way, the projected 2D joints overlay with the captured 2D image, as illustrated in Figure 3(a) and (b).

However, given that the camera is usually far away from the human body, the 3D joint points are usually presented by relative coordinates from the root joint (*i.e.*, pelvis). Therefore, the representation of human pose is defined by $\boldsymbol{J}_{3D} = \boldsymbol{J}_{3D}^{abs} - \boldsymbol{J}_{root}$, which is obtained by subtracting the root joint locations.

From 2D Image to 3D Joints. The goal of our method is to estimate the 3D human joint positions from a 2D image. Given the complex appearance of 2D images in the wild, we cannot assume a perfect projection between the 3D joints, 2D joints and the 2D image. Therefore, we aim to apply a constrained orthographic projection linear regression to robustly associate the estimated 3D joint points, 2D joint points and the appearance of 2D images.

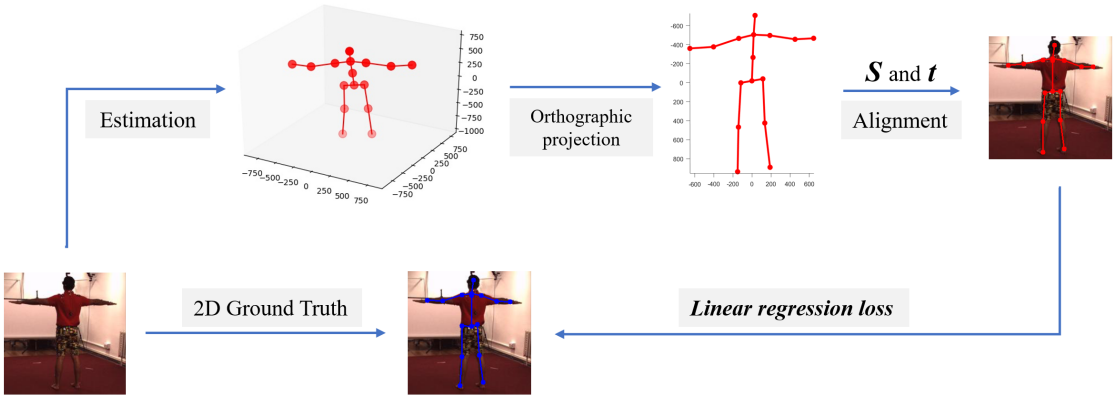


Figure 4: The general idea of matching 3D with 2D poses by the orthographic projection linear regression method. Orthographic projection is used first to reduce the risk of overfitting. Then, the predicted scale and translation parameters $\boldsymbol{S}$ and $\boldsymbol{t}$ are used to align the 3D projections with 2D poses on the image plane. In this way, 2D pose annotations implicitly regularize the estimated 3D pose.

2.3.2 *Orthographic Projection Linear Regression*

Here, we introduce the details of the proposed Orthographic Projection Linear Regression.

Orthographic Projection. We use orthographic projection and linear regression to constrain the estimated 3D joints. This allows us to ignore the depth and constrain only the 3D joints in the $X-Y$ image plane. The depth variance of human joints is smaller than the distance from camera to human body. Therefore, we approximate the perspective projection by an orthographic projection. Such approximation is usually named as the small angle problem. The projection, to extract X and Y value from the 3D coordinate, is defined by:

$$\begin{aligned} \mathbf{p}_{2D} &= \Pi \mathbf{P}_{3D}, \\ \Pi &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \end{aligned} \tag{2.2}$$

Existing methods exploit the projection constraint based on a perspective projection. Later, we will discuss the limitation of a perspective projection for small-angle scenarios, as shown in Figure 3(c).

Constrained Linear Regression. We use linear regression to align the projected 3D joints with the 2D joints. A canonical 2D linear regression can be used to obtain an affine transform, skew and rotation. However, for our problem, these transformations are unnecessary. Therefore, we constrain the linear regression to only scaling and translation.

The constrained orthographic projection is

$$\mathbf{p}_{2D} = [\boldsymbol{S}|\boldsymbol{t}]\Pi \mathbf{P}_{3D}, \tag{2.3}$$

where $\boldsymbol{S}$ and $\boldsymbol{t}$ indicate scale and translation parameters, respectively.

Eq. (2.3) is as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \\ 0 & 0 & 1 \end{bmatrix} \Pi \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \tag{2.4}$$

where (u, v) denotes the 2D pose on the image plane.

The linear regression computes the scaling and translation by minimizing the error between the estimated 3D joints projection and 2D pose $\mathbf{p}_{2D}$:

$$\arg\min_{\boldsymbol{S},\boldsymbol{t}} \left\| [\boldsymbol{S}|\boldsymbol{t}] \Pi \mathbf{P}_{3D} - \mathbf{p}_{2D} \right\|_2^2. \tag{2.5}$$

Later, in Section 2.4, we will introduce the loss functions and training strategy based on the proposed constrained linear regression method, where $\boldsymbol{S}$, $\boldsymbol{t}$ and $\mathbf{P}_{3D}$ are updated simultaneously.

2.3.3 *Limitation of Perspective Projection*

Recent work [51,185] proposes to use a perspective projection to keep the 3D predictions and intermediate 2D poses consistent. However, a perspective projection from 3D to 2D poses may cause problems. As shown in Figure 3(c), to minimize the error between 2D projections and 2D ground truth, the value of depth (*i.e.*, z axis) will deviate to find the optimal solution. This may lead to overfitting. Therefore, an orthographic projection is employed in this work to solve this issue.

2.4 METHODOLOGY

The goal of our method is to regress the root-centered 3D locations of human joints in a camera coordinate system from a single *RGB* image. To improve generalization to in-the-wild images, we propose to associate the 3D human pose, the 2D human pose projection and the appearance of 2D images through a new orthographic projection linear regression method. The overview of our framework is depicted in Figure 5.

2.4.1 *Network Design*

2D Pose Module. ResNet [53] is widely adopted for the human pose detection task [51,203,227]. In this work, ResNet-50 is applied as our backbone network followed by deconvolution layers for pose geometry feature extraction and 2D pose estimation. The estimated 2D poses are represented by heatmaps with a spatial dimension of 64×64.

We optimize the 2D pose module by minimizing the loss between 2D predictions and 2D ground-truth heatmaps. The loss function is defined as

$$\mathcal{L}_{Heatmap} = \sum_{h}^{H} \sum_{w}^{W} \left\| hm_{(h,w)} - hm_{(h,w)}^{GT} \right\|_2^2, \tag{2.6}$$

where H and W denote the resolution of heatmaps, GT means the ground-truth, and hm indicates the probability distribution of each joint.

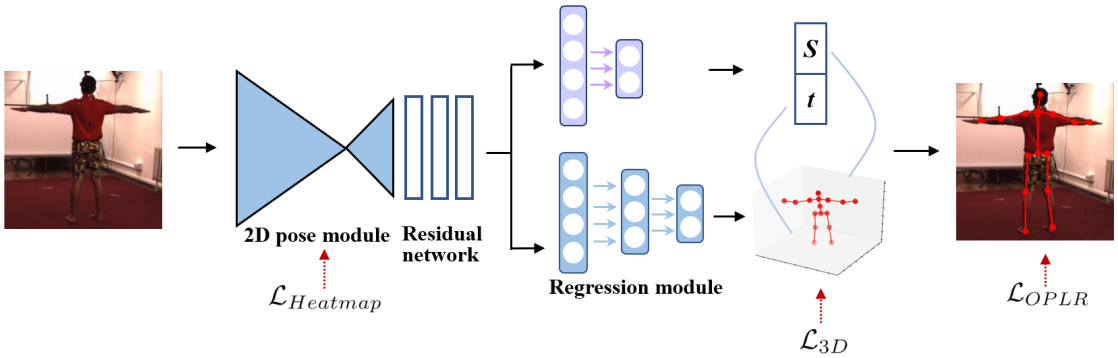


Figure 5: The overview of the proposed framework. We use Resnet as our backbone to detect 2D pose with heatmap representations. The 2D heatmaps are fed into a residual network with attention mechanism to further exploit the information in latent space. Then, we employ a series of fully connected layers to estimate the 3D pose in camera coordinates and re-projection parameters (*i.e.*, scale $\boldsymbol{S}$ and translation $\boldsymbol{t}$). Finally, the estimated 3D pose is transformed into 2D pose on image plane using the predicted $\boldsymbol{S}$ and $\boldsymbol{t}$.

3D Pose Regression Module. A residual neural network is applied to extract the geometric information from the 2D heatmaps. Specifically, the residual network consists of eight residual layers and four max pooling layers. To infer the 3D pose from intermediate features, we employ a fully connected network to regress the 3D human pose $\mathbf{P}_{3\mathrm{D}}$. Specifically, we use m consecutive residual blocks for pose estimation, where each block has two fully connected layers with a width of 1024 and ReLUs activation. To make a trade-off between the accuracy and time cost, the value of m is set to two.

Our 3D pose regression module is optimized by minimizing errors between the estimated 3D pose and 3D ground truth during the training process. The loss function is defined as

$$\mathcal{L}_{3D} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{P}_{3\mathrm{D}\,(i)} - \mathbf{P}^{\mathrm{GT}}_{3\mathrm{D}\,(i)}\right\|_2^2, \tag{2.7}$$

where i and N indicate the index of training samples and the total number of test samples, respectively.

Orthographic Projection Linear Regression Module. We use the proposed orthographic projection linear regression to associate the estimated 3D human pose with the 2D pose $\mathbf{p}_{2\mathrm{D}}$ according to Eq. (2.3). Specifically, scale parameters (s_x, s_y) and translation parameters (t_x, t_y) are predicted by a multi-layer perceptron. Given the 2D ground truth $\mathbf{p}^{\mathrm{GT}}_{2\mathrm{D}}$, the loss of this linear regression is defined as

$$\mathcal{L}_{OPLR} = \frac{1}{N}\sum_{i=1}^{N}\left\|[\boldsymbol{S}|\boldsymbol{t}]\Pi\mathbf{P}_{3\mathrm{D}\,(i)} - \mathbf{p}^{\mathrm{GT}}_{2\mathrm{D}\,(i)}\right\|_2^2. \tag{2.8}$$

The orthographic projection linear regression module predicts camera parameters from different viewpoints, which are used to associate the 3D predictions with the 2D poses. In this way, our method can be trained on 2D pose datasets including in-the-wild images and more viewpoints than 3D pose datasets. Therefore, our method is expected to generalize well to in-the-wild images.

Attention Mechanism. We adopt an attention mechanism in the above residual network to further exploit the information from the latent space. The reason behind this operation is that it is hard to determine which features contribute to the estimation process. Intuitively, geometry-related features are expected to contribute to our goal of regressing 3D human joint positions. However, it is difficult to separate this kind of features from the others. Therefore, an attention mechanism is employed to enhance the role of geometry-related features for estimating 3D human poses.

2.4.2 *Training*

According to Eqs. (2.6) - (2.8), the overall loss of our framework is summarized as

$$\mathcal{L}_{pose} = \lambda_{hm}\mathcal{L}_{Heatmap} + \mathcal{L}_{3D} + \lambda_{OPLR}\mathcal{L}_{OPLR}, \tag{2.9}$$

where λ_{hm} and λ_{OPLR} are loss weights to adjust the combination of 2D heatmap loss, 3D pose loss and orthographic projection linear regression loss.

We divide our training procedure into two stages. Stage 1: initialize the 2D pose module using the MPII dataset with 2D pose annotations. Stage 2: train the 3D pose regression and orthographic projection linear regression modules and fine-tune the 2D pose module on Human3.6M and MPII datasets. As our method with orthographic projection linear regression can be trained on the MPII dataset containing in-the-wild images with various appearances and viewpoints, our method is expected to generalize well to in-the-wild images.

2.5 EXPERIMENTS

Datasets. The datasets — Human3.6M, MPI-INF-3DHP and MPII Human Pose and LSP are used to validate our approach quantitatively/qualitatively.

Human3.6M [64] is the largest dataset for 3D human pose estimation and has been widely used as a benchmark for evaluation. The dataset consists of 3.6 million various 3D human poses. This dataset was collected by a MoCap System in a constrained environment from 11 actors covering 15 daily activities under 4 camera views, containing 2D/3D pose ground truth. MPI-INF-3DHP [127] is the recent dataset consisting of both constrained indoor and challenging outdoor scenes with diverse human appearances. This dataset is used to demonstrate the generalization capabilities of our approach. Particularly, we evaluate the proposed method on this dataset without using the training set. MPII Human Pose [2] and LSP [70] datasets provide a large number of in-the-wild images with only 2D human pose annotations, covering diverse activities and appearances. Therefore, MPII and LSP datasets are adopted here to qualitatively evaluate the cross-domain generalization ability of our approach.

Evaluation Protocols. Following the widely used protocols on Human3.6M [173, 174], we use six subjects (S1, S5, S6, S7, S8) for training and subjects S9 and S11 for testing. Particularly, every 5th and 64th frame of subjects are used for training and testing respectively. The evaluation metric is the Mean Per Joint Position Error (MPJPE), which is calculated after the alignment between the predicted and ground-truth 3D pose with the root joint and measured in millimeters. We refer to this as *Protocol*#1. Another

evaluation metric is PA MPJPE for which the predictions and ground-truth 3D poses are further aligned via a rigid transformation (Procrustes analysis [49], PA), which is referred to as *Protocol*#2. In this chapter, both *Protocol*#1 and *Protocol*#2 are employed for evaluation on Human3.6M dataset.

Implementation Details. We first pretrain our 2D pose module on MPII dataset in terms of the heatmap regression task for 94k iterations with a batch size of 64 and an initial learning rate of 5×10^{-4} with a decay over 70k iterations. In second stage, the network is trained on a mixture of MPII and Human3.6M datasets for 181k iterations. The batch size of this stage is 64. The initial learning rate is set as 5×10^{-4} with a decay over 111k and 160k iterations. The hyperparameters of loss weights of λ_{hm} and λ_{OPLR} are set as 2×10^6 and 5, respectively. The whole training procedures are implemented with two GTX 1080ti GPUs.

Method Comparisons. To evaluate the effectiveness of the proposed method, we conduct experiments to compare with existing state-of-the-art 3D human pose estimation methods. Note that only methods are selected which are related to our work, as there are many different categories of deep learning based methods to estimate 3D human body joint locations. [81, 106, 150, 152] provide 3D human pose estimation using multi-view images to reduce the ambiguity between 2D and 3D pose. [17, 38] propose an unsupervised approach merely using 2D pose annotations to regress coordinates of 3D human pose. These methods are not included in the comparison because of the specific task at hand.

2.5.1 *Quantitative Evaluation*

Human3.6M Dataset. We first evaluate the proposed method on the Human3.6M dataset to verify its effectiveness. Following [51, 143, 233], the global scale is assumed to be known. Table 1 lists the experimental results including the performance of each action in Human3.6M dataset.

Compared to existing 3D human pose estimation methods, the proposed method achieves state-of-the-art performance with an error of 56.2*mm* under *Protocol*#1. Particularly, our method provides the best performance for the five actions in the Human3.6M dataset. For *Protocol*#2, a rigid transformation is employed to align the estimated 3D pose with 3D ground truth. Our method also obtains a higher accuracy than most of state-of-the-art methods, only lower than Yang *et al*. [209]. Specifically, our method achieves the best performance for the actions *Phoning*, *Posing* and *Smoke*.

Table 2 lists the experimental results on Human3.6M dataset compared with other existing 3D human pose estimation methods using more information than ours, such as the training dataset, ordinal depth information, intrinsic camera parameters or other ground-truth information. It is shown in Table 2 that our method still achieves competitive performance without using extra information.

From Tables 1 and 2, it can be seen that our method outperforms the existing re-projection based methods [51, 185]. Because of our orthographic projection linear regression method, our method avoids suffering from small angle problem and achieved better performance.

MPI-INF-3DHP Dataset. To test the generalization of the proposed method, we evaluate our model on another dataset with in-the-wild images, *i.e.*, MPI-INF-3DHP dataset.

Table 1: The quantitative results compared to state-of-the-art 3D human pose estimation methods on Human3.6M under Protocol #1 and Protocol #2.

Protocol #1	Dir.	Dis.	Eat	Gre.	Phon.	Pose	Pur.	Sit	SitD.	Smo.	Phot.	Wait	Walk	WalkD.	WalkP.	Avg ↓
Zhou *et al.* (CVPR'16) [235]	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Chen *et al.* (CVPR'17) [16]	89.9	97.6	90.0	107.9	107.3	93.6	136.1	133.1	240.1	106.7	139.2	106.2	87.0	114.1	90.6	114.2
Pavlakos *et al.* (CVPR'17) [143]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Mehta *et al.* (3DV'17) [127]	57.5	68.6	59.6	67.3	78.1	56.9	69.1	98.0	117.5	69.5	82.4	68.0	55.3	76.5	61.4	72.9
Zhou *et al.* (ICCV'17) [233]	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.1	66.0	51.4	63.2	55.3	64.9
Sun *et al.* (ICCV'17) [173]	52.8	54.8	54.2	54.3	61.8	67.2	53.1	53.6	71.7	86.7	61.5	53.4	61.6	47.1	53.4	59.1
Luo *et al.* (BMVC'18) [122]	53.5	60.9	56.3	59.1	64.3	74.4	55.4	63.4	74.8	98.0	61.1	58.2	70.6	49.1	55.7	63.7
Yang *et al.* (CVPR'18) [209]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	58.6
Zhao *et al.* (CVPR'19) [227]	47.3	60.7	51.4	60.5	61.1	49.9	47.3	68.1	86.2	55.0	67.8	61.0	42.1	60.6	45.3	57.6
Ours	**46.0**	55.3	50.6	**53.5**	**57.5**	**46.3**	49.4	71.7	87.9	56.6	68.4	53.5	**41.4**	57.9	46.6	**56.2**

Protocol #2	Dir.	Dis.	Eat	Gre.	Phon.	Pose	Pur.	Sit	SitD.	Smo.	Phot.	Wait	Walk	WalkD.	WalkP.	Avg ↓
Moreno-Noguer (CVPR'17) [133]	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5	74.6	92.6	69.6	71.5	78.0	73.2	74.0
Sun *et al.* (ICCV'17) [173]	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
Luo *et al.* (BMVC'18) [122]	40.8	44.6	42.1	45.1	48.3	54.6	41.2	42.9	55.5	69.9	46.7	42.5	48.0	36.0	41.4	46.6
Yang *et al.* (CVPR'18) [209]	26.9	30.9	36.3	39.9	43.9	47.4	28.8	29.4	36.9	58.4	41.5	30.5	29.5	42.5	32.2	**37.7**
Zhou *et al.* (TPAMI'18) [236]	47.9	48.8	52.7	55.0	56.8	65.5	49.0	45.5	60.8	81.1	53.7	51.6	54.8	50.4	55.9	55.3
Ours	35.8	41.0	42.3	42.0	**43.4**	**36.3**	36.7	55.1	66.5	**45.0**	49.6	41.2	32.9	43.9	39.0	43.4

Table 2: Comparison with state-of-the-art methods using extra information or different input types for the metrics MPJPE and PA MPJPE on Human3.6M dataset.

Methods	Extra information	MPJPE ↓	PA MPJPE ↓
Using extra information			
Sun *et al*. (ECCV'18) [174]	Camera parameters	49.6	-
Habibie *et al*. (CVPR'19) [51]	Extra training set	65.7	49.2
Want *et al*. (CVPR'19) [185]	Extra training set	89.9	65.1
Sharma *et al*. (ICCV'19) [158]	Ordinal depth	58.0	40.9
Li *et al*. (CVPR'19) [90]	Ground Truth	52.7	42.6
Different input types			
Martinez *et al*. (ICCV'17) [126]	2D pose	62.9	47.7
Fang *et al*. (AAAI'18) [42]	2D pose	60.4	45.7
Ci *et al*. (ICCV'19) [30]	2D pose	52.7	42.2
Ours		56.2	43.4

Table 3: Comparison with state-of-the-art methods using PCK and AUC on MPI-INF-3DHP without training on this dataset.

Methods	Extra information	PCK ↑	AUC ↑
Mehta *et al*. (3DV'17) [127]	-	64.7	31.7
Zhou *et al*. (ICCV'17) [233]	Post-processing	68.2	32.5
Yang *et al*. (CVPR'18) [209]	-	69.0	32.0
Habibie *et al*. (CVPR'19) [51]	Extra training set	70.4	36.0
Want *et al*. (CVPR'19) [185]	Extra training set	81.8	54.8
Ci *et al*. (ICCV'19) [30]	2D Pose	74.0	34.7
Ours (w/o $\mathcal{L}_{OPLR}$)	-	23.9	8.9
Ours (full)	-	66.8	31.9
Using rigid transformation			
Habibie *et al*. (CVPR'19) [51]	Extra training set	82.9	45.4
Ours	-	**84.4**	**46.9**

Particularly, our model is only pretrained on Human3.6M and MPII datasets. PCK and AUC are used as evaluation metrics and the threshold of PCK is 150*mm*.

It should be noted that joint locations are defined differently between Human3.6M and MPI-INF-3DHP dataset, especially at the hip and neck joints. The different joint definitions affect our orthographic projection linear regression module which is used to connect the estimated 3D pose in camera coordinates with the 2D pose on the image plane, resulting in inaccurate 3D pose predictions.

Table 3 lists the experimental results compared with existing 3D human pose estimation methods. It it shown that 1) our method without extra information still achieves competitive results compared with existing 3D estimation methods and generalizes well to in-the-wild images; 2) our method outperforms Habibie *et al*. [51] after using the rigid transformation.

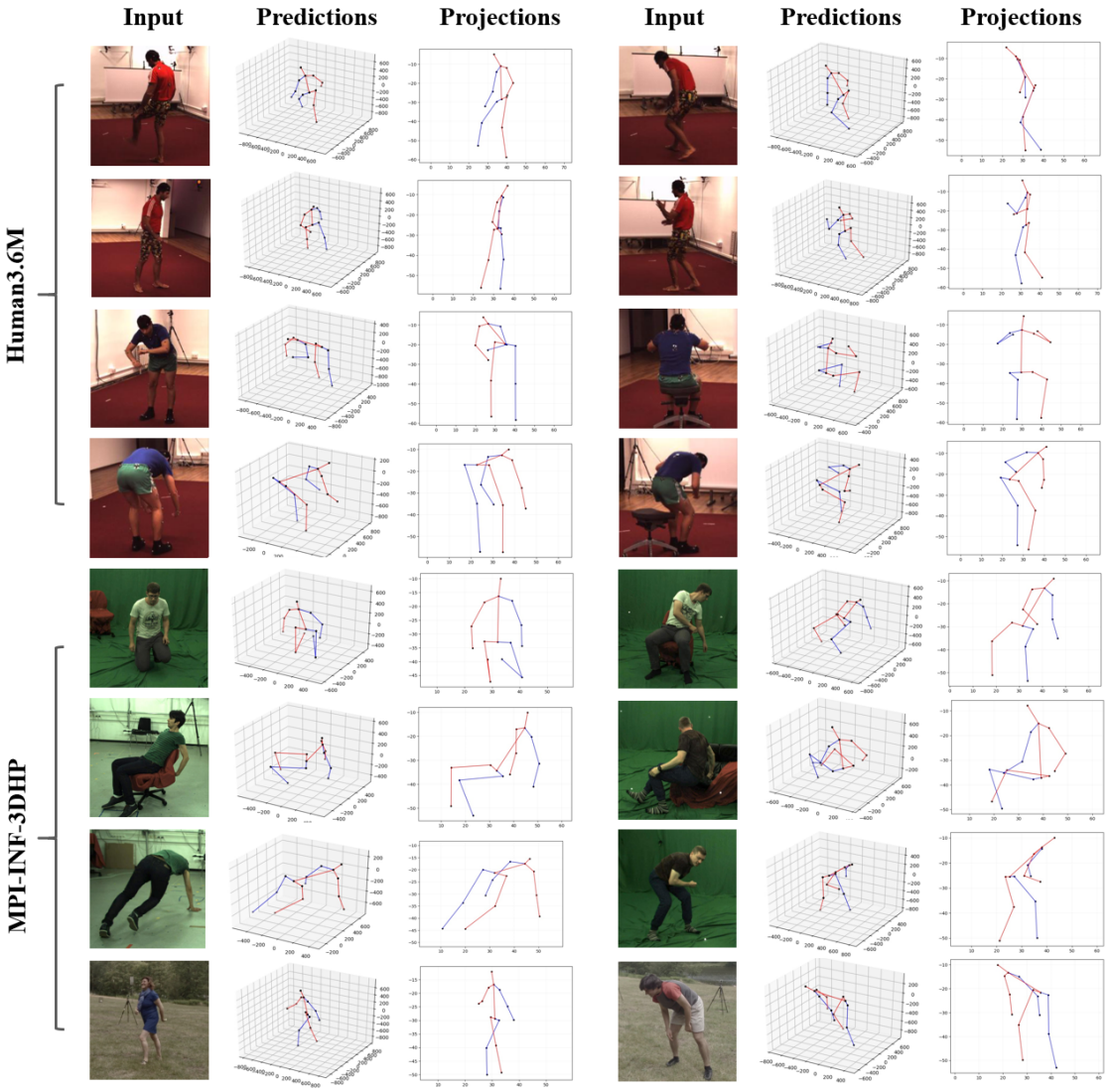


Figure 6: Qualitative results of the proposed network on Human3.6M and MPI-INF-3DHP datasets.

2.5.2 *Ablation Study*

In this section, we conduct an ablation study to measure the contribution of the proposed orthographic projection linear regression module. Specifically, experiments are conducted on the MPI-INF-3DHP dataset to evaluate the performance, where our model is only trained on Human3.6M and MPII datasets.

Table 3 lists the experimental results of our method with/without the orthographic projection linear regression module. It is shown that the method with all the components significantly improves the performance from 23.9% to 66.8%. Therefore, the proposed orthographic projection linear regression method greatly contributes to the generalization capability of our method to in-the-wild images.

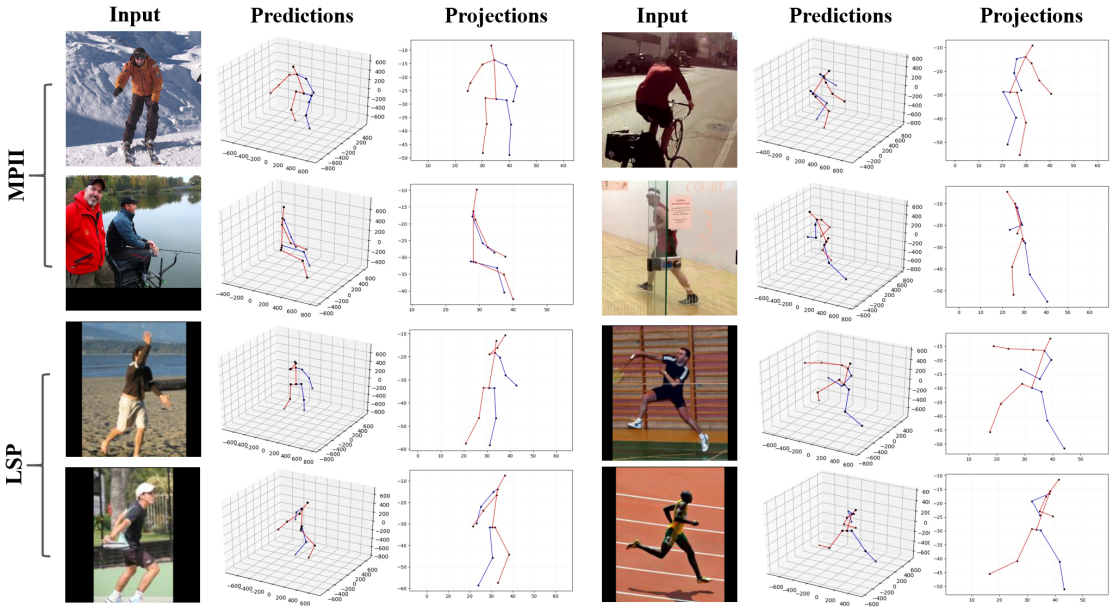


Figure 7: Qualitative results of the proposed network on MPII and LSP datasets.

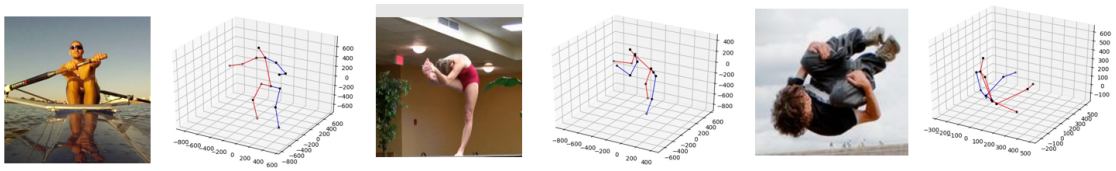


Figure 8: The failure cases generated by the proposed method.

2.5.3 *Qualitative Evaluation*

To further test the generalization, we provide qualitative results on MPII and LSP dataset in Figure 7. Particularly, MPII and LSP datasets consist of in-the-wild images with diverse human poses, backgrounds and appearances, which are hardly included in 3D pose datasets. It is shown in Figure 7 that our method with the proposed orthographic projection linear regression module generalizes well to in-the-wild images, where both estimated 3D poses and 2D projections are accurate.

There are several failure cases of our method as shown in Figure 8. In failure cases, the parts of human body are heavily occluded.

2.5.4 *Discussion*

Our method without $\mathcal{L}_{OPLR}$ cannot predict reasonable 3D poses from in-the-wild images even when the 2D pose module provides accurate 2D poses (with heatmap representations). Intuitively, the 3D pose estimations are expected to be accurate since the inputs of our regression module are the processed 2D heatmaps. The argument is that the 2D heatmap representations contain not only 2D pose information but also other misleading information, such as background information and light conditions.

Recent work [209, 233] shows that using a mixture of the 2D pose dataset (*i.e.*, MPII) and 3D pose dataset (*i.e.*, Human3.6M) increases the performance of the 2D pose module.

A combined training set is the main reason for augmenting the generalization to in-the-wild images.

2.6 CONCLUSIONS

In this Chapter 2, we proposed an orthographic projection linear regression module to construct a relation between the 3D human pose, 2D human pose projection and 2D image appearance. The proposed method first employs orthographic projection to reduce the impact of the depth part. Then, a linear regression is used to align the 3D projections with the 2D ground truth. Our method avoids the small angle problem that perspective projection usually suffers from. Experiments on several datasets validated the effectiveness and generalization ability of the proposed method qualitatively and quantitatively.

3

EGOCENTRIC 3D HUMAN POSE ESTIMATION FROM THE FISHEYE CAMERA

3.1 INTRODUCTION

Egocentric fisheye camera is used for human pose estimation or action recognition in different computer vision applications such as virtual reality (VR) or augmented reality (AR). These applications generally use a head mounted display to transform the user in a virtual world from a first-person viewpoint. Due to the large field of view, pose estimation from the egocentric fisheye viewpoint has many other valuable applications, such as robotics.

Current approaches focus on human pose estimation using pinhole cameras. These methods show significant progress for different benchmarks, such as the Human3.6M [64] and MPI-INF-3DHP [127] datasets. To reduce the ambiguity, many methods estimate the root-relative 3D joint positions in camera coordinates. However, the problem of estimating the egocentric 3D pose for a fisheye camera is to predict the 3D human pose from a first-person viewpoint possibly subject to strong image distortions. These distortions may negatively influence the 3D poses when the 2D poses on the image plane pass through the line of sight of the fisheye lens. For example, as shown in Figure 2, two different 2D poses which are subject to different levels of image distortions correspond to the same 3D pose. Recent works [180, 205] propose methods for 3D human pose estimation from images captured by a fisheye camera to alleviate the problem of self-occlusion. However, their methods ignore the negative influence of the distortions on the 3D pose estimation.

To mitigate the effect of distortions on the 3D human pose estimation, we propose an automatic calibration module with self-correction to regularize 3D predictions. The proposed calibration module automatically estimates the intrinsic and distortion camera parameters with self-correction instead of using a post-processing step [205] to enforce the 3D predictions to be consistent with the corresponding distorted 2D poses. In this way, the effect of distortions on 3D pose estimation is alleviated. To assess the effectiveness of the proposed automatic calibration module, we modified the *x*R-EgoPose dataset [180], a recent public dataset for 3D human pose estimation collected by a fisheye camera, by adding different levels of image distortions. We show that our method outperforms previous state-of-the-art methods and significantly improves the performance by using the proposed automatic calibration.

The contributions of our approach are summarized as follows:

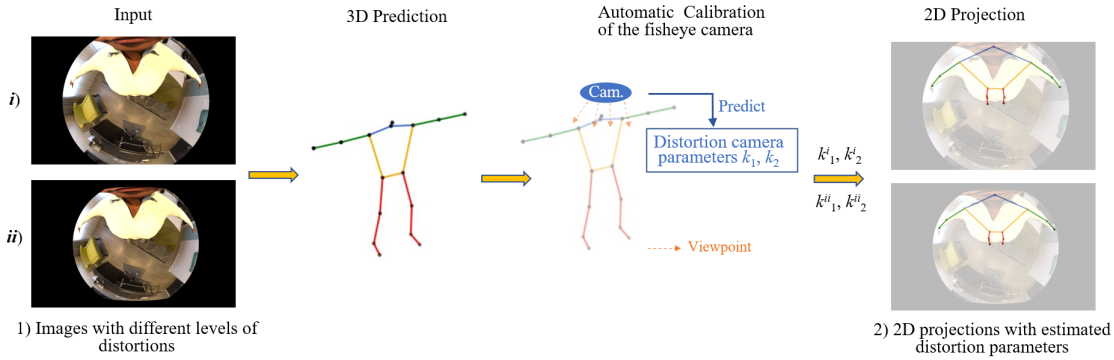


Figure 9: 3D pose prediction from a single image captured by a fisheye camera and 2D projection generated by our method. Note that although image $(\boldsymbol{i})$ and $(\boldsymbol{ii})$ appear different, they correspond to the same 3D pose. The proposed automatic calibration module alleviates the negative impact of image distortions on the 3D human pose estimation.

- We propose a method for egocentric 3D human pose estimation from a single image captured by a fisheye camera.
- We introduce an automatic calibration module with self-correction to mitigate the effect of image distortions for robust 3D human pose estimation.
- Our network shows state-of-the-art performance on the modified *x*R-EgoPose dataset containing images with different levels of distortions.

3.2 RELATED WORK

In this section, we describe monocular 3D human pose estimation methods from a third-person viewpoint, a first-person viewpoint and wearable motion sensors.

Third-person 3D Pose Estimation. Monocular 3D human pose estimation using external cameras shows significant progress with the use of CNNs and with the availability of large-scale 2D [2, 70, 109] and 3D [64, 127] human pose datasets. In general, existing methods are categorised into two types: (i) Direct 3D human pose estimation from images with full supervision [98, 143, 174, 176, 234] and (ii) 3D pose estimation from intermediate 2D pose predictions [16, 42, 90, 126, 133, 133, 142, 190, 199, 227]. As direct pose estimation methods rely on extensive training data with 3D annotations, their generalization capability is limited. To mitigate the above problem, approaches attempt to create synthetic datasets based on Motion Capture (MoCap) systems [18, 183]. Nonetheless, differences still exist between synthetic images and real images, such as backgrounds, appearance and variety of details. On the other hand, using robust 2D pose detectors, 3D pose estimation methods decouple the task into 2D pose prediction and 3D pose lifting step. To reduce the requirement of 3D pose annotations, [173, 233] propose geometric constraints to regularize the 3D estimations. The human pose dataset with only 2D annotations is used to constrain the 3D predictions.

First-person 3D Pose Estimation. A number of methods based on egocentric cameras focuses on hands, arms or torso detection [153, 210]. However, it is considerably more challenging to estimate the full 3D human pose from egocentric cameras. Jiang *et al.* [67]

propose a method for 3D pose estimation based on videos taken from chest-mounted cameras by considering the motion of the surrounding scene. However, the predictions are less accurate and have low confidence. Rhodin *et al.* [149] present an approach for full human body reconstruction captured from a head-mounted camera pair. Only recently, egocentric monocular 3D human pose estimation based on fisheye cameras is proposed. Xu *et al.* [205] design a new head-mounted system, where a fisheye camera is placed at the rig of a standard baseball cap. To reduce the error of the lower body, their methods take two images — one original image and one 2 × zoomed central part of the original image, as input to compute the 3D pose estimation. Tome *et al.* [180] propose an auto-encoder with two branches for egocentric 3D human pose estimation based on a fisheye camera. However, their methods assume images with the same distortion and therefore ignoring the negative impact of different levels of distortions on 3D human pose estimation.

3D Pose Estimation from Wearable Motion Sensors. Inertial Measurement Units (IMUs) are used to perform 3D pose estimation from a first-person viewpoint. However, a large number of sensors may cause the system to become intrusive and require more time to calibrate. Using less sensors becomes more challenging to reconstruct the 3D human pose in this configuration [184]. Shiratori *et al.* [161] introduce an alternative way to estimate the 3D human pose by structure-from-motion (SfM), with 16 cameras mounted at the human body joints. Nonetheless, this approach is difficult to use in real scenes due to motion blur, self-occlusion of limbs and missing textures in the background.

3.3 AUTOMATIC CALIBRATION OF FISHEYE CAMERAS

The inherent problem of image distortions captured by fisheye cameras makes 3D human pose estimation challenging. Two images may correspond to the same 3D pose when 2D different poses on the image plane pass through the line of sight of the fisheye lens. Therefore, it is difficult to regress 3D human joint positions in camera coordinates without the distortion parameters. To alleviate this problem, we propose an automatic calibration module to enforce the 3D predictions to be consistent with the corresponding distorted 2D poses.

3.3.1 *Fisheye Camera Model*

As shown in Figure 10.1, the human pose is represented by a set of joints $\boldsymbol{J}_i = [X_i, Y_i, Z_i, 1]^T$ located in the camera coordinate system. For the fisheye lens, shown in Figure 10.2, the angle of refraction from 3D locations $\boldsymbol{J}$ in Figure 10.1 is decreased from θ to θ_d. Then, the joint location $\boldsymbol{J}$ is projected on the image plane by $\boldsymbol{j} = [x, y, 1]^T$ in Figure 10.3. Particularly, the projected joint $\boldsymbol{j}_o = [x_o, y_o, 1]^T$ represents the projection based on the pinhole camera model. It is because of the distortion that positions $\boldsymbol{j}$ and $\boldsymbol{j}_o$ are different.

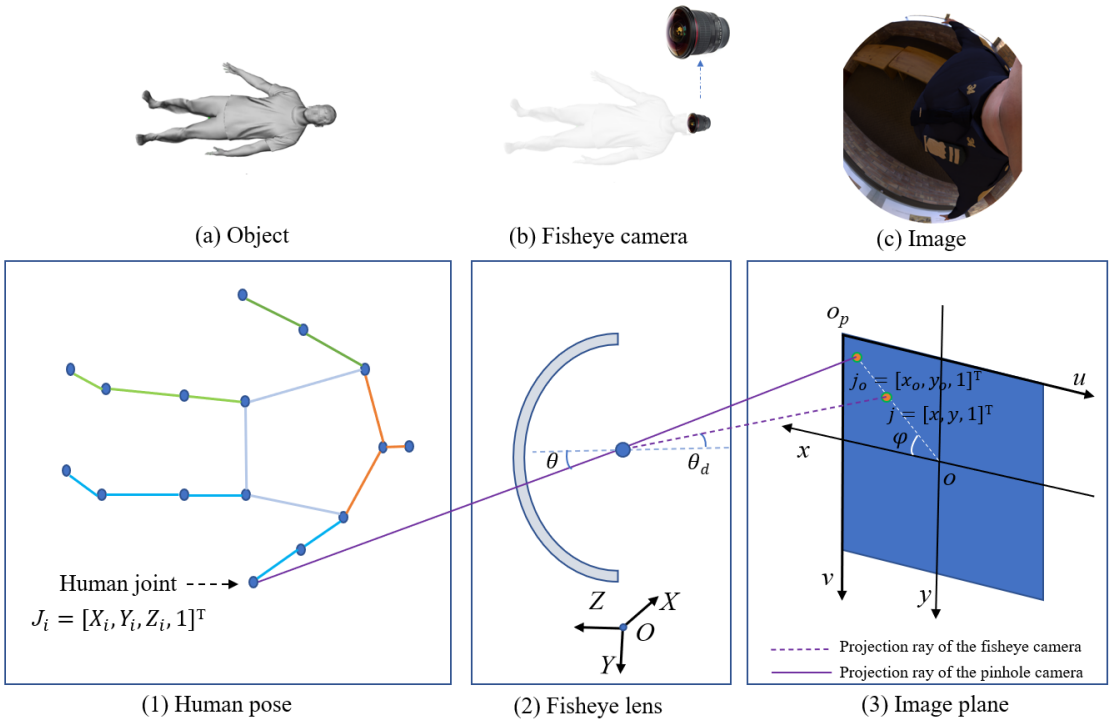


Figure 10: The imagery model of 3D-to-2D projection. The object — human joints $\boldsymbol{J}_i$ are located in camera coordinates $OXYZ$; The projected 2D pose (hand joint as an example) with the pinhole camera $\boldsymbol{j}_o$ and the fisheye camera $\boldsymbol{j}$ is on the image plane oxy; o_puv representing pixel coordinates. θ and θ_d indicate the angle of incidence and refraction with the fisheye lens respectively. φ represents the angle between the projected ray $\vec{oj}$ and x axis on the image plane.

3.3.2 *Egocentric 3D Pose Estimation under a Fisheye Camera*

From a single 2D image to 3D pose. 3D human pose estimation from a single image is an ill-posed geometric problem: there is no depth information. Previous methods attempt to solve this problem by learning the relation between 2D and 3D poses in a data-driven manner. However, with strong image distortions introduced by a fisheye camera, 3D human pose estimation is more challenging.

To alleviate the above issues, we propose an automatic calibration module to regularize 3D predictions. Instead of using a post-processing method [205] or ground truth, the proposed module automatically predicts the distortion camera parameters with self-correction. This is the first attempt to perform egocentric 3D human pose estimation by using automatic calibration of the fisheye camera.

From 3D pose to 2D pose. For a fisheye camera mounted on the head, the relative depth of human joints is comparable to the distance between the camera and the human joints. Therefore, weak perspective projection can not be used to approximate the 2D projections [51, 72, 185]. The 3D-to-2D projection process for the fisheye camera is illustrated in Figure 10.

Let $\mathbf{P}_{3D} = [\boldsymbol{J}_1, \boldsymbol{J}_2, ..., \boldsymbol{J}_n]$ denote the human joint locations in camera coordinates $OXYZ$, where n is the number of human joints and $\boldsymbol{J}_i = [X_i, Y_i, Z_i, 1]^T$. The projected 2D pose from the fisheye camera and pinhole camera is defined by $\mathbf{p}_{2D}$ and $\mathbf{p}_{o2D}$, a 3 by n matrix with $\boldsymbol{j}_i = [x_i, y_i, 1]^T$ and $\boldsymbol{j}_{oi} = [x_{oi}, y_{oi}, 1]^T$ respectively.

Given the intrinsic ($\boldsymbol{K}$) and extrinsic ($\boldsymbol{R}$ and $\boldsymbol{T}$) camera parameters, the 2D projections $\mathbf{p}_{\text{o2D}}$ under the pinhole camera model is as follows:

$$s \cdot \mathbf{p}_{\text{o2D}} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{\text{3D}}. \tag{3.1}$$

where the extrinsic camera parameters $\boldsymbol{R}$ and $\boldsymbol{T}$ are the identity matrix, s represents the scale factor and is equal to the Z value of the corresponding 3D joints in camera coordinates.

As the fisheye lens produces strong image distortions compared to a pinhole camera, distortion matrix $\boldsymbol{D}$ needs to be considered to compute 2D projections from a fisheye camera:

$$s \cdot \mathbf{p}_{\text{2D}} = \boldsymbol{K}\boldsymbol{D}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{\text{3D}}. \tag{3.2}$$

In this chapter, $\boldsymbol{D}$ is defined by

$$\boldsymbol{D} = \begin{bmatrix} \theta_d/l & 0 & 0 \\ 0 & \theta_d/l & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.3}$$

where $l = \frac{\sqrt{X^2+Y^2}}{Z}$, and θ_d indicates the angle of refraction. In this chapter, we refer to [73, 182] to calculate the angle of refraction $\theta_d = \theta(1 + k_1\theta^2 + k_2\theta^4)$, where the angle of incidence $\theta = \arctan(l)$, and the number of radial distortion parameters to be estimated is set to two, *i.e.*, k_1, k_2.

Visually, the 2D projection $\boldsymbol{j}_o$ under the constraint of a pinhole camera is transformed to $\boldsymbol{j}$ for a fisheye camera in Figure 11 using Eq. (3.2) for the distortion camera matrix.

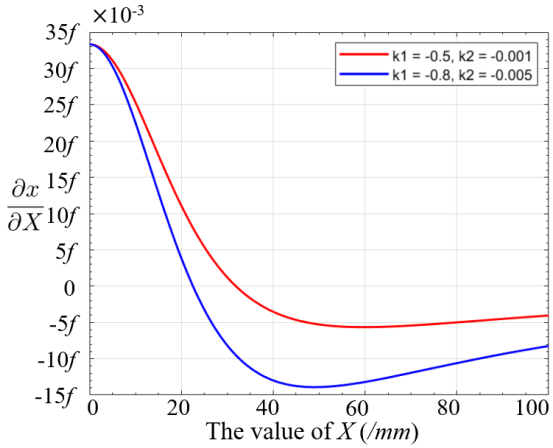
3.3.3 *Error Analysis of Estimated 3D Joints and 2D Projections*

Different from other methods for 3D pose estimation from external viewpoints, *i.e.,* outside-in approaches, in our task, the depth variance of human joints is comparable to the distance between the human joints and the fisheye camera. Therefore, the depth of 3D joint locations has an effect on the 2D re-projection error. Besides, the level of distortions and the distance of 3D joint locations to the optical axis (Z axis) also influence the 2D re-projection error.

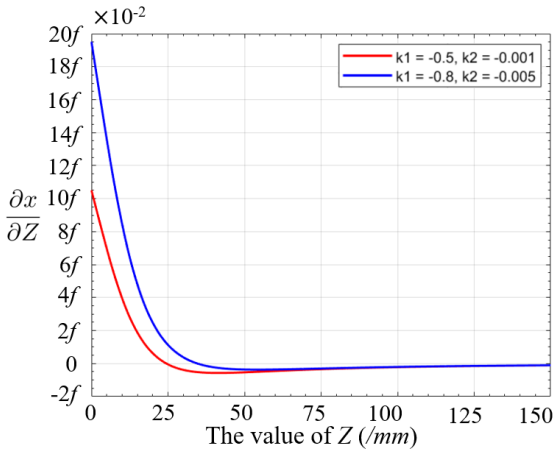
Eq. (3.2) can be detailed as follows,

$$x = f\frac{\theta_d}{l}\frac{X}{Z} = f(\theta + k_1\theta^3 + k_2\theta^5)\frac{X}{\sqrt{X^2+Y^2}}, \tag{3.4}$$

$$y = f\frac{\theta_d}{l}\frac{Y}{Z} = f(\theta + k_1\theta^3 + k_2\theta^5)\frac{Y}{\sqrt{X^2+Y^2}}. \tag{3.5}$$



1) The impact of the value of X and distortion parameters on reprojection error.



2) The impact of the value of Z and distortion parameters on reprojection error.

Figure 11: The impact of value of X and Z on the re-projection error under the fisheye camera with different distortion parameters k_1 and k_2. Due to the large range of hand and elbow joints, we plot the curve setting Z to be $30mm$ as shown in Figure 11.1. Since most joints such as shoulders, hips and knees have similar positions in the XY plane, we plot this curve setting $X = 25mm$ as shown in Figure 11.2.

Without loss of generality, we only study the influence of the level of distortions, the depth Z, and the value of X (the same to Y) on the 2D re-projection error. Before calculating the derivation of Eq. (3.4), we set Y to zero to simplify the formula, *i.e.*,

$$x = f\frac{\theta_d}{l}\frac{X}{Z} = f(\arctan\frac{X}{Z} + k_1\arctan^3\frac{X}{Z} + k_2\arctan^5\frac{X}{Z}). \tag{3.6}$$

The partial derivative of Eq. (3.6) is taken:

$$\begin{aligned}\frac{\partial x}{\partial X} &= f\frac{Z}{X^2+Z^2}(1 + 3k_1\arctan^2\frac{X}{Z} + 5k_2\arctan^4\frac{X}{Z}),\\ \frac{\partial x}{\partial Z} &= -f\frac{X}{X^2+Z^2}(1 + 3k_1\arctan^2\frac{X}{Z} + 5k_2\arctan^4\frac{X}{Z}).\end{aligned} \tag{3.7}$$

Figure 11 shows the impact of distortion parameters, the value of X and Z on re-projection error according to Eq. (3.7): 1) The value of X and Z have different influences on the re-projection error with various distortion parameters. Specifically, the larger the image distortions, the larger the influence of 3D locations on the 2D re-projection error. 2) Under the same level of distortions, the 3D joint locations with larger distances to the camera (such as ankles, toes and hips joints in lower body) or with larger X (such as elbows and hands joints) are expected to cause smaller errors on the 2D projections. In other words, the error of 3D poses is larger for joints with larger distances to the camera in the depth or larger distances to the optical axis under the same error of 2D projections.

3.3.4 *Self-correction for Calibrating the Fisheye Camera*

An automatic calibration module is proposed to regularize the 3D predictions. Our calibration module predicts the intrinsic camera parameters $\boldsymbol{K}$ and the distortion camera

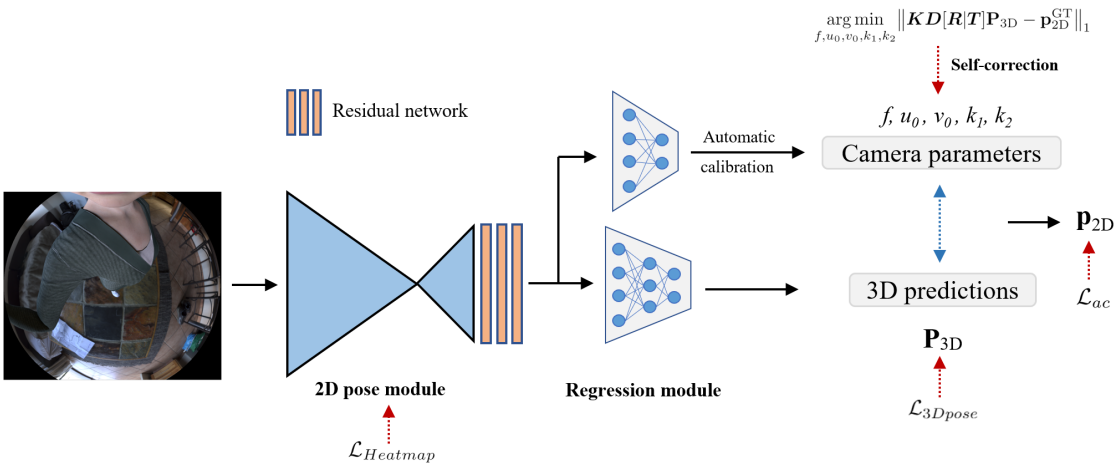


Figure 12: Overview of the proposed framework. We use *ResNet-50* as our backbone to detect 2D poses with heatmap representations. The 2D heatmaps are fed into a residual network with attention mechanism to further exploit the information in latent space. Then, we employ a series of fully connected layers to estimate the 3D pose in camera coordinates and camera parameters (*i.e.*, focal length f, principal coordinate u_0 and v_0 and distortion parameters k_1 and k_2). Finally, the estimated 3D pose are enforced to be consistent with 2D poses on the image plane using the predicted camera parameters.

parameters $\boldsymbol{D}$ automatically. Specifically, $\boldsymbol{K}$ includes focal length (f) and principal coordinates (u_0, v_0) while $\boldsymbol{D}$ contains the distortion parameters (k_1, k_2).

As discussed in Section 3.3.3, the re-projection error depends on the level of distortions, the depth, and the distance to the optical axis of the estimated 3D joints. Therefore, the commonly used $L2$ loss that constrains the camera parameters in the outside-in approaches [51, 185] cannot be used to update our automatic calibration module. The optimization process will focus on the upper body estimation, especially for neck and arm joints, due to the larger re-projection error. This may result in inaccurate estimation of hands, elbows and joints in lower body. We will verify this issue in Section 3.5.

To optimize our automatic calibration module, we minimize the absolute error between the projected 3D pose and 2D pose annotations $\mathbf{p}_{2D}^{GT}$. This avoids the optimization process to focus on the joints with larger re-projection errors:

$$\arg\min_{f,u_0,v_0,k_1,k_2} \left\|\boldsymbol{K}\boldsymbol{D}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{3D} - \mathbf{p}_{2D}^{GT}\right\|_1. \tag{3.8}$$

Note that the camera parameters (f, u_0, v_0, k_1, k_2) and $\mathbf{P}_{3D}$ are updated simultaneously.

3.4 NETWORK AND TRAINING DETAILS

Given a single image captured by a fisheye camera, our method aims to regress 3D human joint locations in camera coordinates. In this section, we introduce our network design and training strategy of our network.

Table 4: Comparison with existing methods on the modified *x*R-EgoPose dataset.

Approach	Gaming	Gesticulating	Greeting	Lower Stretching	Patting	Reacting	Talking	Upper Stretching	Walking	Average ↓
Martinez [126]	98.3	85.3	65.6	83.0	74.7	97.2	53.7	77.2	79.2	79.7
Ours (w/o $\mathcal{L}_{ac}$)	80.7	66.4	61.0	74.8	65.6	80.2	44.4	83.8	76.4	78.6
Ours	**75.3**	**66.0**	**54.1**	**68.7**	**65.4**	**78.3**	**43.0**	**67.4**	**69.2**	**67.7**

[1] Ours (w/o $\mathcal{L}_{ac}$) indicates that our method is trained without using the proposed automatic calibration module.

3.4.1 *Network Design*

Our framework consists of three modules as shown in Figure 12. First, we employ a 2D pose module to detect 2D heatmaps of human joint positions on the image plane. Second, a 3D pose regression module takes the fused features from 2D heatmaps as input to estimate 3D joint locations in camera coordinates. Finally, we use the proposed automatic calibration of the fisheye camera to enforce 3D predictions to be consistent with the corresponding 2D poses under the distortions.

2D Pose Module. Considering the accuracy and computational costs, we adopt *ResNet-50* followed by three deconventional layers as our 2D pose module. Given a single image with a resolution of 256×256, the 2D pose module infers 2D poses with heatmap representations $\mathbf{HM} \in \mathbb{R}^{16\times64\times64}$, where 16 indicates the number of human body joints with the space dimension of 64×64.

To train the 2D pose detector, we use the mean square error (MSE) to calculate the loss between the estimated $\mathbf{HM}$ and 2D ground-truth heatmaps $\mathbf{HM}^{\mathrm{GT}}$. The loss function is defined by:

$$\mathcal{L}_{Heatmap} = \sum_{h}^{H}\sum_{w}^{W}\left\|\mathbf{HM}_{(h,w)} - \mathbf{HM}^{\mathrm{GT}}_{(h,w)}\right\|_2, \tag{3.9}$$

where H and W indicate the resolution of the heatmaps. Specifically, we generate ground-truth heatmaps by using Gaussian distributions with kernel size of 13×13 and standard deviation of 2 pixels on each joint locations on the image plane.

2D-to-3D Regression Module. To regress the 3D human pose $\mathbf{P}_{3\mathrm{D}}$ in camera coordinates, we employ several residual blocks with fully connected layers followed by batch normalization, ReLU non-linearity and Dropout. Considering the inference time and prediction accuracy, we use two residual blocks for 2D-to-3D regression.

We optimize the 3D pose regression module by minimizing the MSE error between 3D predictions $\mathbf{P}_{3\mathrm{D}}$ and 3D pose ground truth $\mathbf{P}^{\mathrm{GT}}_{3\mathrm{D}}$. Given the dataset with the number of N samples, the loss function is defined by:

$$\mathcal{L}_{3Dpose} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{P}_{3\mathrm{D}\,(i)} - \mathbf{P}^{\mathrm{GT}}_{3\mathrm{D}\,(i)}\right\|_2, \tag{3.10}$$

where i represents the index of the training set.

Automatic Calibration Module. As shown in Figure 12, there are two branches in the regression module. The first branch is the lifting module regressing 3D locations of human joints while a multi-layer perception is employed in the second branch to perform automatic calibration of the fisheye camera. Specifically, the second branch estimates the

Table 5: Average error for per joint performed by our method with $L1$ and $L2$ loss on the modified xR-EgoPose dataset.

Joint	Error (/*mm*) ↓			Joint	Error (/*mm*) ↓		
	Ours	Ours_*L2*	Improvement		Ours	Ours_*L2*	Improvement
Head	32.90	27.03	-5.87	Neck	20.53	17.20	-3.33
Left Arm	31.91	35.61	**3.69**	Right Arm	33.23	36.97	**3.74**
Left Elbow	47.69	51.82	**4.13**	Right Elbow	52.22	58.24	**6.02**
Left Hand	82.99	93.83	**10.84**	Right Hand	85.49	100.19	**14.71**
Left Hip	56.86	65.03	**8.17**	Right Hip	56.77	65.13	**8.35**
Left Knee	79.33	87.54	**8.21**	Right Knee	79.87	89.84	**9.97**
Left Foot	100.31	110.84	**10.53**	Right Foot	103.21	115.41	**12.20**
Left Toe	109.39	119.61	**10.22**	Right Toe	110.73	122.09	**11.36**

[1] Ours_*L2* denotes our method use $L2$ loss to update the proposed automatic calibration module.

intrinsic camera parameters consisting of focal length (f), principal coordinate (c_x, c_y) and distortion parameters (k_1, k_2). Then we use Eq. (3.2) to obtain the 2D projections, where 3D predictions are constrained by the 2D poses under the distortions. In this way, the impact of image distortions on 3D human pose estimation is alleviated. In this chapter, automatic calibration module is only applied during the training phase.

As discussed in Section 3.3.3, the level of distortions, the depth and distance to the optical axis of estimated 3D joint locations have an influence on the errors of the corresponding 2D projections. Therefore, we minimize the absolute error (*i.e.*, $L1$ loss) between the projected 3D pose and 2D ground truth avoiding the optimization to focus on joints with large re-projection errors. The loss function is defined by:

$$\mathcal{L}_{ac} = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{K}\boldsymbol{D}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{3\mathrm{D}\,(i)} - \mathbf{p}^{\mathrm{GT}}_{2\mathrm{D}\,(i)}\right\|_1, \tag{3.11}$$

where $\boldsymbol{R}$ and $\boldsymbol{T}$ are the identity matrix.

3.4.2 *Training*

According to Eqs. (3.9) - (3.11), we train our full network by minimizing the following cost function:

$$\mathcal{L}_{pose} = \lambda_{HM}\mathcal{L}_{Heatmap} + \mathcal{L}_{3Dpose} + \lambda_{ac}\mathcal{L}_{ac}, \tag{3.12}$$

where λ_{HM} and λ_{rep} are loss weights to adjust the combination of the 2D heatmap loss, the 3D pose loss and the loss of automatic calibration module.

During training, we first pre-train the 2D pose module on the external perspective MPII dataset, because we found pre-trained 2D pose module obtains higher accuracy of 2D pose estimations for images captured by a fisheye camera. Then, we fine-tune the whole network on the modified xR-EgoPose dataset.

Table 6: Experimental results of our network on the modified *x*R-EgoPose dataset under less 3D ground truth.

Approach	3D ground truth	MPJPE(*/mm*) ↓
Martinez *et al.* [126]	100%	79.7
Ours	100%	67.7
Ours	**80%**	76.9

Table 7: Comparison with existing methods on *x*R-EgoPose dataset.

Approach	Gaming	Gesticulating	Greeting	Lower Stretching	Patting	Reacting	Talking	Upper Stretching	Walking	Average ↓
Martinez [126]	109.6	105.4	119.3	125.8	93.0	119.7	111.1	124.5	130.5	122.1
Tome [180]	56.0	50.2	44.6	51.1	59.4	60.8	43.9	**53.9**	57.7	58.2
Ours	**36.8**	**34.1**	**36.7**	**50.1**	**57.2**	**34.4**	**32.8**	54.3	**52.6**	**50.0**

3.5 EXPERIMENTS

Datasets. Recently, two datasets for egocentric 3D human pose estimation for a fisheye camera are released — *x*R-EgoPose [180] and Mo2Cap2 [205] datasets. Both datasets consist of a large number of frames of daily activities for different environments and lighting conditions. Considering images from current datasets with the same distortion, we modified the *x*R-EgoPose dataset using Eq. (3.4) and Eq. (3.5) to randomly add image distortions. For fast evaluation, the total number of images in the modified dataset is one-fifth of the total size in original *x*R-EgoPose dataset.

Evaluation Metrics. We use the Mean Per Joint Position Error (MPJPE) as the evaluation metric in the experiments. Note that we do not need to align the root joint for the evaluation as in the outside-in approaches.

Implementation Details. The proposed network regresses 16 human body joints including the head joint. The head joint is estimated based on the position of head-mounted display from 2D images. We first pre-train our 2D pose module on the MPII dataset [2] and then train our full network for 36 epochs on the modified *x*R-EgoPose dataset using Adam [78] for optimization. The learning rate is set to 5×10^{-4}. The model is trained on two GTX 1080ti GPUs with a batch size of 64. The weights in the overall loss function are set to $\lambda_{HM} = 10^7$ and $\lambda_{ac} = 50$.

Method Comparisons. To assess the effectiveness of our method, we conduct experiments on the modified *x*R-EgoPose dataset compared with Martinez *et al.* [126], a simple but effective 3D pose estimation method from external camera viewpoints. Furthermore, we evaluate our method on the *x*R-EgoPose and Mo2Cap2 datasets compared with current state-of-the-art methods [180, 205] for egocentric 3D human pose estimation for fisheye cameras.

3.5.1 *Evaluation on Modified xR-EgoPose Dataset*

Overall Performance. We first evaluate the proposed approach on the modified *x*R-EgoPose dataset. Since the existing methods [180, 205] do not release their code, it is

Table 8: Comparison with existing methods on indoor set of Mo2Cap2 dataset.

Approach	Walking	Sitting	Crawling	Crouching	Boxing	Dancing	Stretching	Waving	Average ↓
3DV'17 [127]	48.76	101.22	118.96	94.93	57.34	60.96	111.36	**64.50**	76.28
VNect [130]	65.28	129.59	133.08	120.39	78.43	82.46	153.17	83.91	97.85
Xu* [205]	38.41	70.94	94.31	81.90	48.55	55.19	99.34	60.92	61.40
Tome* [180]	38.39	61.59	69.53	51.14	37.67	42.10	58.32	44.77	48.16
Ours	**41.16**	**76.58**	**73.04**	**89.67**	**52.96**	**58.90**	**92.21**	71.55	**62.13**

[1] $*$ means the method uses extra information.

hard to make a fair comparison with them. Therefore, we compared our method with a state-of-the-art method [126] for 3D human pose estimation from external camera viewpoints.

Table 4 lists the experimental results showing that our method achieves the best performance in all activities, leading to an improvement of 15.1% in overall performance.

Effectiveness of Automatic Calibration Module. We perform an ablation study on the modified *x*R-EgoPose dataset to assess the influence of our proposed automatic calibration module. The MPJPE of all activities are reported in Table 4, in which Ours (w/o $\mathcal{L}_{ac}$) refers to the proposed method without automatic calibration module. Our method obtained better performance than Ours (w/o $\mathcal{L}_{ac}$) with a 10.9*mm* improvement. The results show the effectiveness of the proposed automatic calibration module.

Update Strategy of Automatic Calibration Module. As discussed in Section 3.3.3, the level of distortions, the depth, and the distance to the optical axis of 3D joint locations have an influence on the error of the 2D projections. Based on the error analysis, we employ the $L1$ loss to train our automatic calibration module instead of the commonly used in the outside-in approach — $L2$ loss. In this way, our update strategy avoids the optimization process to focus on the estimated 3D joints with larger 2D re-projection errors. Otherwise, an inappropriate update strategy of our automatic calibration module may lead to overfitting of these joints and a decrease in overall performance.

We conduct a comparative experiment on the modified *x*R-EgoPose dataset to validate this strategy. Particularly, we denote our method using $L2$ loss as Ours_$L2$. Table 5 reports the average error for each estimated joint and the improvement by our method. It is shown that the proposed method achieves better performance for each joint except head and neck joints. Note that the error of joints in 1) lower body, such as knee, foot and toes, and 2) joints with large distances to the optical axis, such as hands and elbows in the 3D space are reduced significantly by our method, which validates our assumption.

3.5.2 *Mixed 2D and 3D Ground Truth Datasets*

Another advantage of the proposed method is that our network can be trained on a mixture of 2D and 3D pose datasets. Due to our automatic calibration module, the estimated 3D pose can be partially constrained by the 2D ground truth, alleviating the needs of 3D ground-truth labels. We test our model on the modified *x*R-EgoPose dataset with 80% of 3D annotations while the 2D ground truth labels are available in the training phase.

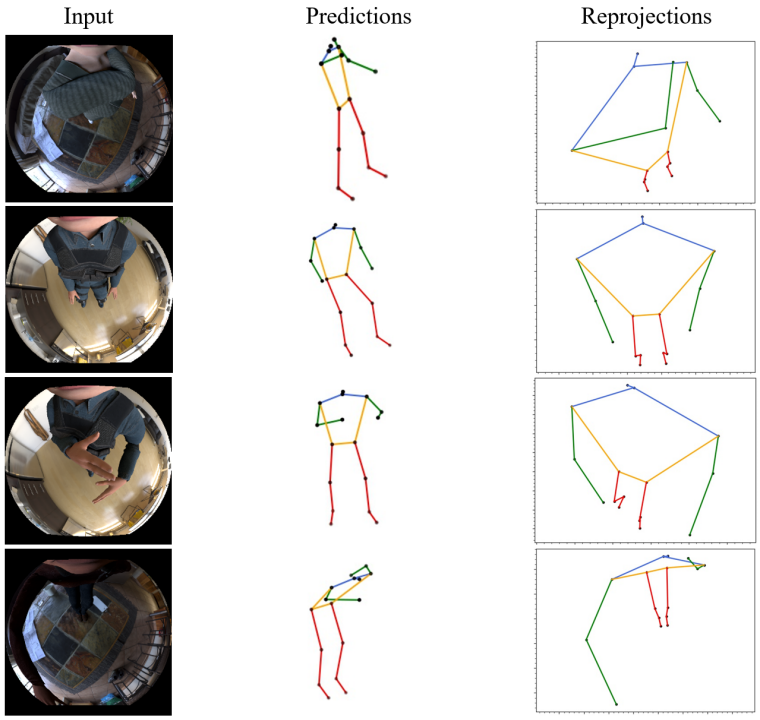


Figure 13: The visual results on the modified *x*R-EgoPose dataset predicted by the proposed method.

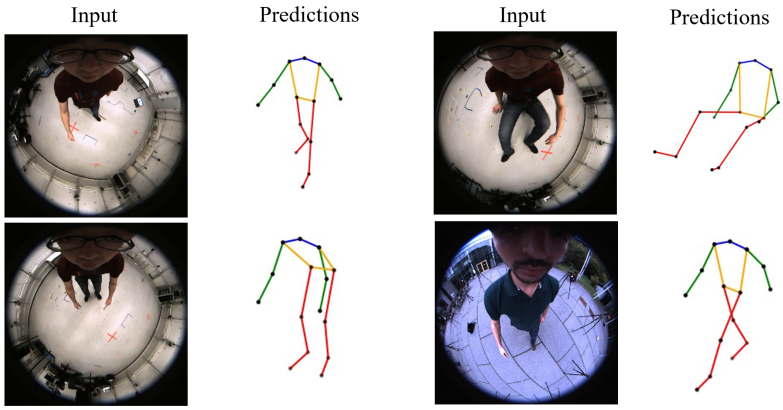


Figure 14: The visual results on Mo2Cap2 dataset predicted by the proposed method.

Table 6 lists the experimental results. Our method still outperforms Martinze *et al.* [126] (79.7*mm*) with an error of 76.9*mm*.

3.5.3 *Evaluation on Current Datasets*

Evaluations on *x*R-EgoPose Dataset. We also validate our method on the original *x*R-EgoPose dataset. Specifically, the proposed method is compared with Martinez *et al.* [126] and Tome *et al.* [180] — the state-of-the-art egocentric pose estimation method. Table 7 shows MPJPE on this dataset, including the error on each activity and the average error. Our method shows the best performance with an average error of 50.0*mm*, leading to an improvement of 14.1% on average compared to the state-of-the-art results.

Evaluations on Mo2Cap2 Dataset. We further compare our method with current methods on the Mo2Cap2 dataset. Table 8 shows the experimental results, where

3DV'17 [127] and VNect [130] focus on pose estimation from external camera viewpoint while Xu *et al.* [205] and Tome *et al.* [180] are the current state-of-the-art egocentric pose estimation methods. Note that Xu *et al.* (i) take two images — one original image and one $2\times$ zoomed central part of the original image to regress 3D poses while we only use a single image as input; (ii) need the toolbox for calibration of the fisheye camera to obtain distortion camera parameters while we directly estimate the distortion camera parameters with self-correction in our framework. On the other hand, Tome *et al.* uses the estimated 2D heatmaps from Xu *et al.* to implement the evaluation. From Table 8, the proposed method achieves competitive results with an error of 62.13*mm* on the indoor set of Mo2Cap2 dataset, even with only a single image as input.

3.6 CONCLUSIONS

We presented a novel method for egocentric 3D human pose estimation from a single image captured by a fisheye camera. To alleviate the impact of image distortions on 3D human pose estimation, we proposed an automatic calibration module to enforce the 3D predictions to be consistent with the corresponding 2D projections under the distortions. Experimental results showed that our method obtained state-of-the-art performance on the modified *x*R-EgoPose and current datasets compared with existing methods.

4

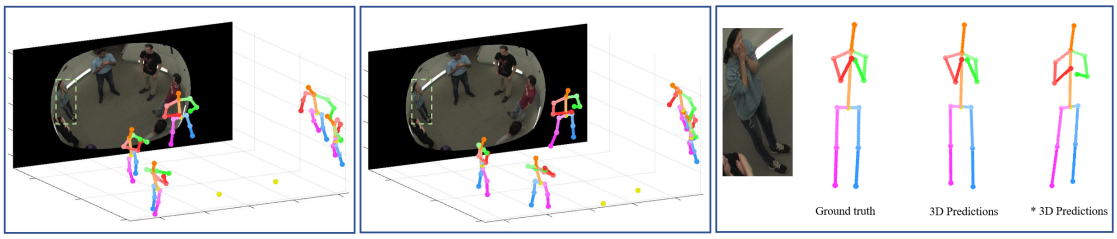
MULTI-PERSON 3D POSE ESTIMATION FROM THE FISHEYE CAMERA

4.1 INTRODUCTION

Due to the wide angle, fisheye cameras have been widely used in various practical scenarios such as video surveillance [77], virtual reality [149] and automotive applications [62]. 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 [28, 50, 131, 155].

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 [30, 51, 72, 82] use a re-projection method to establish a relationship between 2D and 3D poses with predicted scale and translation parameters, they 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 because the distance between humans and cameras is not fixed. Recent methods [179, 180, 187, 205] 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 [94, 107, 131, 230] 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 chapter, 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 module is proposed



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

Figure 15: 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.

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.

We evaluate the proposed approach on two public datasets including CMU Panoptic [71] and Shelf [5] 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.

4.2 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 [12, 55, 68, 80, 135, 138] simultaneously detect all human joints and then collect them for each person. Top-down approaches [22, 41, 132, 140, 172, 203] first employ a detector to predict bounding boxes of humans and then estimate a single 2D human pose from the cropped images.

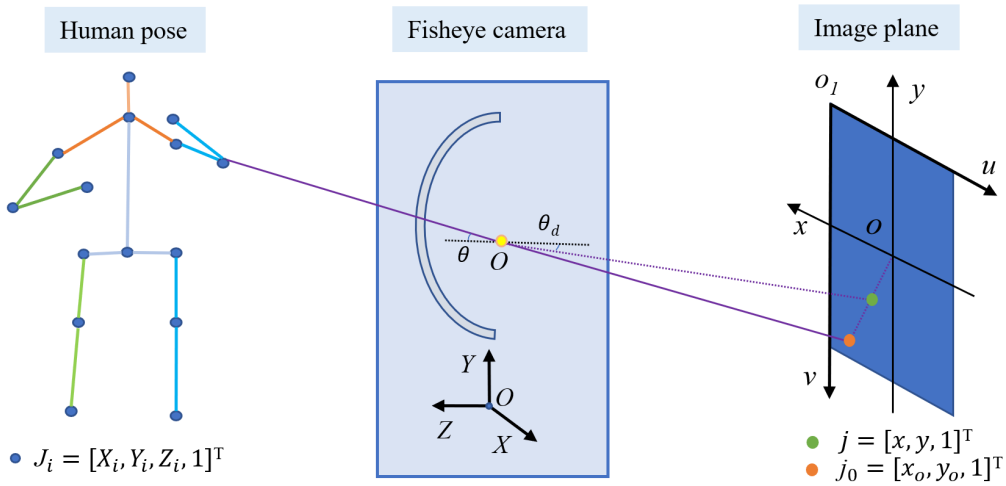


Figure 16: The process of 3D-to-2D projection for the fisheye camera model. This figure consists of a 3D human pose represented by a set of joints in camera coordinates $OXYZ$, a fisheye camera, and 2D projections on the image plane o_1uv. The angle of refraction θ is decreased to θ_d.

Multi-person 3D Pose Estimation. There are many methods [32,129,154,155,212,213] 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 [32,154,155] or correspondences between semantic representations [212] to obtain absolute joint locations in real spaces. Recently, some methods [94, 107, 131, 230] adopt the learning based manner to obtain absolute depths of root joints. [131] introduces 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. [107] considers 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 [24, 50, 94, 107, 131, 230] 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 procedure. On the other hand, existing methods mainly focus on pose estimation from a perspective camera or multi-view perspective images [36, 108, 200]. No research exits 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 [63, 67] or head [149, 180, 205]. Recently, [205] takes original and auxiliary images that focus on the lower body as inputs to improve the performance of egocentric pose estimation. [180] and [179] propose a method that includes two branches for 2D and 3D pose regression to estimate egocentric 3D poses. However, these methods are based on the root-relative single-person pose estimation and the camera is placed fixedly on the human head. Further, [29] proposes 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.

4.3 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.

4.3.1 *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 the appendix for more analysis). This makes multi-person 3D human pose estimation more challenging when camera parameters are not provided (known). In this chapter, 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.

Fisheye Camera Model. Figure 16 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.

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 $\mathbf{P}_{\text{3Dabs}} = [\boldsymbol{J}_1, \boldsymbol{J}_2, ..., \boldsymbol{J}_n]$ 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, $\mathbf{P}_{\text{3Drel}}$ denotes the root-relative human joint locations. Pelvises are defined as the root joint in this work. 2D projections $\mathbf{p}_{\text{2D}}$ and $\mathbf{p}_{\text{o2D}}$, 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 $\mathbf{p}_{\text{o2D}}$ under the perspective camera are obtained by:

$$s \cdot \mathbf{p}_{\text{o2D}} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{\text{3Dabs}}, \tag{4.1}$$

where s is a scale factor and is equal to the value of Z in $\mathbf{P}_{\text{3Dabs}}$. Because $\mathbf{P}_{\text{3Dabs}}$ 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. (4.1). Specifically, Eq. (4.1) is modified by adding a distortion matrix $\boldsymbol{D}$:

$$s \cdot \mathbf{p}_{\text{2D}} = \boldsymbol{K}\boldsymbol{D}\boldsymbol{I}_{3\times4}\mathbf{P}_{\text{3Dabs}}, \tag{4.2}$$

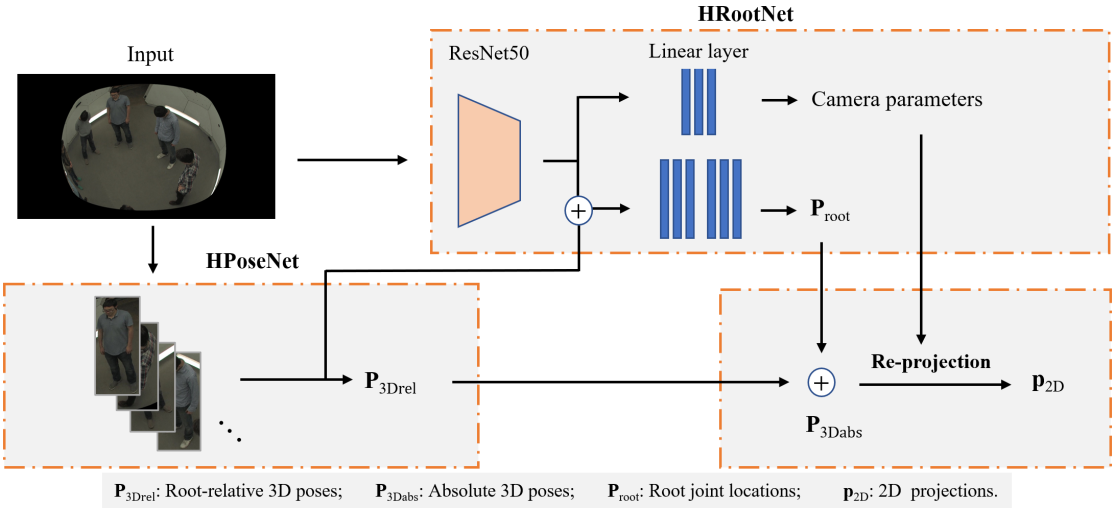


Figure 17: Overview of the proposed framework. There are two branches, *i.e.,* HPoseNet and HRootNet, to estimate root-relative 3D poses and absolute depths of root joints. Finally, we use a re-projection module to connect the two branches, enforcing the estimated 3D human poses to be consistent with the 2D poses under distortions by minimizing the re-projection error.

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

$$\boldsymbol{D} = \begin{bmatrix} \theta_d/l & 0 & 0 \\ 0 & \theta_d/l & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{4.3}$$

where $l = \frac{\sqrt{X^2+Y^2}}{Z}$. Following previous works [73, 182], the angle of refraction $\theta_d = \theta(1 + k_1\theta^2 + k_2\theta^4 + k_3\theta^6 + ...)$, where $\theta = \arctan(l)$, and two of distortion parameters (k_1, k_2) are used for simplification.

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 $\mathbf{P}_{3Dabs}$ and 2D ground truths $\mathbf{p}_{2D}^{GT}$.

$$\underset{f,c_x,c_y,k_1,k_2}{\arg\min} \left\| \boldsymbol{K}\boldsymbol{D}\boldsymbol{I}_{3\times4}\mathbf{P}_{3Dabs} - \mathbf{p}_{2D}^{GT} \right\|_1. \tag{4.4}$$

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.

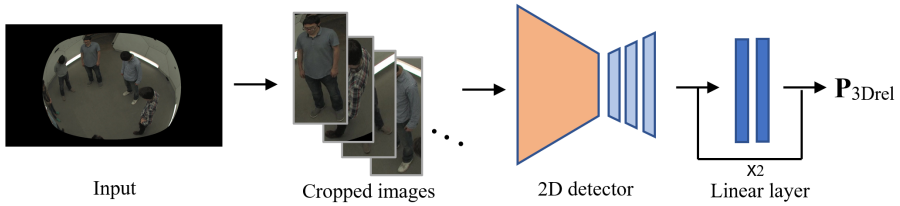


Figure 18: The design of our HPoseNet. HPoseNet takes images cropped by human bounding boxes as inputs to estimate root-relative 3D human poses.

Inspired by [107] and [227], 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.

4.4 NETWORK AND TRAINING DETAILS

We adopt a top-down pipeline to estimate multi-person 3D poses with absolute depths as shown in Figure 17. 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.

4.4.1 *Network Design*

HPoseNet. HPoseNet is to estimate root-relative joint locations for each person. Following [222], the design of HPoseNet is shown in Figure 18. 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 $\mathbf{P}_{\mathrm{3Drel}}$ and ground-truth 3D pose $\mathbf{P}^{\mathrm{GT}}_{\mathrm{3Drel}}$:

$$\begin{aligned} \mathcal{L}_{HM} &= \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{HM}_{(i)} - \mathbf{HM}^{\mathrm{GT}}_{(i)}\right\|_2, \\ \mathcal{L}_{3D} &= \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{P}_{\mathrm{3Drel}\,(i)} - \mathbf{P}^{\mathrm{GT}}_{\mathrm{3Drel}\,(i)}\right\|_2, \end{aligned} \tag{4.5}$$

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 [57] 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

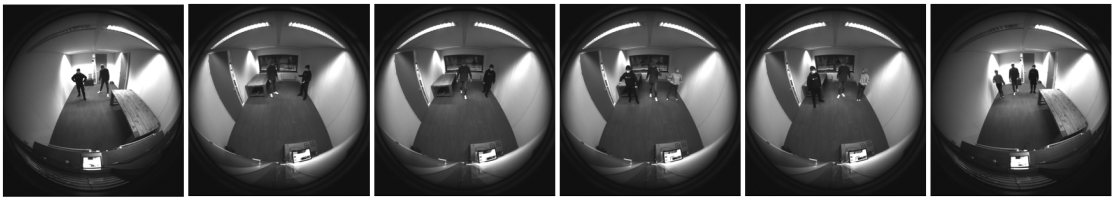


Figure 19: Example images in 3DhUman dataset. Actions from left to right: posing, talking and walking with 2 persons (first 3 pictures) and 3 persons (last 3 pictures).

optimize the MSE error between estimated root joint locations $\mathbf{P}_{\text{root}}$ and the ground truth $\mathbf{P}^{\text{GT}}_{\text{root}}$. The loss function is given by:

$$\mathcal{L}_{root} = \left\| \mathbf{P}_{\text{root}} - \mathbf{P}^{\text{GT}}_{\text{root}} \right\|_2. \tag{4.6}$$

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 scale changes caused by unknown distortions. The loss function is as follows:

$$\mathcal{L}_{rep} = \frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{KDI}_{3\times4} \mathbf{P}_{\text{3Dabs}\,(i)} - \mathbf{p}^{\text{GT}}_{\text{2D}\,(i)} \right\|_1. \tag{4.7}$$

4.4.2 *Training*

According to Eqs. (4.5) - (4.7), the overall loss function is given by:

$$\mathcal{L}_{pose} = \lambda_{HM}\mathcal{L}_{HM} + \mathcal{L}_{3Drel} + \lambda_{rep}\mathcal{L}_{rep} + \lambda_{root}\mathcal{L}_{root}, \tag{4.8}$$

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

4.5 EXPERIMENTS

4.5.1 *Experimental Setup*

Current datasets. We use CMU Panoptic [71] and Shelf [5] 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. (4.2). 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

Table 9: 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. [131]	100.26*	96.79*	99.88*	125.09*	102.83*	783.42
Lin et al. [107]	100.26*	96.79*	99.88*	125.09*	102.83*	367.47
Ours	**79.98**	**55.26**	**79.12**	**80.26**	**66.76**	**182.94**

Methods	MPJPE ↓	MRPE ↓
Moon et al. [131]	300.18*	696.10
Lin et al. [107]	300.18*	793.11
Ours	**132.45**	**589.19**

* As [131] and [107] use the same architecture for root-relative 3D human estimation and [107] does not release code for this part, the values of MPJPE are considered to be the same.

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 Figure 18. 2D/3D annotations and camera parameters are given by this dataset. Following [71], 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 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) [131] 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 [107, 131]. 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 [180] and [205].

Following existing approaches [32, 94, 107, 131], we first attempt to use Mask R-CNN [52] to detect each person in the input image. However, it fails to detect accurate

Table 10: 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**

Table 11: MPJPE on the Human3.6m dataset [64].

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

bounding boxes for each person. To avoid the influence of the person detector, ground-truth bounding boxes are used for evaluation.

4.5.2 *Quantitative Evaluation*

Modified CMU Panoptic Dataset. We first compare our approach with existing state-of-the art methods [107, 131] on the modified CMU Panoptic dataset. Table 9 (bottom) lists experimental results including the MPJPE of four activities and MRPE. It is shown that our approach achieves the best performance and obtains 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 *et al.* [107]. Moon *et al.* [131] 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.* [131] fail to achieve desirable performance.

Modified Shelf Dataset. We then test all approaches on the modified Shelf dataset. Table 9 (top) 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 10 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.

Perspective Images. We compare our HPoseNet with [131] and [107] for 3D human pose estimation on perspective images. HPoseNet estimates root-relative 3D human poses. Therefore, Human3.6m dataset, a large-scale dataset and a commonly used benchmark

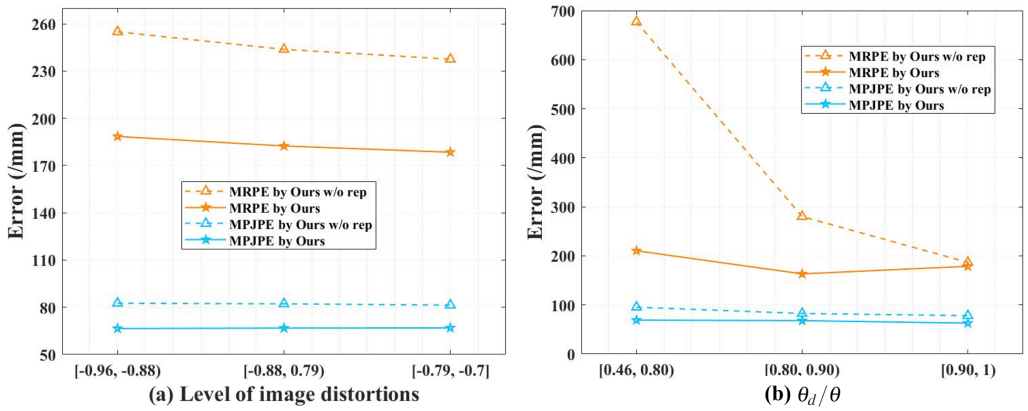


Figure 20: Analysis of the sensitivity of our method and our method without using proposed re-projection module (Ours w/o rep) on a) image level defined by image distortions and b) instance level defined by θ_d/θ for each human appeared in images on the modified CMU Panoptic dataset.

Table 12: 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**

for 3D human pose estimation, is used to evaluate all methods. Following the same setting as in [107, 131], 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 11 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.

4.5.3 *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 [51, 107, 233]. The experimental results are listed in Table 12.

From Table 12, 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

Table 13: 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

Table 14: 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**

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.

4.5.4 *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 Figure 20.

Image Level. We first analyze the sensitivity of our method on each image for different levels of image distortions shown in Figure 20 (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 Figure 20 (b).

It is shown in Figure 20 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.

4.5.5 *Discussion*

Performance for other fisheye camera settings. In this chapter, 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 4.5.1. The results are listed in Table 13. It is shown that: 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 over-fitting on our 3DhUman dataset, we additionally define training and testing sets by whether the image is taken by the same fisheye camera. Table 14 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.

4.6 CONCLUSIONS

In this chapter, 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 to be 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 have been further regularized. Compared with existing work, our method showed the state-of-the-art performance on both synthesized and real-world datasets.

4.7 APPENDIX

In this section, we provide additional experimental results and training details. Specifically, Section 4.7.1 gives training details on the modified CMU Panoptic and our newly collected 3DhUman datasets. In Section 4.7.2, we conduct experiments to analyze the effect of proposed re-projection module on sensitivity of different levels of image distortions. Additional visual results of applying our approach on modified CMU Panoptic and 3DhUman datasets are presented in Section 4.7.3.

4.7.1 *Training Details*

Please refer to Section 4.5.1 for the details of CMU Panoptic [71] and 3DhUman datasets. Here, we provide additional training details on these two datasets.

Modified CMU Panoptic. The training set in our experiments includes: 160224_mafia1, 160224_mafia2, 160224_ultimatum1, 160224_ultimatum2 and 160226_mafia2. The test set contains: 160226_haggling1, 160422_haggling1, 160226_mafia1, 160422_mafia2, 160422_ ultimatum1 and 160906_pizza1.
3DhUman. Following CMU Panoptic dataset, our dataset consists of three common activities: posing, talking and walking. To avoid the negative impact of grayscales on [131] and [107], all compared methods are first trained on modified CMU Panoptic dataset with grayscales and then finetuned on 3DhUman dataset.

For joint annotations, we manually labeled 2D poses in perspective images captured by pinhole cameras and converted them to 3D poses based on the corresponding depths obtained by the LiDAR cameras. We then projected 3D poses by fisheye camera parameters to obtain the distorted 2D poses. Particularly, images were captured simultaneously by pinhole, fisheye and LiDAR cameras.

4.7.2 *Analysis of Image Distortion*

Due to the existence of image distortions, persons at different locations on images may cause various distortion strengths. Therefore, even when the persons express different 2D poses, they may be originated from the same 3D pose. Here, we verify the robustness of our approach on images with various image distortions. Specifically, Figure 21 shows 3D predictions performed by our method from the same image with two levels of image distortions in modified CMU Panoptic dataset, where three paired examples are provided. It can be seen that our method generates reasonable visual results and is not affected by different levels of image distortions, even the person scale at the edge of the image changes significantly.

4.7.3 *Visual Results*

In this section, we present additional visual results of applying our method on modified CMU Panoptic and 3DhUman datasets shown in Figures 22 and 23.

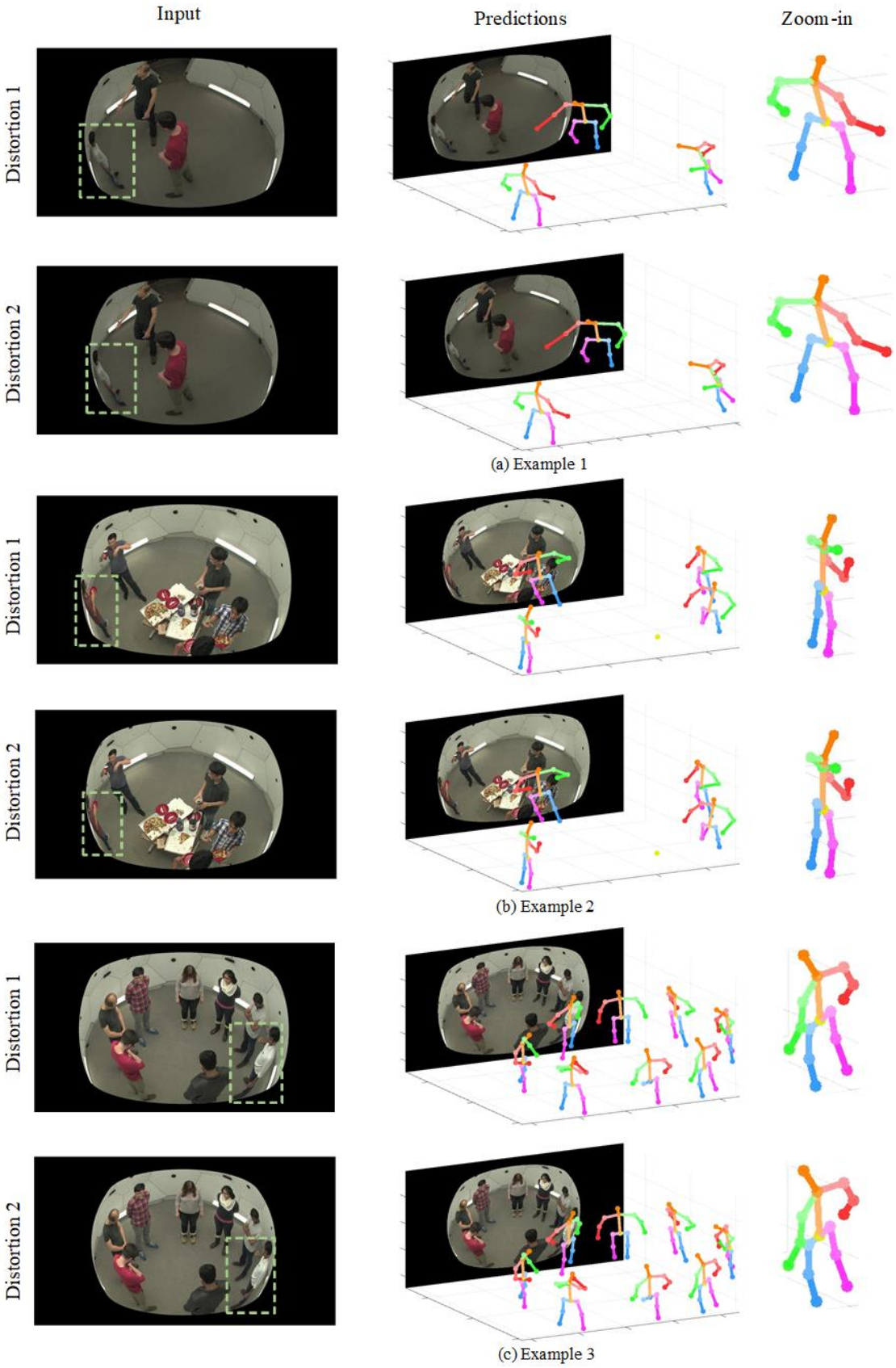


Figure 21: The visual results of applying our method on images with different levels of image distortions, where Distortion 2 is larger than Distortion 1. Notice scales of the same person with different levels of image distortions changed. Particularly, human scales change the most at the border of the image.

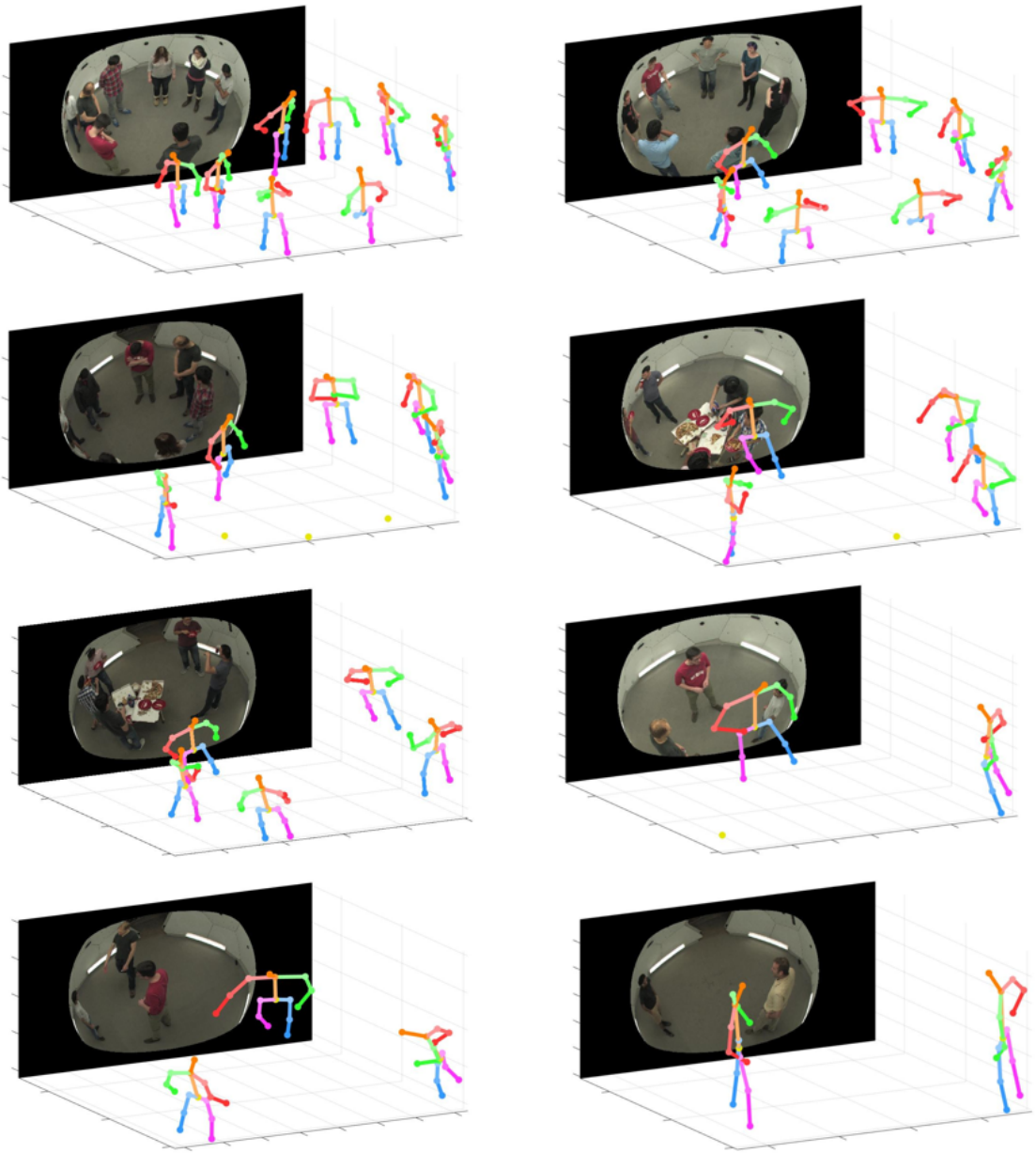


Figure 22: Visual results of applying our method on the modified CMU Panoptic dataset.

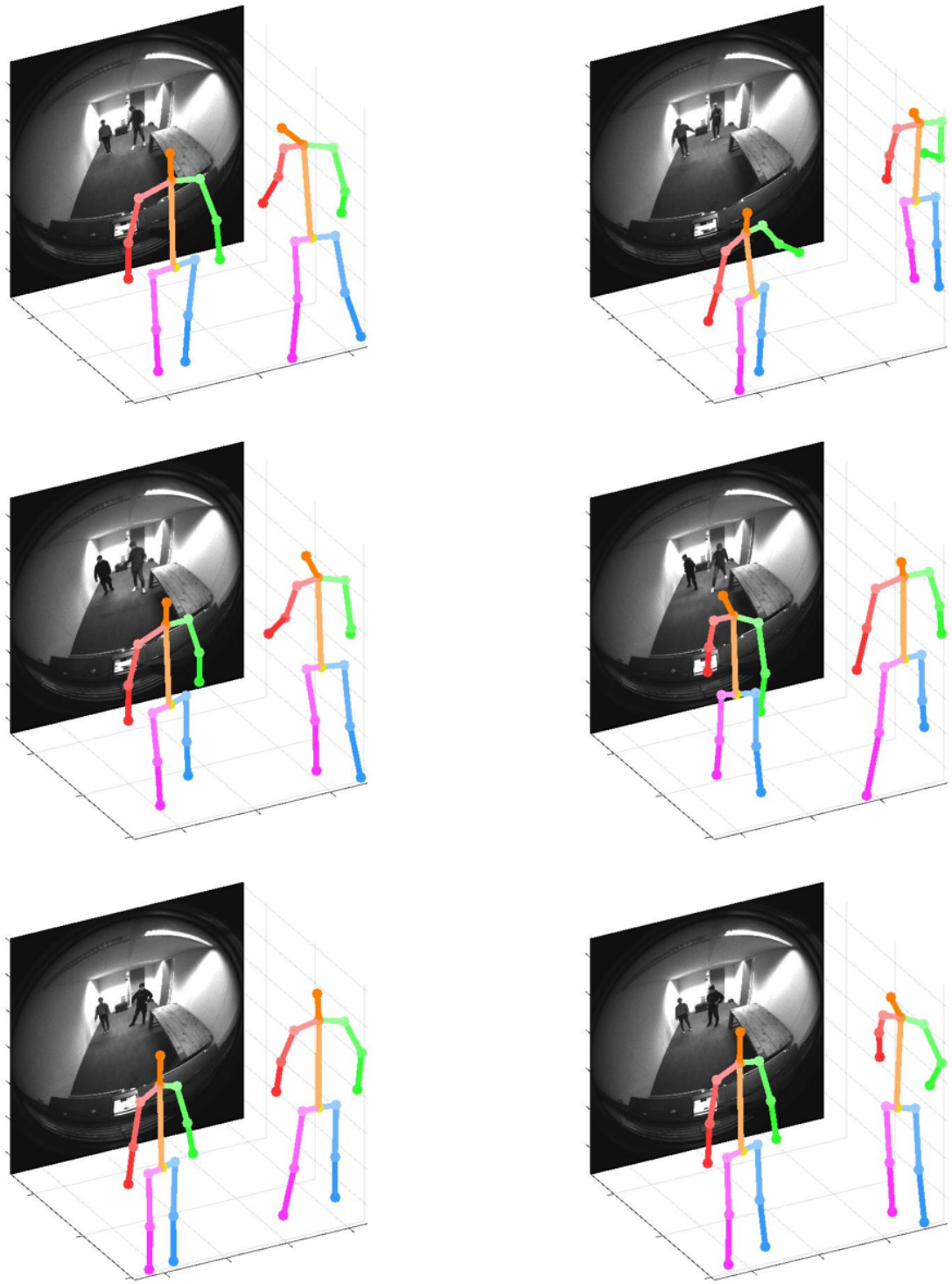


Figure 23: Visual results of applying our method on the 3DhUman dataset.

5

A BENCHMARK FOR 3D HUMAN POSE ESTIMATION & ACTION RECOGNITION

5.1 INTRODUCTION

3D human pose estimation (3D HPE) from a single image is an active research field in computer vision with many applications such as augmented reality (AR) [229], virtual reality (VR) [149], and human-robot interaction [195]. The goal is to infer 3D human joint locations from a single image. With the emergence of deep learning and large scale datasets, 2D human pose estimation made significant progress recently. However, 3D human pose estimation from a single image is (still) an ill-posed problem due to the inherent depth ambiguity and changing imaging conditions introducing variations in (human) appearance and self-occlusions. It is labor-intensive and time-consuming to annotate 3D labels to create 3D pose datasets. 3D pose datasets are usually collected in constrained environments limited by a motion capture device. Considering the difference between constrained and in-the-wild environments, it is still a challenge for existing methods to generalize well to in-the-wild images and unseen poses.

Human action recognition (HAR) is also a very active research field in computer vision with many applications such as medical diagnosis and security. The aim is to recognize actions performed by persons in videos. In general, HAR methods exploit (single) images, videos or skeletons as their input. In this chapter, we focus on skeleton-based methods. As it is hard and time-consuming to annotate human poses, HPE is beneficial for HAR. However, existing research mainly focuses on 3D HPE and HAR from perspective cameras. Due to a wider field-of-view of fisheye cameras, they are widely used in applications such as surveillance, photography, and sports. However, as shown in Tables 15 and 16, there are only a few existing methods that use fisheye cameras. Moreover, no datasets exist for fisheye-based 3D HPE and HAR. To this end, we propose a new dataset for multi-person 3D pose estimation (F-M3DHPE), and skeleton-based HAR (F-HAR) captured by a fisheye camera. There are several surveys on 3D HPE and HAR [20, 37, 83, 206]. However, they all focus on perspective cameras and ignore egocentric 3D HPE. To provide a complete picture of this important research area, our aim is to provide a comprehensive survey on the recent advances of 3D HPE and HAR for both perspective and fisheye cameras.

The remainder of this chapter is organized as follows. Camera models are given in Section 5.2. Section 5.3 and Section 5.4 provide a survey on deep learning (DL)-based single-/multi-person 3D pose estimation and human action recognition. In Section 5.5, a comparative study is conducted with existing methods on public datasets (perspective cameras) and on our newly collected dataset (fisheye cameras). Section 5.6 discusses

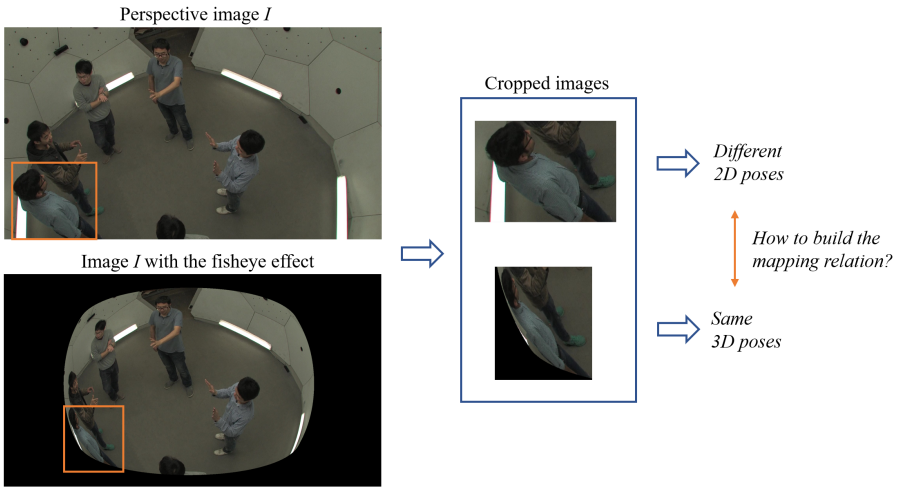


Figure 24: The challenge of fisheye images for 3D HPE. Although cropped images around the same human exhibit different appearances and 2D poses, they correspond to the same 3D poses.

Table 15: Summary of methods for 3D HPE considering cameras, viewpoints and the number of humans in the images.

Existence		Camera		Viewpoint		The number of person	
Method	Dataset	Pinhole	Fisheye	First-person (egocentric) view	Third-person view	Single person	Multi persons
✓	✓	✓			✓	✓	
✓	✓	✓			✓		✓
✓	✓		✓	✓		✓	
✗	✗		✓		✓		✓

Table 16: Summary of skeleton-based methods for HAR considering cameras and viewpoints.

Existence		Camera		Viewpoint	
Method	Dataset	Pinhole	Fisheye	First-person (egocentric) view	Third-person view
✓	✓	✓		✓	
✓	✓	✓			✓
✓	✓	✓			✓
✗	✗		✓		✓

future perspectives on 3D human pose estimation and skeleton-based human action recognition.

5.2 CAMERA MODELS

In this section, we outline the details of perspective and fisheye camera models.

5.2.1 *Definition*

Figure. 25 shows the projection of a 3D human pose onto a 2D pose. The human pose is represented by a set of discrete joints $\boldsymbol{J}_i$ in the camera coordinate system $OXYZ$. The middle part is a perspective or fisheye camera. Finally, the 2D projection of a joint is denoted as $\boldsymbol{j}_o$ for a perspective camera, otherwise it is represented by $\boldsymbol{j}$ where the angle of refraction is reduced from θ_d to θ.

Let $\mathbf{P}_{3D} = [\boldsymbol{J}_1, \boldsymbol{J}_2, ..., \boldsymbol{J}_n]$ denotes a 3D human pose in the camera coordinate system, where n is the number of human joints. The human joints $\boldsymbol{J}_i = [X_i, Y_i, Z_i, 1]^T$ are represented by a 4 by n matrix. $\mathbf{p}_{o2D}$ and $\mathbf{p}_{2D}$, denoted by a 3 by n matrix with $\boldsymbol{j}_o = [x_o, y_o, 1]^T$ and $\boldsymbol{j} = [x, y, 1]^T$, are the 2D projections of human poses using a perspective camera and fisheye camera model, respectively.

5.2.2 *Perspective Camera Model*

$\boldsymbol{j}_o$ in Figure. 25 is the 2D projection using a perspective camera model. In this setting, the angle of incidence and refraction is the same, *i.e.,* θ. According to the triangular similarity, the formulation is obtained by:

$$s \cdot \mathbf{p}_{o2D} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{3D}, \tag{5.1}$$

where $\boldsymbol{R}$ and $\boldsymbol{T}$ are extrinsic camera parameters representing rotation and translation parameters, respectively, $\boldsymbol{K}$ denotes the intrinsic camera parameters, and s is a scale factor and equals the depth in the camera coordinate system. As 3D joint locations $\mathbf{P}_{3D}$ are usually defined in the camera coordinate system, $\boldsymbol{R}$ and $\boldsymbol{T}$ are the identity matrices in Eq. (5.1).

5.2.3 *Fisheye Camera Model*

In contrast to perspective cameras, fisheye cameras contain wide-angle lenses capturing a wider field-of-view (FOV) i.e. capturing more of a scene, but cause image distortions. As shown in Figure. 25, 2D projections are displayed for a fisheye camera. Particularly, the farther the object is from the center of the image, the stronger the distortion will be.

For the fisheye camera model, the relationship between the angle of incidence and refraction, *i.e.,* θ and θ_d need to be considered to calculate the 2D predictions. Therefore, the distortion matrix $\boldsymbol{D}$, to compute 2D projections using Eq. (5.1), is given by:

$$s \cdot \mathbf{p}_{2D} = \boldsymbol{K}\boldsymbol{D}[\boldsymbol{R}|\boldsymbol{T}]\mathbf{P}_{3D}. \tag{5.2}$$

$\boldsymbol{D}$ is defined by:

$$\boldsymbol{D} = \begin{bmatrix} \theta_d/l & 0 & 0 \\ 0 & \theta_d/l & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{5.3}$$

where $l = \frac{\sqrt{X^2+Y^2}}{Z}$.

Based on [182] and [73], the angle of refraction is calculated by:

$$\theta_d = \theta(1 + k_1\theta^2 + k_2\theta^4 + k_3\theta^6 + ...), \tag{5.4}$$

where $\theta = \arctan(l)$.

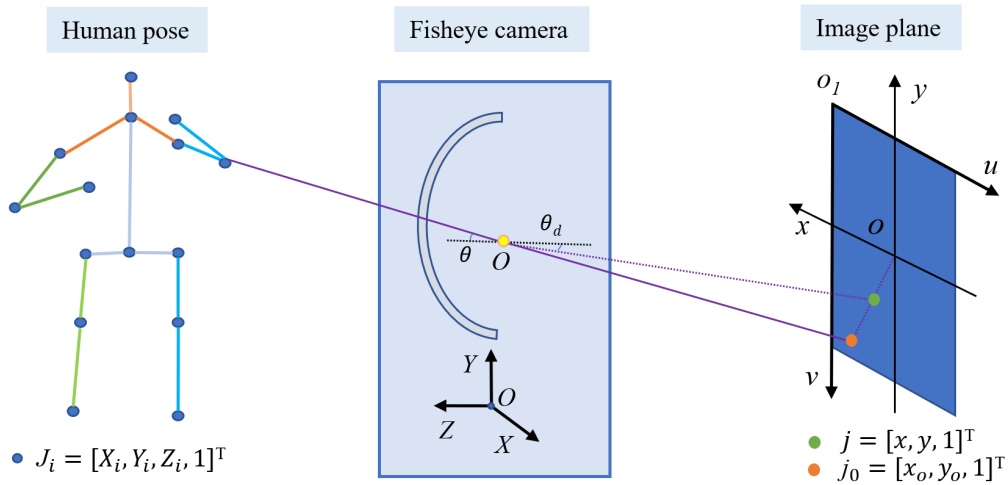


Figure 25: Projection from 3D joint locations in camera coordinates $OXYZ$ to 2D keypoints in image plane o_1uv [222]. This figure consists of 1) a human pose represented by a series of discrete joints J_i; 2) a camera; 3) an image plane. θ denotes the angle of incidence while θ_d represents the angle of refraction when there is a fisheye camera. j and j_o indicate 2D projections.

5.2.4 *Discussion*

The perspective camera model as defined by Eq. (5.1) depends on extrinsic and intrinsic camera parameters. For simplicity, we only consider the common case that the 3D joint locations are in the camera coordinate system. Hence, the projection only relies on the intrinsic camera parameters, *i.e.,* focal length and principal coordinates. The distortions parameters (fisheye lens) are defined in Eq. (5.3). For 3D human pose estimation from perspective cameras, existing methods [51] usually focus on cropped images to estimate the focal length and principal coordinates. However, 2D joint locations of the entire 2D image are needed when *i)* applying the fisheye projection and *ii)* estimating the distortion parameters when the camera parameters are not provided.

5.3 SURVEY ON 3D HUMAN POSE ESTIMATION

In this chapter, we focus on 3D human pose estimation from a single image. This task is ill-posed due to challenging imaging conditions such as inherent depth ambiguities, appearance changes and occlusions. Most of the existing methods perform 3D pose estimation from perspective cameras. Fisheye cameras (wide-angle lenses) capture a wider field-of-view (FOV) than perspective cameras. Therefore, fisheye cameras are widely used in various applications. In this survey, we focus on the following categories: single-person 3D pose estimation, multi-person 3D pose estimation, and 3D pose estimation from fisheye cameras.

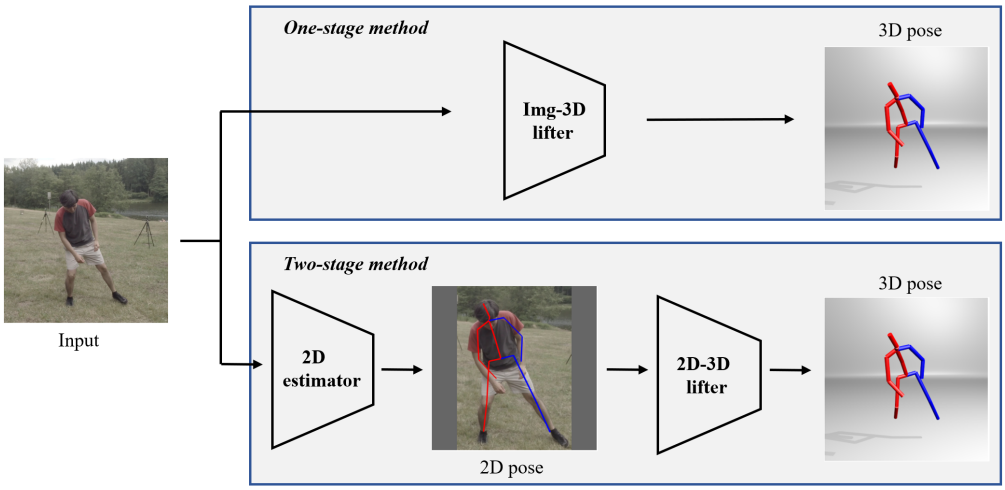


Figure 26: Overview of single-person 3D human pose estimation. Part of the figure is from [51] and [82].

5.3.1 *Single-Person 3D Pose Estimation*

Recently, single-person 3D pose estimation shows significant improvement with the use of deep learning and large-scale 3D human pose datasets. 3D single-person pose estimation can be divided into one-stage and two-stage approaches.

1) One-stage approaches

One stage approaches directly perform single-person 3D pose estimation from a single input image to 3D joint locations without intermediate supervision (*e.g.* 2D keypoints). Li and Chan [98] design a multi-task learning framework for 3D single-person pose estimation, where pose regression and body part detectors are simultaneously trained. Li *et al.* [101] propose a method taking 3D poses and images as input to compute the score if the two inputs match well. This method can be used as a refinement module for other methods. To alleviate the dependency on 3D poses and input images, Tekin *et al.* [176] use an auto-encoder to learn structured pose representations in latent space. The latent pose representations are beneficial for 3D human pose estimation. Sun *et al.* [173] present a regression-based method for 3D pose estimation, where the loss is based on the bone length and vector. Pavlakos *et al.* [143] propose to use volumetric representation for 3D human poses. They estimate the likelihoods of each voxel in discretized space around the subject. To reduce the computational cost, they adopt a coarse-to-fine scheme. Pavlakos *et al.* [142] suggest ordinal depth as extra information to infer 3D human joint locations. The method alleviates the restrictions that 3D pose annotations are prerequisite, and utilizes 2D pose datasets with ordinal depth annotations during the training process to improve the performance.

2) Two-stage approaches

Two-stage approaches first use off-the-self detectors to estimate 2D keypoints, and then infer 3D human joint locations from 2D predictions with/without features extracted from the input images.

Inference from 2D Poses to 3D Poses. Due to the robustness of 2D keypoint detectors, many methods focus on 3D pose estimation directly from the estimated 2D poses. Chen and Ramanan [16] propose a matching scheme to obtain 3D poses from a 3D pose

library by estimating 2D poses and depths from a single image. Martinez *et al.* [126] present a method by considering 2D keypoints followed by a series of fully connected layers to obtain 3D human joint locations in camera coordinates. Moreno-Noguer [133] formulates 3D human pose estimation from a single image as a regression problem of a 2D-to-3D distance matrix. Then, 3D poses are obtained by multidimensional scaling [7]. Tekin *et al.* [177] aggregate features extracted from original images and intermediate 2D poses with heatmap representations to infer 3D human joint locations. Sun *et al.* [174] introduce an integral operation (also referred to as soft-argmax) to obtain 3D human joint locations. Their framework is differentiable with the inference of image-2D poses (heatmap representations)-3D poses, reducing the quantization error caused by extracting keypoints from the heatmaps. Zhou *et al.* [232] utilize the representation of three joint heatmaps to learn the local relations between human body parts based on 2D keypoints and relative depth information. Jahangiri and Yuille [65], Sharma *et al.* [158], and Li *et al.* [91] generate multiple 3D pose candidates to solve the inherent depth ambiguity problem. Li *et al.* [93] present a regression-based method by exploring maximum likelihood estimation to deal with the uncertainty of the distribution for 2D and 3D human pose estimation. Chen *et al.* [24] propose an efficient method to search an optimized architecture for 3D human pose estimation.

Graph Convolutional Network (GCN)-Based Methods. Recently, researchers focus on GCN-based methods for 3D single-person pose estimation focusing on skeletons to obtain pose representations. Ci *et al.* [30] introduce a locally connected network combining GCN and a fully connect network for 3D human pose estimation. The aim is to alleviate the limitations of GCN on learning pose representations caused by a weight sharing scheme. Similarly, Zhao *et al.* [227] propose a semantic graph convolutional network (SemGCN) to deal with the limitations of GCN for the regression problem. SemGCN attempts to learn semantic relationships of nodes by combining 2D keypoints with features extracted from input images at 2D keypoints. Liu *et al.* [113] provide a comprehensive study on the impact of sharing weight schemes with feature transformations. They show that pre-aggregation in GCN, *i.e.,* applying transformations to the 2D input and then aggregating them, is beneficial for 3D human pose estimation. Xu *et al.* [204] propose a graph stacked hourglass network, aggregating multi-scale and multi-level feature information and 2D keypoints.

Geometric Constraint-Based Methods. Since public datasets for 3D human pose estimation are usually captured in an constrained environment, generalization is limited to deal with in-the-wild images. To this end, different methods propose geometric constraints to explicitly use information from 2D in-the-wild datasets. Zhou *et al.* [233] propose a geometric constraint based on the ratio of upper and lower limbs. In this way, the designed geometric loss is used to constrain 3D predictions when trained on a 2D in-the-wild pose dataset even without 3D annotations. Yang *et al.* [209] present an adversarial learning method for 3D human pose estimation. They adopt the network, introduced by Zhou *et al.* [233], as the generator and then employ a discriminator to enforce the 3D predictions to be plausible. A perspective camera model is an alternative way to provide relationships between 2D and 3D poses [17, 51, 185]. Habibie *et al.* [51] propose a method by explicitly using 2D and 3D features in latent space. They add constraints to ensure consistency between projected 3D predictions following the estimated camera parameters and 2D poses. Wandt *et al.* [185] propose a weakly

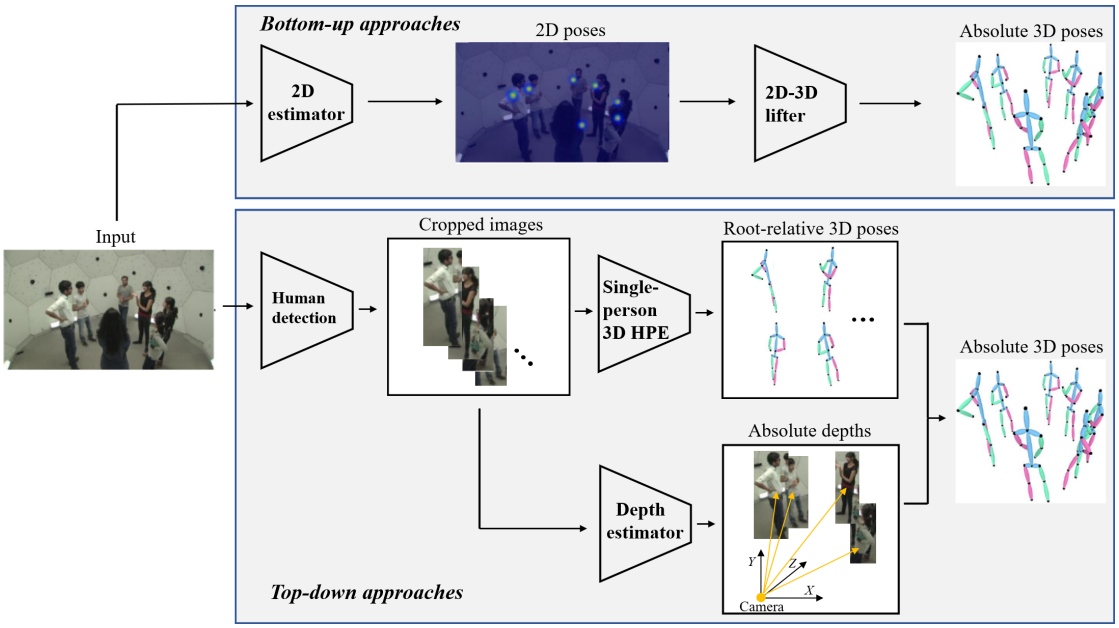


Figure 27: Overview of multi-person 3D human pose estimation. Part of the figure is from [230].

supervised method based on an adversarial re-projection network. Their method can be trained with unpaired 2D-3D poses and generalizes well to in-the-wild images. Drover *et al.* [38] and Chen *et al.* [17] attempt to regress 3D human joint locations from 2D pose landmarks. They use a discriminator to assess the projected 3D projections by minimizing the re-projection error. Zeng *et al.* [214] split the skeletons into local poses and recombine them to learn the pose representations with the aim to improve the generalization to unseen poses.

3) Comparison

The advantage of one-stage approaches is their efficiency. They directly estimate 3D poses without extra intermediate supervision. However, one-stage approaches heavily rely on fully annotated 3D pose datasets. In contrast, two-stage approaches can make use of in-the-wild images even without 3D annotations increasing their generalization capabilities. In general, two-stage approaches outperform one-stage approaches.

5.3.2 *Multi-Person 3D Pose Estimation*

A more general scenario is to estimate multi-person 3D poses. With development of single-person 3D pose estimation and related deep learning techniques, researchers recently pay more attention to multi-person 3D pose estimation. Generally, there are two categories: bottom-up approaches and top-down approaches. Furthermore, methods can be divided into optimization-based or learning-based methods depending on how the absolute depths are obtained. Optimization-based methods attempt to obtain absolute depths by minimizing the distance between projected 3D poses and 2D ground truths, while learning-based methods exploit the extracted features and geometric constraints to perform absolute depth estimation.

1) Bottom-up approaches

Bottom-up approaches firstly estimate all human joint locations followed by grouping all joints. Zanfir *et al.* [213] propose a bottom-up network (MubyNet) to perform 3D multi-person pose estimation as well as 3D shape estimation. MubyNet groups the joints and limbs based on 2D/3D information by limb scoring, transforming the grouping problem into an integer program. Metha *et al.* [129] propose occlusion-robust pose-maps to infer 3D poses by estimating 2D poses and Part Affinity Fields [12], even for strong occlusions. Metha *et al.* [128] propose a real-time approach (XNect) first estimating 2D and 3D pose features for all visible joints. Then, a fully connected neural work is used to infer 3D poses from 2D and 3D pose features. Finally, they use a space-time skeletal model to maintain temporal consistency. Zhen *et al.* [230] design a network aggregating various information cues to perform 3D multi-person pose estimation. They first regress 2.5D representations, *i.e.,* root depth map and part relative-depth maps. Then, 3D poses are inferred based on 2.5D representations and 2D keypoints.

2) Top-down approaches

Top-down approaches start by detecting each person and then performing singe-person 3D pose estimation to localize human joint positions. Rogez *et al.* propose LCR-Net [154] and LCR-Net++ [155] to estimate 2D/3D multi-person pose estimation. LCR-Net and LCR-Net++ consist of localization, classification and regression branches. The localization branch detects candidate regions for each person. The classification branch determines the pose classes and divides the detected poses into several 2D-3D anchor-poses. Finally, the 2D/3D poses are refined by the regression branch. Dabral *et al.* [32] present a Mask-RCNN based network for 3D multi-person pose estimation. They first detect 2D keypoints from each Region of Interest and then use a hourglass architecture to perform single-person 3D pose estimation from 2D keypoints. Finally, the absolute 3D poses are obtained by minimizing the distance of the projected 3D predictions and 2D keypoints. Recent methods adopt a learning-based manner to compute the absolute depths. Moon *et al.* [131] introduce a novel depth measure combined with a correction factor to obtain the depth. [50] adopts a data augmentation scheme to deal with the occlusion problem for depth estimation. Lin and Lee [107] use a similar architecture for 3D single-person pose estimation as [131]. To obtain the absolute depths, they consider the depth regression problem as a classification problem to perform depth estimation and localization of root joints. Wang *et al.* [186] utilize hierarchical multi-person ordinal relations (HMOR) to perform 3D multi-person pose estimation. HMOR encodes three levels of information: joint, body part and human instances. Finally, a coarse-to-fine strategy is adopted to infer the absolute depths. Cheng *et al.* [27] first use graph convolutional networks to infer multi-person 3D joint locations with absolute depths. Then, a temporal convolutional network is used to refine the 3D predictions by using temporal constraints. Cucchiara and Fabbri [31] discuss different ways to deal with the occlusion problem and perspective constraints caused by top-down viewpoint in video surveillance.

3) Comparison

In terms of computational costs, bottom-up approaches are more efficient than top-down approaches. The former makes a trade-off between computational cost and accuracy. The computational cost of top-down approaches depends on the detection module and the number of humans. Regarding the estimation accuracy, top-down approaches per-

form better than bottom-up approaches. The reason is that top-down approaches split the task into two subtasks, *i.e.,* human detection and 3D single-person pose estimation. The 3D pose estimator focuses on salient regions to predict 3D human poses. Recently, Cheng *et al.* [28] combine top-down and bottom-up networks to perform 3D multi-person pose estimation. They introduce an interaction-aware discriminator to integrate two kinds of 3D predictions from top-down and bottom-up network to obtain refined 3D predictions.

5.3.3 *3D Human Pose Estimation from Fisheye Cameras*

Fisheye cameras are widely used in various applications such as virtual reality, video surveillance and automatic driving. Surprisingly, there are only a few methods focusing on 3D human pose estimation from fisheye cameras. Based on the viewpoint of cameras, the methods are categorized into first-person view (egocentric) 3D HPE and third-person view 3D HPE.

1) Egocentric 3D HPE

Rogez *et al.* [153] and Yonemoto *et al.* [210] propose a method for hand, arm and torso pose inference from RGB-D data. Jiang and Grauman [67] infer full body 3D joint locations from cameras mounted on the chest with the aim to estimate unseen 3D poses. Rhodin *et al.* [149] first propose a method for full body 3D human pose reconstruction from a pair of fisheye cameras mounted on a helmet. Shiratori *et al.* [161] present an approach based on structure-from-motion (SFM) for 3D pose estimation from wearable devices, where 16 limb-mounted cameras are used.

To capture the 3D poses of full human body, Xu *et al.* [205] propose a disentangled method for egocentric 3D pose estimation from a camera mounted on a baseball cap. To improve the estimation performance of the lower body, occupying less area than the upper body, their framework consists of two branches to perform 2D pose estimation from an original image and an image of the central part. Finally, 3D poses are obtained by known camera parameters and estimated 2D keypoints. Tome *et al.* [180] design a novel two-stage method for egocentric 3D pose estimation from a camera mounted on user's head. They first detect 2D keypoints followed by an auto-encoder with two branches to obtain 3D human joint locations and 2D keypoints. To alleviate the requirement of ground-truth camera parameters, Zhang *et al.* [222] propose an automatic calibration method for egocentric 3D pose estimation. They introduce a re-projection module to predict the camera parameters. The influence of image distortions is alleviated by minimizing the error between projected 3D predictions and 2D keypoints. Wang *et al.* [187] perform a spatio-temporal optimization to compute smooth 3D human poses.

2) Third-person view 3D HPE

No methods exist for 3D HPE from third-person view for fisheye cameras. In egocentric settings, the distance between the camera and human head joint is fixed. In other words, the position of the head joint in the image is almost the same regardless of the human pose. Therefore, existing methods may fall short when applied to single/multi-person 3D pose estimation from a fisheye camera using a third viewpoint.

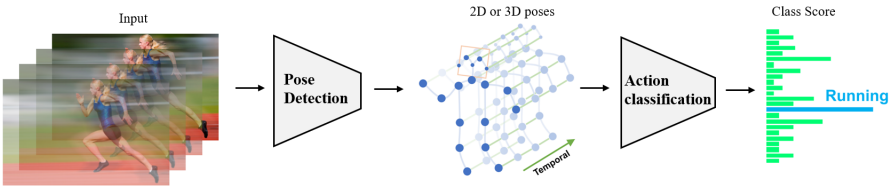


Figure 28: Overview of human action recognition based on skeletons. Part of the figure is from [207].

5.4 SURVEY ON HUMAN ACTION RECOGNITION

Human action recognition generally consists of two elements: action representation and classification. Action representation is provided by a feature extractor. The classifier is used to predict an action label from the extracted features. Regarding feature extraction, methods for human action recognition are divided into traditional and DL-based methods. Due to the progress of deep learning techniques, DL-based methods for human action recognition received more attention recently. In this chapter, we mainly focus on DL-based methods. DL-based methods can be grouped into video-based and skeleton-based methods. Video-based methods rely on the objects in the video. Skeleton-based methods focus on the change of 2D/3D skeletons detected in the videos. In this section, we focus on skeleton-based methods.

The input of skeleton-based methods is a sequence of 2D/3D human joint locations. Generally, skeleton-based methods can be divided into three categories: CNN-based, RNN-based, and GCN-based.

CNN-based Methods. 2D ConvNets achieve superior performance to extract features, while skeleton-based methods take human joint locations as their input [10, 21, 39, 56, 59, 76, 87, 89, 102, 115, 116, 123, 148, 202, 237]. To bridge this gap, skeleton sequences are converted into image formats by projecting 3D joint locations on three orthogonal planes [39]. Then, 2D ConvNets are applied to extract features from the new representation followed by a classifier for action recognition. Liu *et al.* [115] propose a view-independent approach for action recognition mitigating the impact of skeleton variations caused by different viewpoints. Ren *et al.* [148] present a multi-stream method based on 2D ConvNets for human action recognition. They focus on action-relation joints to reduce the impact of noise in converted images. Li *et al.* [87] transform 3D skeletons into translation-scale invariant images for action recognition to deal with variations of input skeletons. Ke *et al.* [76] transform skeleton sequences into cylindrical coordinates followed by 2D ConvNets to extract temporal dynamics. Chen *et al.* [21] propose a color-coding strategy to covert skeleton sequences into pseudo-color images, aiming to learn discriminate features for human action recognition. Li *et al.* [102] design several learning blocks of elastic units (Else-Net) to explore the relation between pose sequences.

RNN-based Methods. To exploit temporal information, RNNs are used to process sequential data. There are different methods based on RNNs for human action recognition [3, 40, 43, 100, 111, 112, 137, 157, 217, 221, 228]. Du *et al.* [40] and Shahroudy *et al.* [157] group the skeleton into five body parts to model temporal dynamics. Liu *et al.* [111] propose a LSTM-based method to extract features at both spatial and temporal domains, named ST-LSTM, by analyzing hidden sources of contextual information. Zhang *et al.* [221] design geometric features based on the relationships of all joints to learn

spatial information for action recognition. Zhao *et al.* [228] present a Bayesian-based LSTM method to learn spatial information from joints, temporal information from poses and variation of subjects. To alleviate the variation of sequence skeletons caused by various viewpoints, Zhang *et al.* [217] introduce a view adaptive module to transform automatically input skeletons into representations under a new viewpoint for action recognition.

GCN-based Methods. Since human skeletons can be seen as a graph structure, GCN-based methods received most attention for human action recognition. There are several types for GCN-based methods: modifications of GCNs, combination of GCNs and LSTMs, and multi-streams-based GCNs.

Firstly, there are methods to directly use GCNs to learn spatio-temporal representations for human action recognition [1, 46, 120, 141, 145, 193, 198, 208, 216]. [207] is the pioneer work to apply graph convolution networks to human action recognition. They propose a network named ST-GCN to model spatio-temporal relations from skeleton sequences. Li *et al.* [96] present an encoder-decoder GCN network to learn action-related representations named AS-GCN, where actional and structural links are used by the GCN to extract spatial features. AS-GCN achieves promising performance on both human action recognition and pose prediction. Ghosh *et al.* [48] extend the ST-GCN by applying stacked hourglass networks for action segmentation to increase the generalization ability.

Secondly, Graph RNN frameworks are used for human action recognition [19, 25, 26, 60, 88, 121, 146, 162, 197, 215, 219]. Si *et al.* [162] propose an attention mechanism using a LSTM architecture, named AGC-LSTM, to effectively extract spatio-temporal features for HAR. Chen *et al.* [19] propose a pooling strategy and a point-wise attention mechanism to learn action-related features from skeleton sequences. Instead of capturing skeleton differences between sequential skeletons, Ding *et al.* [34] focus on inherent differences in terms of spatio-temporal and context to determine the human actions.

Thirdly, multi-stream GCNs are proposed to fuse spatial and temporal features from different streams for human action recognition. For two-stream GCN-based methods, joints and bones are fed into GCNs to learn spatio-temporal representations [44, 45, 159, 160, 175, 181]. Multi-stream GCN, including more than two branches, use GCNs dealing with different cues (*e.g.,* joints, bones, motion and relative positions) to extract temporal information. Then, the extracted features are aggregated to perform human action recognition [23, 35, 92, 97, 103, 114, 169]. Chen *et al.* [23] exploits joints, bones, and corresponding motion to learn typologies for skeleton-based HAR.

In addition, there are different unsupervised methods to perform human human action recognition from skeleton sequences. Their common characteristic is to adopt an encoder-decoder structure to learn temporal information. Zheng *et al.* [231] propose an adversarial method based on an encoder-decoder structure to model temporal dynamics. Su *et al.* [171] present a fully-unsupervised method for human action recognition without the requirement of action labels. The aim is to weaken the role of decoders by enhancing the encoder to learn discriminative information.

5.5 DATASETS AND BENCHMARKS

Many datasets are available for 3D human pose estimation and human action recognition. In this section, we describe the publicly available datasets, our new dataset F-M3DHPE, the evaluation metrics, and the performance of existing methods on the public and proposed datasets.

5.5.1 *Dataset and Benchmarks for 3D Human Pose Estimation*

In this section, we provide a comparison of existing methods on public datasets. Particularly, we focus on single-/multi-person 3D pose estimation from a single image.

1) Datasets for 3D human pose estimation

There are different 3D pose datasets for single-person 3D pose estimation for perspective cameras including HumanEva [168], Human3.6M [64], and MPI-INF-3DHP [127]. For 3D multi-person pose estimation, CMU Panoptic [71], the MuCo-3DHP, and MuPoTS-3D [129] datasets are commonly used for quantitative evaluation. In addition, the Mo2Cap2 [205] and *x*R-EgoPose [180] datasets are recently released for egocentric 3D human pose estimation captured by a fisheye camera. The details of the datasets are listed in Table 17.

Human3.6M. Human3.6M is a large-scale 3D human pose dataset for single-person 3D pose estimation. The dataset contains 3.6 million human poses captured by a marker-based motion capture from 4 calibrated cameras at 50Hz in an indoor environment. There are 11 participants, 6 males and 5 females, performing 17 daily activities such as walking, posing, taking photo, sitting down, etc. The annotation includes 3D poses, body part labels, 3D body surface, depth information and bounding boxes. Generally, three protocols are used to train and test methods. *Protocol*#1 uses human subjects S1, S5, S6, S7 and S8 for training and human subjects S9 and S11 for testing; *Protocol*#2 adopts the same training and testing split as *Protocol*#1 but employs an alignment between predictions and ground truths for evaluation; *Protocol*#3 uses human subjects S1, S5, S6, S7, S8 and S9 for training and human subject S11 for testing with the alignment between predictions and ground truths.

MPI-INF-3DHP. MPI-INF-3DHP is collected by a marker-less motion capture for single-person pose estimation. The training set consists of more than 1.4M frames captured by 13 cameras. 3D pose annotations and universal skeleton are provided in this dataset. The test set includes both indoor and outdoor images with participants performing various activities such as standing, siting, exercise, sports, etc. Due to the existence of in-the-wild images and complexity, the test set is usually used to verify the generalization of existing methods (normally without fine-tuning on the training set).

CMU Panoptic. CMU Panoptic is used for multi-person 3D pose estimation from single-/multi-view image(s). It is captured by a multi-view system, where 480 synchronized cameras at 25Hz are installed at the surface of a dome with a radius and height of 5.59m and 4.15m, respectively. Various games, that people are playing, are included in this dataset such as haggling, ultimatum, mafia, playing musical instruments, etc. The number of participants ranges from 1 to 8. 3D poses, 3D trajectory stream and camera information are provided.

Table 17: Public dataets for 3D pose estimation.

Dataset	Year	Camera Type	Environment	Size	The number of persons
HumanEva [168]	2010	Pinhole	Indoor	6 subject, 7 actions, 40k frames	Single
Human3.6M [64]	2014	Pinhole	Indoor	11 subjects, 17 actions, 3.6M frames	Single
MPI-INF-3DHP [127]	2017	Pinhole	Indoor	8 subjects, 8 actions, 1.3M frames	Single
Occlusion-Person	2020	Pinhole	Indoor	73k frames	Single
Mo2Cap2 [205]	2018	Egocentric fisheye	Indoor & Outdoor	8 actions, 530k frames	Single
*x*R-EgoPose [180]	2019	Egocentric fisheye	Indoor	46 subjects, 9actions, 383K frames	Single
CMU Panoptic [71]	2016	Pinhole	Indoor	8 subjects, 1.5M frames	Multi
MuPoTS-3D [129]	2018	Pinhole	Indoor & Outdoor	8 subjects, 8k frames	Multi
F-M3DHPE (Our dataset)	2022	Fisheye	Indoor	11 subjects, 10 actions, 2.8k frames	Multi

Figure 29: Example images of our F-M3DHPE dataset for multi-person 3D pose estimation. The actions from left to right are Conversation, Taking photo, Phoning, Walking and Stretching. The first row and second row represent two viewpoints for the same pose.

MuCo-3DHP and MuPoTS-3D. MuCo-3DHP is a synthesised dataset build on the MPI-INF-3DHP dataset. The number of subjects in the image ranges from 1 to 4. Since 3D poses are known in the MPI-INF-3DHP dataset, MuCo-3DHP provides overlapping scenarios with correct depth ordering.

MuPoTS-3D is used for evaluation of multi-person 3D pose estimation captured by a multi-view marker-less motion capture system. This dataset consists of more than 8000 frames and 8 participants wearing different clothes performing daily activities. The dataset provides 3D pose annotations of up to 3 people collected for 20 scenes, where 5 indoor sequences are taken at 30fps and 15 outdoor sequences are captured at 60fps.

Mo2Cap2. Mo2Cap2 is an egocentric 3D human pose dataset with both indoor and outdoor environments. In this dataset, a fisheye camera is mounted on a baseball cap. The synthetic training set includes 530k images that are rendered from around 700 different characters performing 3K actions. Different from the synthetic training set, the test set is captured in both real indoor and outdoor environments. Participants dress general cloths and perform 8 common actions such as sitting, crawling and boxing. The number of images in the test set is 5591. 3D annotations are obtained by a multi-view marker-less mocap system. 2D/3D poses with 15 joints and camera information are included in this dataset.

***x*R-EgoPose.** *x*R-EgoPose is a synthetic dataset for egocentric 3D human pose estimation. It consists of 383K frames with a resolution of 1024×1024 pixels captured by a fisheye camera mounted on a human head. There are 46 characters (23 males and 23 female) performing 9 actions including gaming, lower/upper stretching and so on. In this dataset, characters dress different clothes with various textures and colors. The rendered images have superior quality. The character models are taken from real mocap data. 2D/3D poses, human body parts, normal maps and camera pose are provided.

F-M3DHPE. The proposed F-M3DHPE dataset, for multi-person 3D pose estimation, is captured in an indoor environment with two Insta360 cameras, and one LiDAR camera. 11 participants with different clothes perform 10 common activities such as posing, discussion, walking, etc. Our dataset includes more than 2000 frames with a resolution of 1920×1920 pixels. Example images are shown in Figure. 29. To obtain 3D ground-truth poses, we manually label 2D poses for the images captured by a LiDAR camera. Then, the 3D poses are obtained by labeling 2D poses and their corresponding depth values.

Table 18: The quantitative results on MuPoTS-3D dataset with the metric of 3DPCK for root-relative 3D poses.

Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg ↓
Bottom-up approaches																					
Mehta *et al.* [129]	81.0	59.9	64.4	62.8	68.0	30.3	65.0	59.2	64.1	83.9	67.2	68.3	60.6	56.5	69.9	79.4	79.6	66.1	66.3	63.5	65.0
Mehta *et al.* [128]	88.4	65.1	68.2	72.5	76.2	46.2	65.8	64.1	75.1	82.4	74.1	72.4	64.4	58.8	73.7	80.4	84.3	67.2	74.3	67.8	70.4
Zhen *et al.* [230]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	80.5
Top-down approaches																					
Rogez *et al.* [154]	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
Rogez *et al.* [155]	87.3	61.9	67.9	74.6	78.8	48.9	58.3	59.7	78.1	89.5	69.2	73.8	66.2	56.0	74.1	82.1	78.1	72.6	73.1	61.0	70.6
Dabral *et al.* [32]	85.1	67.9	73.5	76.2	74.9	52.5	65.7	63.6	56.3	77.8	76.4	70.1	65.3	51.7	69.5	87.0	82.1	80.3	78.5	70.7	71.3
Moon *et al.* [131]	94.4	77.5	79.0	81.9	85.3	72.8	81.9	75.7	90.2	90.4	79.2	79.9	75.1	72.7	81.1	89.9	89.6	81.8	81.7	76.2	81.8
Wang *et al.* [186]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.0
Chen *et al.* [27]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	87.5
Cheng *et al.* [28]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.6
Bottom-up approaches																					
Mehta *et al.* [129]	81.0	64.3	64.6	63.7	73.8	30.3	65.1	60.7	64.1	83.9	71.5	69.6	69.0	69.6	71.1	82.9	79.6	72.2	76.2	85.9	69.8
Mehta *et al.* [128]	88.4	70.4	68.3	73.6	82.4	46.4	66.1	83.4	75.1	82.4	76.5	73.0	72.4	73.8	74.0	83.6	84.3	73.9	85.7	90.6	75.8
Zhen *et al.* [230]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	73.5
Top-down approaches																					
Rogez *et al.* [154]	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
Rogez *et al.* [155]	88.0	73.3	67.9	74.6	81.8	50.1	60.6	60.8	78.2	89.5	70.8	74.4	72.8	64.5	74.2	84.9	85.2	78.4	75.8	74.4	74.0
Dabral *et al.* [32]	85.8	73.6	61.1	55.7	77.9	53.3	75.1	65.5	54.2	81.3	82.2	71.0	70.1	67.7	69.9	90.5	85.7	86.3	85.0	91.4	74.2
Moon *et al.* [131]	94.4	78.6	79.0	82.1	86.6	72.8	81.9	75.8	90.2	90.4	79.4	79.9	75.3	81.0	81.0	90.7	89.6	83.1	81.7	77.3	82.5
Lin and Lee *et al.* [107]	94.4	79.6	79.2	82.4	86.7	73.0	81.6	76.3	90.1	90.5	77.9	79.2	78.3	85.5	81.1	91.0	88.5	85.1	83.4	90.5	83.7

* The top row reports values for all annotated poses (All), while the values in bottom row are calculated by matched poses (Matched).

Table 19: Quantitative results of existing methods on the MuPoTS-3D dataset for absolute multi-person 3D poses.

Methods	Input	$3DPCK_{Matched}$ ↑	$3DPCK_{All}$ ↑
Optimization-based methods			
-	-	-	-
Learning-based methods			
Moon *et al.* [131]	A single image	31.8	31.5
Zhen *et al.* [230]	A single image	38.7	35.4
Wang *et al.* [186]	A single image	-	43.8
Lin and Lee [107]	A single image	35.2	-
Chen *et al.* [19]	Videos	45.7	-
Cheng *et al.* [28]	Videos	48.0	-

Finally, 3D poses for Insta360 cameras are acquired by the transformation matrix of Insta360 camera and LiDAR camera. We provide 3D pose annotations for up to 3 subjects in this dataset.

The advantage of our proposed F-M3DHPE dataset is that our source frames provide different views from one insta360 camera, for the participants, from the panorama cameras, *i.e.,* Insta360 cameras, during collection. Each Insta360 camera contains two fisheye lenses. We provide images taken from one of the fisheye cameras in our dataset. We can project the spherical panorama at any camera angle within a 360-degree field of view to obtain fisheye images that looks like it is taken by a single fisheye camera as shown in Figure. 31. In other words, our dataset can be extended to the multi-view set-ups. In this chapter, we only analyze images captured by one fishsye lens from the Insta360 camera.

2) Evaluation Metrics

Mean Per Joint Position Error (MPJPE). MPJPE is commonly used for the evaluation of 3D human pose estimation. MPJPE is calculated by the Euclidean distance between predicted 3D poses and the ground truth after aligning the root joint, as defined by:

$$\text{MPJPE} = \frac{1}{N}\sum_{i=1}^{\text{N}} \left\| \mathbf{J}_i^{rel} - \mathbf{J}_i^{rel*} \right\|_2, \tag{5.5}$$

where N denotes the number of test frames, i represents the index of the samples, $\mathbf{J}$ and $\mathbf{J}^*$ indicate the predicted and ground-truth joint locations, *rel* means the joint locations are root-relative. In general, MPJPE is measured in millimeters.

Procrustes Analysis Mean Per Joint Position Error (PA MPJPE). PA MPJPE is the MPJEP after further alignment, *i.e.,* the Procrustes transformation [49].

Mean of the Root Position Error (MRPE). MRPE is the metric proposed by Moon *et al.* [131] for absolute root joint estimation. Given predicted and ground-truth root joint locations in the camera coordinate $\mathbf{R}$ and $\mathbf{R}^*$, MRPE is as follows:

$$\text{MRPE} = \frac{1}{N}\sum_{i=1}^{N} \left\| \mathbf{R}_i - \mathbf{R}_i^* \right\|_2. \tag{5.6}$$

Table 20: The quantitative results on the proposed F-M3DHPE dataset.

Methods	Conversation	Direction	Eating	Greeting	Phoning	Photo	Posing	Purchasing	Stretching	Walking	MPJPE↓	MRPE↓
Zhen *et al.* [230]	574.85	554.99	548.39	657.46	536.91	606.70	690.47	603.03	782.33	594.05	616.23	5662.76
Moon *et al.* [131]	476.44	446.46	514.00	603.70	509.52	570.30	639.20	553.47	743.41	547.08	560.99	1655.50
Lin and Lee [107]	476.44	446.46	514.00	603.70	509.52	570.30	639.20	553.47	743.41	547.08	560.99	427.02
Zhang *et al.* [222]	300.39	247.72	307.26	411.14	324.31	393.01	464.37	397.05	573.02	380.21	381.19	384.90

3D Percentage of Correct Keypoints (3DPCK). 3DPCK is used in the MPI-INF-3DHP and MuPoTS-3D datasets as the evaluation metric. Specifically, the estimated 3D joint is considered as correct if the distance between predictions and ground truths is within a threshold. Following [127], the threshold is equal to 150mm.
Area Under the Curve (AUC). AUC is an another metric for MPI-INF-3DHP dataset. AUC is defined as the area under the 3DPCK curve with different thresholds.

3) Evaluation on Public dataset

We summarize the performance of existing methods for multi-person 3D human pose estimation on the MuPoTS-3D dataset. Particularly, the comparison for root-relative multi-person 3D pose estimation is shown in Table 18. Table 19 provides the comparison of absolute depth estimation for multi-person 3D poses.

In general, top-down approaches achieve better performance than bottom-up approaches. The reason is that top-down approaches divide the multi-person 3D pose estimation into single-person 3D person estimation for each detected person making the task simpler. However, absolute depth estimation is more challenging since the input image is cropped and hence ignores the relation between each person. Bottom-up approaches are faster than top-down approaches, because the former directly predicts 2D keypoints and absolute depths of all humans. However, the main weakness of bottom-up approaches is that the accuracy tends to be lower than top-down approaches.

Moon *et al.* [131] use MRPE on Human3.6M dataset to demonstrate that learning-based methods perform better to estimate the absolute depth than optimization-based methods.

4) Evaluation on our F-M3DHPE dataset

In contrast to existing datasets, our dataset is captured from Insta360 cameras consisting of two fisheye lenses to capture 3D human poses. Depending on the availability of code, we evaluate four methods for 3D multi-person pose estimation from perspective cameras ([131], [107] and [230]) and one method for egocentric 3D pose estimation ([222]) from a fisheye camera on our dataset. We choose images containing six specific subjects as the training set. The remaining images are used as the test set.

Table 20 lists the experimental results of the existing methods on the F-M3DHPE dataset. It is shown that all methods perform similarly for root-relative 3D human pose estimation. However, the method that considers distortion parameters outperforms other methods. Note that the distortion parameters are not provided in the evaluation. Therefore, taking image distortions into account contributes positively to the performance. On the other hand, bottom-up methods provide worse performance compared to top-down methods. The reason is that 2D-to-3D lifting module follows the same mapping relation to obtain all 3D poses from the detected 2D keypoints for bottom-up methods. However, 2D keypoints located in different regions may suffer from different levels of distortions.

5.5.2 *Dataset and Benchmarks for Action Recognition*

In this section, we describe public datasets for human action recognition and our new dataset F-HAR captured by a fisheye camera. Then, we analyze the performance of existing methods on these datasets.

1) Datasets for human action recognition

Table 21: Public dataets for human action recognition.

Dataset	Year	Camera Type	Modality	Size	Action
HMDB51 [85]	2011	Pinhole	Image	6,766	51
UCF101 [170]	2012	Pinhole	Image	13,320	101
Sports-1M [74]	2014	Pinhole	Image	1,133,158	487
ActivityNet [9]	2015	Pinhole	Image	28,000	203
YouTube-8M [74]	2016	Pinhole	Image	8,000,000	4716
Kinetics 400 [75]	2017	Pinhole	Image	306,245	400
Kinetics 600 [13]	2018	Pinhole	Image	495,547	600
Kinetics 700 [14]	2019	Pinhole	Image	650,317	700
MSR-Action3D [104]	2012	Pinhole	Skeleton	567	20
UTKinect-Action3D [201]	2012	Pinhole	Image + Skeleton	200	10
MSRDailyActivity3D [188]	2012	Pinhole	Image + Skeleton	320	16
Northwestern-UCLA [189]	2014	Pinhole	Image + Skeleton	1,475	10
SYSU-3D [58]	2015	Pinhole	Image + Skeleton	480	12
NTU RGB+D [157]	2016	Pinhole	Image + Skeleton	56,880	60
Kinetics-Skeleton [207]	2017	Pinhole	Image + Skeleton	266,440	400
NTU RGB+D 120 [110]	2019	Pinhole	Image + Skeleton	114,480	120
F-HAR(Our dataset)	2022	Fisheye	Image + Skeleton	1,000	16

Table 22: The quantitative results of existing methods for human action recognition on public datasets.

Methods	NTU CS↑	NTU CV↑	NTU120 CS↑	NTU120 CV↑	KS Top-1↑	KS Top-5↑
CNN-based methods						
SkeleMotion [10]	76.5	84.7	67.7	66.9	-	-
Chained Net [237]	80.8	-	-	-	-	-
Deep Bilinear [59]	85.4	90.7	-	-	-	-
2D/3D pose [123]	85.5	-	-	-	-	-
Posemaps [116]	91.7	95.3	64.6	66.9	-	-
Else-net [102]	91.6	96.4	-	-	-	-
LSTM-based methods						
GCA-LSTM [112]	76.1	84.0	61.2	63.3	-	-
Ind-RNN [100]	81.8	88.0	-	-	-	-
Geometric [43]	97.0	98.5	90.6	86.7	-	-
GCN-based methods						
ST-GCN [207]	81.5	88.3	-	-	30.7	52.8
SR-TSL [163]	84.8	92.4	-	-	-	-
AS-GCN [96]	86.8	94.2	-	-	34.8	56.5
2S-AGCN [160]	88.5	95.1	-	-	36.1	58.7
SGN [218]	89.0	94.5	79.2	81.5	-	-
AGC-LSTM [162]	89.2	95.0	-	-	-	-
MS-G3D [120]	91.5	96.2	86.9	88.4	38.0	60.9
Channel-wise [23]	92.4	96.8	88.9	90.6	-	-

There are two types of datasets for human action recognition, *i.e.,* with/without skeletons and depths data. Datasets without skeletons are used for video-based methods. Datasets with skeletons and or depth maps are used by skeleton-based methods. Existing datasets are captured by perspective cameras. In this chapter, we propose a newly dataset for human action recognition captured by a fisheye camera.

UCF101. UCF101 is a large-scale dataset for action recognition including 101 action classes. The dataset is divided into 5 types such as Human-Object Interaction and Sports. UCF101 consists of 13320 clips downloaded from YouTube at 25FPS with the resolution of 320×320 pixels. The dataset contains diverse viewpoints, subject appearances, dynamic backgrounds, etc.

Sports-1M. Sports-1M comprises of 1 million videos collected from YouTube. It includes 487 sport categories annotated by using the metadata of videos. For each class, this dataset includes 1000-3000 clips and there are 50K videos with more than one label including aquatic sports, ball sports, etc.

YouTube-8M. YouTube-8M is one of the largest datasets for action recognition. YouTube-8M consists of 8 million videos with the duration of 500K hours downloaded from YouTube. This dataset includes 4716 classes and each video contains 1.8 classes (on average).

Kinetics 400/600/700 and Kinetics-Skeleton. Kinetics dataset consists of 400/600/700 human action categories and more than 400 videos per class downloaded from YouTube. This dataset covers different human actions including human-object interactions and human-human interactions. Kinetics-Skeleton is an extension of Kinetics 400 dataset. Due to no skeleton information in Kinetics dataset, [207] adopt a 2D pose detector OpenPose [11] to obtain 2D skeletons for skeleton-based methods for human action recognition.

NTU RGB+D and NTU RGB+D 120. NTU RGB+D and NTU RGB+D 120 are large scale datasets with depth information for action recognition. They consists of 56,880 and 114,480 clips captured in an indoor environment, respectively. NTU RGB+D 120 includes 120 action classes, while NTU RGB+D comprises of 60 action classes. There are three types, *i.e.,* daily actions, health-related actions and mutual actions in the datasets. Both datasets include RGB videos with the resolution of 1920×1080 pixels, depth map with the resolution of 512×512, 3D skeletons with 25 body joints.

F-HAR. The proposed F-HAR dataset is captured by an Insta360 camera, consisting of two fisheye lenses, capturing 10 scenes with both indoor and outdoor environments as shown in Figure. 30. There are 13 participants (7 males and 6 females), and each participant performs actions in five scenes of an indoor or outdoor environment. Furthermore, 7 participants are present in all indoor scenes while 6 participants are in all outdoor scenes. There are 14 action categories consisting of taking off coat/backpack, wearing coat/backpack, walking, talking, moving, phoning, drinking water, using a computer, taking photos, waving hands, writing, reading, clapping, and eating. Our dataset consists of 1000 clips. Each clip includes one action class lasting between 3 to 10 seconds. The videos are taken at 30fps with the resolution of 1920×1920 pixels. Labels, camera information and skeletons are provided. Skeletons are generated by an off-the-shelf model [95].

Different from previous datasets, F-HAR dataset provides a fisheye view to perform human action recognition. Moreover, F-HAR provides different viewpoints as shown

Figure 30: Example images of the proposed F-HAR dataset for human action recognition. 13 participants perform 10 actions in 5 indoor scenes and 5 outdoor scenes.

Table 23: Quantitative results of existing methods on our F-HAR dataset for human action recognition.

Methods	F-HAR-CS↑	F-HAR-CP ↑
ST-GCN [207]	12.21%	23.31%
AS-GCN [96]	10.57%	6.42%
2s-AGCN [160]	19.14%	30.74%
MS-G3D [120]	53.16%	70.27%

in Figure. 31 because of the panorama cameras as described in Section 5.5. Therefore, F-HAR is beneficial for 1) human action recognition from a single image captured by a fisheye camera; 2) multi-view human action recognition; and 3) human action recognition from a single panorama image. In this survey chapter, we only focus on videos captured by one fisheye lens from the Insta360 camera.

2) Evaluation Metrics

The mean accuracy is used to evaluate the methods for human action recognition, *i.e.,* the proportion of the number of correct predictions to all samples.

3) Evaluation on public datasets

We summarize the performance of existing skeleton-based methods on the commonly used NTU, NTU120 and Kinetics-Skeleton datasets. For the NTU and NTU120 datasets, existing methods use 3D information (2D poses with depth maps or 3D joint locations). In terms of Kinetics-Skeleton dataset, the sequences of detected 2D skeletons are used for action classification. The experimental results are shown in Table 22. GCN-based methods tend to perform better than the other two kinds of methods. Instead of 3D skeletons, GCN-based methods can also use 2D skeletons for human action recognition. GCN-based methods have the advantage to exploit relationships between nodes presented by the 2D keypoints in the skeleton-based methods for human action recognition.

4) Evaluation on our F-HAR dataset

Depending on the availability of code, four methods are used to compare the performance on our dataset for human action recognition using a fisheye camera. Based on whether the training set includes both indoor and outdoor environments, there are two split protocols including F-HAR-CS and F-HAR-CP. Specifically, F-HAR-CS represents the training set that only includes videos captured in indoor or outdoor environments.

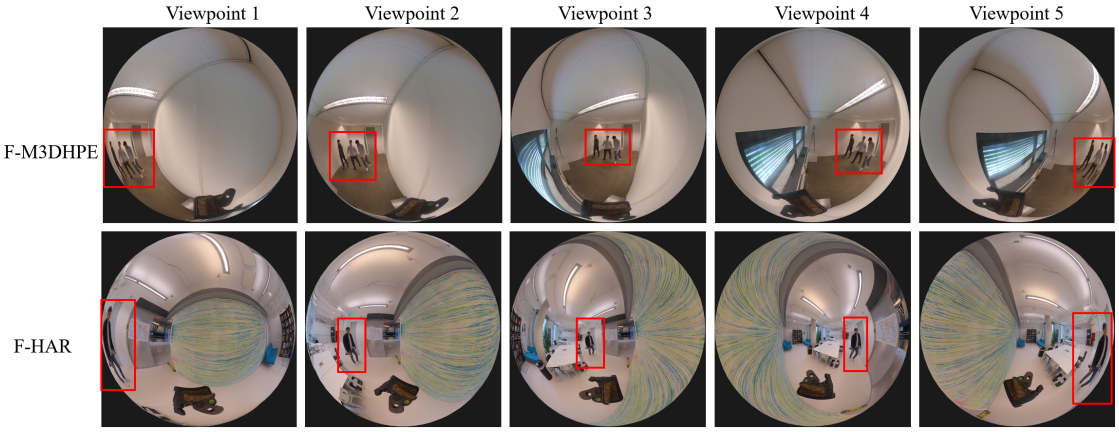


Figure 31: The frames in F-M3DHPE and F-HAR datasets can be viewed from different angels in one Insta360 camera. It can be seen that the subject at the edge of the image suffers from large distortions, causing 3D HPE and HAR more challenging.

F-HAR-CP indicates the training set consisting of both indoor and outdoor environments using videos performed by half the number of subjects as the training set. The experimental results are shown in Table 20. Liu *et al.* [120] outperforms the other three methods on both protocols. The reason is that this method alleviates the biased weighting problem in GCNs and that it combines spatial-temporal connections.

5.6 DISCUSSION

In this chapter, a survey is presented on 3D human pose estimation and human action recognition. A new dataset is collected using a fisheye camera for multi-person 3D pose estimation and human action recognition. Despite significant progress over the past few years, there are still a number of challenges. In this section, we provide possible future directions.

3D Human Pose Estimation. Existing methods for 3D pose estimation mainly focus on perspective cameras, and there are only a few methods for multi-person 3D pose estimation using fisheye cameras. Methods exploiting a fisheye camera model to alleviate the negative influence caused by the fisheye lens [222] outperforms other methods. From the experimental results of existing methods on our new dataset, the performance of multi-person 3D pose estimation is still far from perfect. The main challenge is that humans, located at different positions, suffer from different levels of distortions. Therefore, combining human positions and distortion parameters to regularize root-relative 3D poses and absolute depths is a promising direction for multi-person 3D pose estimation.

Human Action Recognition. Human poses are essential for human action recognition. However, it is difficult to obtain 2D or 3D ground-truth skeletons especially for in-the-wild images. A common strategy is to use off-the-shelf 2D/3D estimators to obtain 2D/3D skeletons. Although skeleton-based methods for human action recognition made good progress, they only use 2D or 3D skeletons. In terms of skeletons, estimated 2D poses are more robust but also more ambiguous. Estimated 3D poses provide depth information and reduce ambiguity but are also less stable. Therefore, the combination of 2D and 3D poses may be beneficial for human action recognition. Moreover, how to combine skeletons with input images is also an interesting future research topic.

6

HUMAN MOTION TRANSFER WITH POSE CONSISTENCY

6.1 INTRODUCTION

Animating a person, in an image or video, from its source pose to a novel (target) pose, referred to as human motion transfer, is a topic receiving more and more attention. It has many applications in computer vision including movie production, entertainment and education. With the introduction of Generative Adversarial Networks (GANs), GAN-based methods provide promising results for image synthesis in general. However, these methods usually focus on global or style transformations and ignore geometric relations [119, 191, 192]. Therefore, they may fall short to generate high quality synthesized images of humans with novel (target) poses.

Recently, a number of methods uses extracted 2D human poses as a condition to guide the animation process [15, 124]. A drawback is that this type of methods usually needs large-scale training data for a certain person and therefore limiting its applicability. To this end, researchers adopt off-the-shelf human pose and/or shape models [12, 72, 156] to relate humans in source and target images. Particularly, the warping process exploiting estimated 2D poses [4, 164–167], 3D poses [79], 3D human parametric models [117], 3D implicit volumetric representations [147], is an essential part for existing methods to transfer human motion.

Existing methods perform well in terms of reconstruction quality. However, the challenge, to generate high quality synthesized images with accurate animated poses, still remains. Specifically, *i)* existing methods, using a 3D human model as an intermediate presentation to build the relation between source and target images, are based on (estimated) 3D human parametric models. This type of methods tends to provide higher image quality than methods using 2D poses to compute the motion transfer. This is because (projected) 3D human models are able to compute more accurate correspondences than sparse 2D keypoints. However, the quality of the human pose transfer has received less attention although it is an important part of the motion transfer process. *ii)* articulated persons usually exhibit self-occluded poses in 2D images. Existing methods attempt to use either the warped 2D or 3D features extracted from the source images to perform this task. However, for 2D-based methods, it's hard to solve the self-occlusion problem. For 3D-based methods, the estimated 3D information is less robust and precise than 2D information.

To alleviate the problem caused by inaccurate 3D human models, we propose to use 2D information for robustness. Our method uses human masks and 2D human keypoints to regularize the animated persons in the synthesized images. Although the use of human masks is able to constrain the pose of the animated person in a specific region, it may

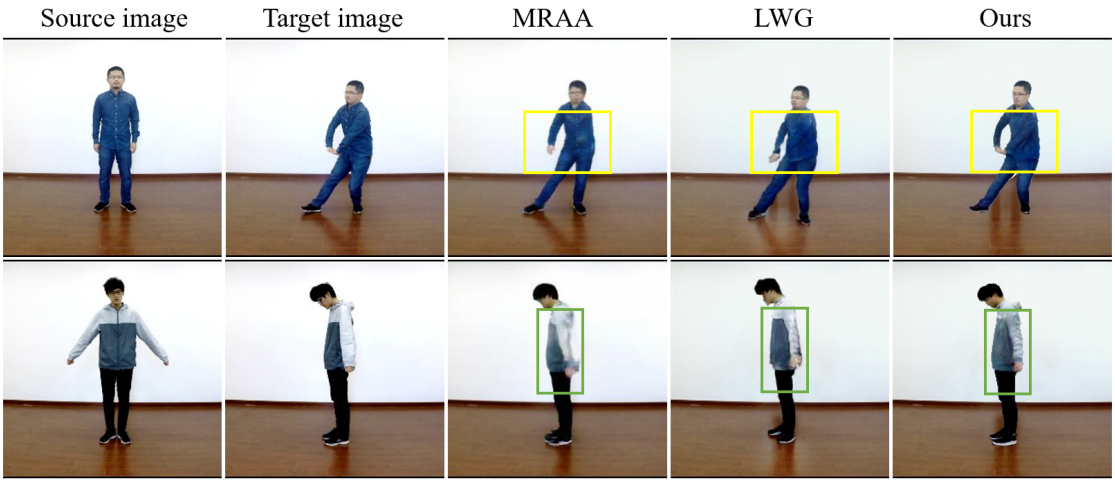


Figure 32: The proposed method generates realistic images of an animated person driven by the person in the target image. Comparing with recent state-of-the-art methods, our method shows improvements in both image quality and pose accuracy.

fail in case of self-occlusions exhibited by target poses. Therefore, in this chapter, 2D keypoints are used to retain semantic consistency between the target and reconstructed images.

To mitigate the self-occlusion problem, we introduce a novel strategy to exploit the relation between 2D and 3D representations during the animation process. In contrast to 2D information, 3D models contain more information such as depth and 3D structure. However, as opposed to 3D, 2D information is more robust. To this end, we propose to use the best of the two worlds by combining 2D and 3D information to steer the generation procedure. Specifically, we first calculate the 2D optical flow based on the projected 3D human models and 3D flows based on estimated 3D poses. Then, warped 2D and 3D features are used by our generator module to synthesize images of animated persons. As the warped 2D and 3D features are complementary, we design an attention module to estimate an occlusion map to reduce the redundancy of the combination of the warped 2D and 3D features. To the best of our knowledge, this is the first method to exploit the relation of 2D and 3D information to perform human motion transfer.

Our contributions are summarized as follows:

- We propose a novel end-to-end method by *i)* regularizing generated animations of humans based on 2D information, and *ii)* exploiting the relation between 2D and 3D information for human motion transfer.
- To reduce the dependency on the accuracy of the estimated 3D human models, we propose to use 2D information to regularize synthesized humans by constraining the generated regions and ensuring semantic consistency at arm/leg local regions between synthesized and target images.
- A novel strategy is proposed to combine 2D (robust) and 3D (depth/structure) features to alleviate the self-occlusion problem.
- Experiments demonstrate that the proposed method outperforms, both quantitatively and qualitatively, existing methods.

6.2 RELATED WORK

Methods for human motion transfer can be categorised into 2D- and 3D-based methods. Human motion transfer aims to animate a person in the source image with the pose of the person detected in the target image. Both image quality and animated pose accuracy are important factors to evaluate human motion transfer methods.

2D-based Methods. The three types of information, used to perform human motion transfer for 2D space based methods, are: conditional animation, affine transformations, and optical flow.

Conditional animation: the combination of a source image, a target image and the target pose is taken as an input [15, 124]. The target pose is regarded as a condition to guide the generation process followed by a discriminator to improve the image quality.

Affine transformation: a few methods [4, 166] compute the relation of source and target images based on the extracted 2D poses by existing human pose estimators [12, 95]. Then, the spatial transformation is computed to deform the source image and steer the generation process.

Optical flow: Siarohin *et al.* [164, 165, 167] propose a series of work to perform human motion transfer based on optical flow between source and target images. Keypoints are detected in an unsupervised manner to estimate the optical flow by modeling the relation of human co-parts. Other works [105, 117] adopt the 3D human model as a representation to estimate the optical flow. Specially, an off-the-shelf approach [72] is usually taken to obtain the 3D human mesh followed by a neural render to model the relation between source and target images.

3D-based Methods. Knoche *et al.* [79] propose to extend the affine transformation approach from 2D to 3D spaces. The method first uses a 3D estimator to obtain 3D poses followed by calculating the 3D affine transformation parameters and 3D masks. Instead of using 3D poses, Ren *et al.* [147] adopt an implicit volumetric representation to perform human motion transfer. The extracted representation is warped in 3D space. A decoder is designed to generate the synthesized image.

Pose Consistency. Pose consistency assumes that the animated person and the person in the target image have the same appearance including ratios of limbs and body sizes. The approach is an un-/semi-supervised dense correspondence estimation. Keypoints detection is used to improve robustness. Other methods [66, 125, 151, 178] apply a thin-plate spline (TPS) transformation to force the detector to be equivalent on images with different transformations. A few methods [86, 99, 134] retains semantic consistency at keypoint regions from different frames, e.g., eyes, necks and hips, to ensure the detectors to be invariant.

In general, there are two categories for human motion transfer based on *Pose Consistency*. *i)* [4, 15, 124, 147, 164–167] perform human motion transfer by absolute poses in target images. The scale of the animated human is the same as the human in the target image. *ii)* [79, 105, 117] animate a human by the relative motion of the human in the target image. The scale of the animated human is retained. The animated human generated by the latter is more beneficial since the appearance information, including body sizes and ratios of limbs, is preserved.

6.3 POSE GUIDED GENERATION

The goal of this work is to animate a human in the source image based on the pose exhibited in the target image with high reconstruction quality as well as pose accuracy. The main challenge for methods using 3D information is to maintain pose consistency between synthesized images and target images, since estimated 3D information (3D human poses or models) may not be accurate. For 2D-based methods, the ambiguity in source poses makes it difficult to obtain superior quality of synthesised images due to ambiguous human co-parts as shown in Figures 32 and 37. In this section, we present details of our solution to solve the above challenges.

6.3.1 *3D Human Model*

Skinned Multi-Person Linear (SMPL) model is commonly used as the human prior model. SMPL model decomposes the 3D human model into pose parameters $\theta \in \mathbb{R}^{24\times3}$ and shape parameters $\beta \in \mathbb{R}^{10}$. Pose parameters indicate 3D relative rotations of defined 23 joints and the orientation of the root joint. Shape parameters are PCA coefficients to represent the human shape. The SMPL model can be defined as follows: $M(\theta,\beta) \in \mathbb{R}^{N\times3}$, where N denotes vertices at the triangulated mesh and $N = 6890$.

Based on the above definitions, 3D human models of source and target images are described as $M_s(\theta_s,\beta_s)$ and $M_t(\theta_t,\beta_t)$. Ideally, the 3D human model of an animated person is $\hat{M}_t(\theta_t,\beta_s)$. One of the limitations of SMPL models is that it is not able to model hair and clothes. This may have a negative influence on the alignment of the rendered 3D models, especially near the body contours. Therefore, $\hat{M}_t(\theta_t,\beta_s)$ needs to be refined in the generation procedure.

6.3.2 *Pose Consistency*

As 2D poses lack depth information, for methods for human motion transfer with 2D poses, it is hard to preserve the scale of the person in the source image. Instead, 3D human parametric models or 3D poses contain the absolute scale information of humans. The scale of an animated person can be retained by transferring the pose parameters θ in SMPL models (i.e., $\hat{M}_t(\theta_t,\beta_s)$) or scaling skeletons based on ratios of 3D human poses from source and target images. In this chapter, we utilize the 3D human parametric model to guide the process of human motion transfer. However, there is no ground truth available in public datasets. Therefore, following previous work [117], we use an off-the-self model to obtain 3D human models as pseudo ground truth. One of the limitations is that the accuracy of the 3D models can not be guaranteed. This may cause existing methods to fail to synthesize the animated person with high pose accuracy. To alleviate this problem, we propose to use 2D information including 2D human poses and body masks to regularize the animation of a person. The aim of using body masks is to constrain the position of these masks after the person has been animated when there are no self-occlusions of target poses. Furthermore, inspired by methods for unsupervised keypoint detection, we propose a strategy to maintain the semantic consistency at regions (e.g., arms and legs) to deal with self-occlusions of target poses.

6.3.3 *Ambiguity*

Existing methods either use warped 2D or 3D features extracted from source images to perform human motion transfer. However, *i)* the former methods perform the warping process in 2D space and therefore ignoring 3D structural information. Hence, ambiguous human co-parts and the invisible parts of humans in images make it difficult to generate plausible textures of synthesized humans; *ii)* the latter methods are not stable since the estimated 3D information (3D poses, 3D human models, etc.) is less robust and accurate than 2D information (human masks, 2D human poses, etc.). Besides, 3D human parametric models do not include hair and clothes information, so it is challenging to animate a human with high quality.

Therefore, we propose to combine the warped 2D and 3D features to steer the generation procedure. For this, there are two issues to be addressed: *i)* warped 2D features may contain inaccurate pose information because the estimated 3D models may contain imperfections to compute 2D warping flows; *ii)* warped 2D and 3D features share the same parts of representation. Hence, naively concatenating warped 2D and 3D features may increase the influence of the shared parts but ignoring other useful information. To this end, we propose a feature fusion map to combine them. The aim is to reduce the redundancy of concatenated features and enhance the impact of useful representation. Finally, the weighed combination of 2D and 3D features is used by the decoder to synthesize the animated person. In this way, the combination exploits the robustness of 2D features together with enriched (such as 3D depth, pose and structure) 3D features to guide the image synthesis.

6.4 NETWORK AND TRAINING DETAILS

Our human motion transfer framework consists of four steps as shown in Figures 33 and 34: optical flow calculation, feature fusion, generation and discrimination. In this section, we present the details of our framework and training strategy.

6.4.1 *Network Design*

Optical Flow Calculation. Two branches are used to compute the 2D and 3D warping flows, respectively. Specifically, an off-the-shelf model HMR [72] is used to estimate the 3D parametric model in the first branch. Following [117], the 3D models (M_s and M_t) are then projected onto the 2D images to compute the 2D optical flow $T_{2d} \in \mathbb{R}^{H\times W\times 2}$ based on projected vertex indexing. In the second branch, we follow [79] to obtain the 3D flows $T_{3d} \in \mathbb{R}^{H\times W\times D\times 3}$ by extending affine transformations from 2D to 3D spaces. 3D human poses ($\mathbf{P}_s$ and $\mathbf{P}_t$) are estimated by the off-the-self 3D estimator [156]. Finally, we warp the extracted 2D and 3D features based on 2D and 3D warping flows. In this work, $H = W = D = 64$.

Feature Fusion. To benefit from both robust 2D and structural 3D information, we propose to combine them to perform the generation procedure. Particularly, both 2D and 3D features are extracted from the warped images and projected 3D human models. Then, a fusion map $m_F \in \mathbb{R}^{H\times W}$ is estimated by the fusion module, similar to the hourglass

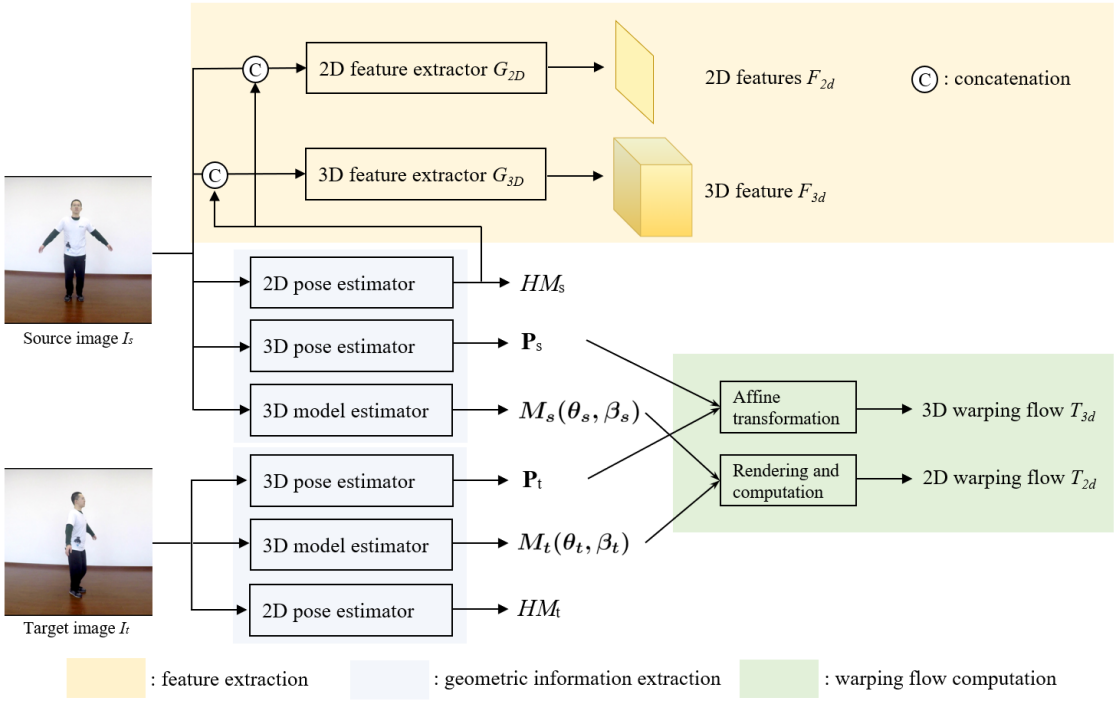


Figure 33: Geometric information and 2D/3D feature extraction. Given a pair of source and target images, 2D and 3D feature extraction is used to obtain their representation in 2D- and 3D-space. 2D/3D pose and human model estimation is then adopted to compute geometric details with 2D heatmaps, 3D poses, and 3D human mesh representation. With pairs of 3D human models and 3D poses, 2D and 3D warping flows are computed. Note that the source mask is used during the feature extraction procedure.

network, to concatenate the warped 2D and 3D features to reduce the redundancy between them. The fusion module takes the warped source image I_{syn} by 2D optical flow and the difference between heatmaps ΔHM of source poses $\mathbf{p}_s$ and target poses $\mathbf{p}_t$.

$$\begin{aligned} m_F &= G_{map}\left(I_{syn}, \Delta HM\right), \\ I_{syn} &= w_{2D}(I_t, T'_{2d}), \\ \Delta HM(\mathbf{p}) &= \exp\left(-\frac{(\mathbf{p}-\mathbf{p}_t)^2}{2\sigma^2}\right) - \exp\left(-\frac{(\mathbf{p}-\mathbf{p}_s)^2}{2\sigma^2}\right), \end{aligned} \tag{6.1}$$

where G_{map} denotes the fusion module, w_{2D} is the 2D warping procedure, and T'_{2d} represents the resized 2D optical flow and $T'_{2d} \in \mathbb{R}^{256\times256\times2}$. σ is set to 2 pixels.

Generation. Based on the 2D and 3D flows and feature fusion map, we adopt the hourglass network to generated synthesized images. Specifically, we first extract 2D and 3D appearance features by networks G_{2D} and G_{3D} followed by being warped in 2D and 3D spaces. Then, we concatenate the warped 2D and 3D features based on the fusion

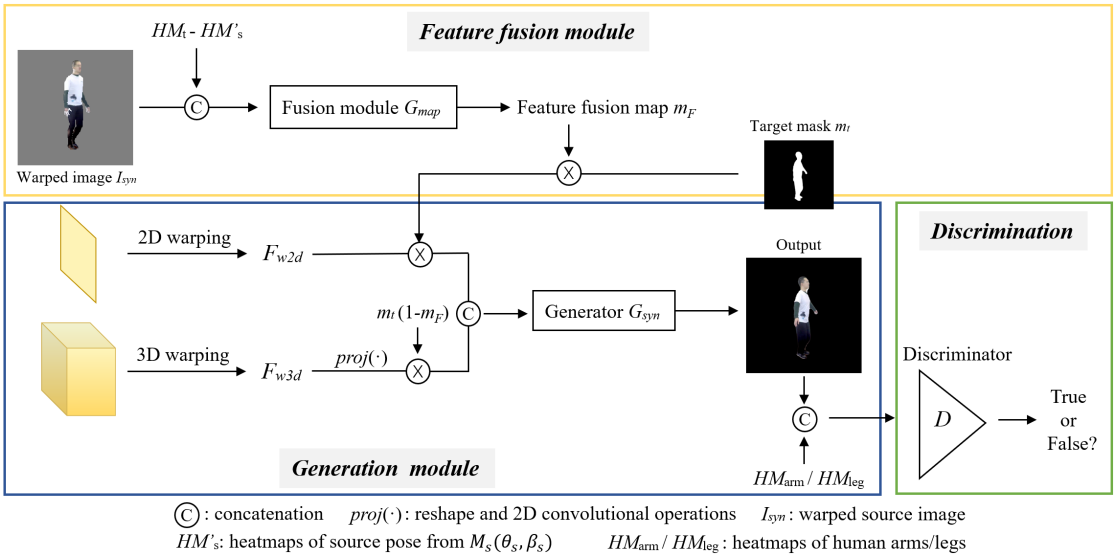


Figure 34: Overview of our generation module and discriminator. The extracted 2D and 3D features are warped by computing 2D and 3D warping flows. Then, a feature fusion map is computed to concatenate warped 2D and 3D features. The aim is to reduce (1) the redundancy between the 2D and 3D features and (2) inaccurate pose information. The input is the warped source image and the difference between the source and target poses. Finally, a generator converts the combination of weighted warped 2D and 3D features to an animated image. To maintain semantic consistency between outputs and target images, discriminators are used with arm/leg regions to ensure the animated person is located at the same positions as the target pose.

map m_F. Finally, the concatenated features are used by the decoder to generate images of animated humans $\hat{I}_t$.

$$\begin{aligned} \hat{I}_t &= G_{syn}\left([F_{w2d} \odot m_F \odot m_t, F'_{w3d} \odot (1 - m_F) \odot m_t]\right), \\ F_{2D} &= G_{2D}\left(I_s, HM_s\right), F_{3D} = G_{3D}\left(I_s, HM_s\right), \\ F_{w2d} &= w_{2D}(F_{2D}, T_{2D}), F_{w3d} = w_{3D}(F_{3D}, T_{3D}), \\ F'_{w3d} &= proj(F_{w3d}), \end{aligned} \tag{6.2}$$

where HM_s represents the source pose with heatmap representation, $w(\cdot)$ denotes warping operation, $proj(\cdot)$ indicates warped 3D features are reshaped then followed by 2D convolutional operations, and m_t is the mask of the target image.

Discrimination. We follow [117] to design our discriminators. To regularize the animated pose, we propose to enforce the semantic consistency between $\hat{I}_t$ and I_t at arms/legs. To this end, we combine the generated image $\hat{I}_t$ with regions into the specified discriminator determining whether it is an arm/leg region in $\hat{I}_t$. Specifically, consecutive joints are used to represent the regions, i.e., shoulders, elbows, and wrists for arm regions; hips, knees, and ankles for leg regions which are denoted by HM_{Arm} and HM_{Leg} respectively. $\hat{I}_t$ is separately taken as input to the discriminator. Therefore, our discriminator consists of three components: D_{Arm} for arm regions, D_{Leg} for leg regions, and D_{Human} for human body.

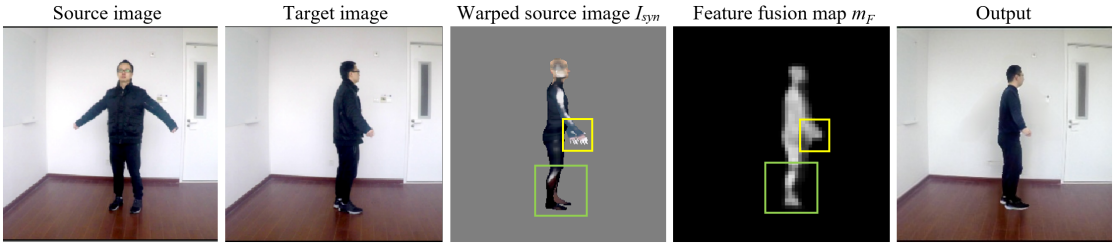


Figure 35: Visualization of the feature fusion map m_F. m_F is used to balance the warped 2D and 3D features to reduce (1) feature redundancy and (2) inaccurate pose information. It is shown that the inaccurate regions and edges in I_{syn} have smaller weights.

6.4.2 *Training*

Image Loss. A perceptual loss [69] is adopted to enforce the reconstructed $\hat{I}_t$ and target I_t images to be similar for both image and feature levels extracted by VGG19.

$$L_{image} = \left\| VGG19\left(\hat{I}_t\right) - VGG19\left(I_t\right) \right\|_1 . \tag{6.3}$$

Face Identity Loss. Following previous works [15,117], we apply a face loss to preserve the face identity information during the generation process. SphereFaceNet (SFN) [118] is used to extract face identity information between reconstructed and target images:

$$L_{face} = \left\| SFN\left(\hat{I}_t\right) - SFN\left(I_t\right) \right\|_1 . \tag{6.4}$$

Semantic Loss. A discriminator is used to retain semantic consistency at arm or leg locations between reconstructed and target images. Specifically, the reconstructed image is taken $\hat{I}_t$ with arm and leg regions with heatmap representations (HM_{arm} and HM_{leg}) as inputs to the discriminator to enhance pose consistency. To preserve image quality, $\hat{I}_t$ is separately fed into the discriminator.

For the generator, the loss is:

$$\begin{aligned} L_{adv}^{G} = & \sum D_{Human}\left(\hat{I}_t\right)^2 + \sum D_{Arm}\left(\hat{I}_t, HM_{arm}\right)^2 \\ & + \sum D_{Leg}\left(\hat{I}_t, HM_{leg}\right)^2 . \end{aligned} \tag{6.5}$$

For the discriminator, the loss is:

$$\begin{aligned} L_{adv}^{D} = & \sum \left[D_{\text{Human}}\left(\hat{I}_t\right) + 1\right]^2 + \sum \left[D_{\text{Human}}\left(I_t\right) - 1\right]^2 \\ & + \sum \left[D_{\text{Arm}}\left(\hat{I}_t, HM_{arm}\right) + 1\right]^2 + \sum \left[D_{\text{Arm}}\left(I_t, HM_{arm}\right) - 1\right]^2 \\ & + \sum \left[D_{\text{Leg}}\left(\hat{I}_t, HM_{leg}\right) + 1\right]^2 + \sum \left[D_{\text{Leg}}\left(I_t, HM_{leg}\right) - 1\right]^2 . \end{aligned} \tag{6.6}$$

Full Objectives. From Eqs. (6.3) - (6.6), the overall loss function of our framework is given by:

$$L = \lambda_{image} L_{image} + \lambda_{face} L_{face} + L_{adv}^{G} + L_{adv}^{D}, \tag{6.7}$$

where λ aims to balance the weights of each loss term.

Table 24: Comparison with existing methods on iPER dataset.

Approach	SSIM ↑	LPIPS↑	Pose Error↓	FID↓
PG2* [124]	0.854	0.865	-	-
SHUP* [4]	0.832	0.901	-	-
DSC* [166]	0.829	0.871	-	-
LWG** [117]	0.840	0.907	10.912	54.291
MRAA** [167]	**0.862**	0.911	**5.049**	107.443
Ours	0.861	**0.933**	6.010	**52.615**

[1] $*$ indicates the values are from [117].
[2] $**$ denotes the compared models are re-trained from scratch for comparison.

Table 25: Comparison with existing methods on the FashionVideo dataset.

Approach	SSIM ↑	LPIPS↑	Pose Error↓	FID↓
LWG [117]	0.882	0.930	3.310	18.164
MRAA [167]	**0.932**	0.931	2.386	29.971
Ours	0.916	**0.949**	**2.298**	**17.735**

[1] LWG and MRAA are re-trained from scratch for comparison.

6.5 EXPERIMENTS

Datasets. The iPER [117] and FashionVideo [211] datasets are adopted for evaluation. The iPER dataset contains 30 subjects with different types of cloth and background. There are 164 training video sequences and 42 test video sequences with a resolution of 512×512. The FashionVideo dataset consists of 600 video sequences, where 500 videos are defined as training samples and the remaining videos serve as the test set. Each sequence contains around 300∼350 frames with different resolutions. People wear different clothes against a white background.

Implementation Details. We normalize the images between the range of [-1, 1] with a resolution of 256×256. We adopt HMR [72] and Metrabs [156] to compute the 3D human parametric models and human poses, respectively. Adam optimizer is adopted with a batch size of 12. The weights of the loss term is set to: $\lambda_{image} = 10$ and $\lambda_{face} = 5$. The model is trained with 30 epochs, where L_{adv} is only applied in the last 25 epochs to stabilize the training.

Metrics. Commonly used metrics are: Structural Similarity (SSIM) [194] and Learned Perceptual Similarity (LPIPS) [220] for cases of self-imitation, where the source and target images are from the same video; Fréchet Inception Distance (FID) [54] is adopted for cases of cross-imitation, where source and target images are from different videos. As the focus is on both the reconstruction quality as well as pose accuracy, an additional metric is used — Pose Error to evaluate the accuracy of the human pose transfer, i.e., the pose accuracy between animated humans and humans in the target images. Specifically, the Pose Error is calculated by the Euclidean distance of the estimated 2D poses from AlphaPose [95] in image space. Appearance information (e.g., limbs ratios and body

Table 26: Ablation study on iPER dataset.

Type	Approach	SSIM ↑	LPIPS↑	Pose Error↓	FID↓
Feature Fusion	Ours w/o 2D	0.850	0.920	7.086	59.654
	Ours w/o 3D	0.855	0.928	6.256	53.120
	Ours w/o fusion map	0.849	0.921	7.425	52.629
Pose Constraint	Ours w/o mask+sem	0.825	0.898	9.368	55.796
	Ours w/o mask	0.826	0.899	9.016	55.818
	Ours w/o sem	0.859	0.931	6.323	53.171
	Ours (Full)	**0.861**	**0.933**	**6.010**	**52.615**

sizes) needs to be preserved during the animation process. Therefore, the Pose Error is used for cases of self-imitation instead of cross-imitation [61].

Comparative Methods. For comparison, based on the availability of the code, five state-of-the-art methods are selected including methods with 2D human poses [4,124,166,167] as well as 3D human parametric models [117]. For a fair comparison, we re-train [117] and [167] from scratch following the same train-test split. Because the task here is human motion transfer instead of in-painting, we adopt the same model as LWG [117] to perform the in-painting task for the background during the evaluation to eliminate the possible negative impact caused by the in-painting process.

6.5.1 *Comparative Study*

iPER Dataset. We quantitatively compare our method with five state-of-the-art methods on iPER dataset as shown in Table 24. Our method outperforms all other methods for both self- and cross-imitation except for 2D-based method MRAA. Although MRAA shows good results for self-imitation, it falls short in the case of cross-imitation. The reason can be that MRAA is over-fitting on this dataset. Compared with LWG, a method which is similar to our method as it also uses 3D human models as an intermediate representation, our method clearly outperforms LWG in terms of Pose Error metric. Qualitative comparison is shown in Figures 37 and 38 for self- and cross-imitations respectively. It can be seen in Figure 37 that our method can alleviate the self-occlusion problem, while MRAA suffers from pose ambiguities and self-occlusion problems. On the other hand, our method preserves scale information of the source person to synthesize images, which is ignored by MRAA as shown in Figure 38.

FashionVideo Dataset. Another comparison is conducted on the FashionVideo dataset. We use the available code from LWG and MRAA to re-train from scratch and compute the metrics. Compared with LWG and MRAA, our method provides competitive results. Since the type of gestures in FashionVideo dataset is limited, there is no significant discrepancy between the performance of the three methods regarding the Pose Error metric.

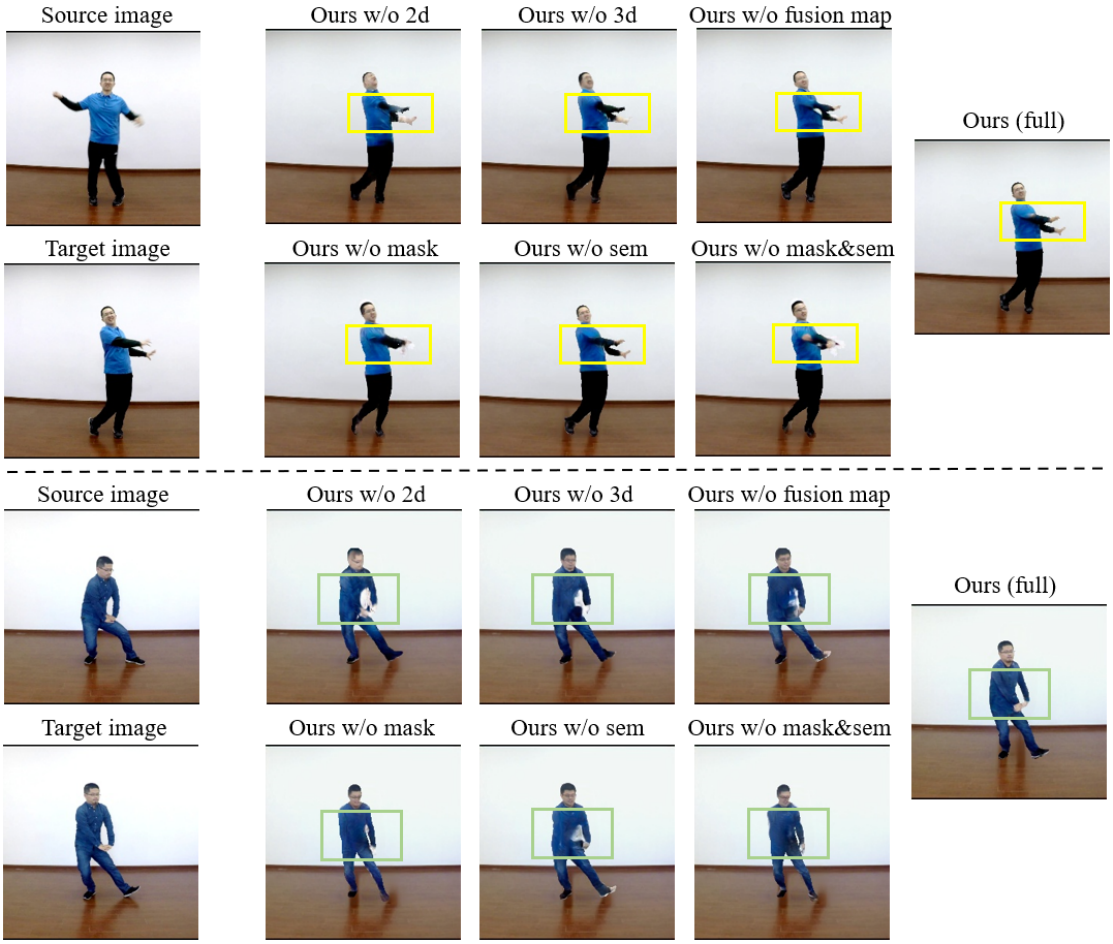


Figure 36: Qualitative comparison of ablation study. Our method obtains superior performance on pose accuracy of animated person and alleviates the self-occlusion problem when source and target poses are self-occluded.

6.5.2 *Ablation Study*

The ablation study is conducted on the iPER dataset because it contains more subjects, body movements and backgrounds than the FashionVideo dataset. The results are shown in Table 26.

Feature Fusion. As discussed in Section 6.3.3, a weighted combination of 2D and 3D features may benefit from both 2D (robustness) and 3D (structure/pose) information. To verify this, a comparative study is done with our method considering (1) without using 2D features (Ours w/o 2D), (2) without using 3D features (Ours w/o 3D), and (3) without using feature fusion map (Ours w/o fusion map). In Table 26, it is shown that: *i)* our method with a weighted combination of 2D and 3D features outperforms other model versions for both self- and cross-imitation; *ii)* our method without using a feature fusion map drops in performance in the case of self-imitation especially for the Pose Error metric. The reason is that the warped 2D features include inaccurate pose information due to the imperfections induced by estimated 3D human models; *iii)* our method without 2D features falls short in the case of cross-imitation. Hence, 2D features are more robust to synthesize the images.

Figure 37: Qualitative comparison of self-imitation with existing methods on iPER dataset.

Pose Consistency. The aim is to use 2D information including human masks and 2D keypoints to constrain the synthesised poses. Specifically, 2D keypoints are used to maintain pose consistency between the animation and the person in the target image. To validate pose consistency, experiments are conducted with our method focusing on (1) without using human masks (Ours w/o mask) or/and (2) semantic loss (Ours w/o sem). In Table 26, it is shown that *i)* human masks considerably improve the accuracy of the animated poses as they regularize the animated poses in non-overlapping regions of the target poses; *ii)* our method with semantic loss, aiming to deal better with self-occluded target poses, further increases the pose accuracy. On the other hand, synthesized images with inaccurate animated humans tend to exhibit inferior performance in the metrics of SSIM, LPIPS and FID.

A qualitative comparison of ablation study is shown in Figure 36, where the persons in source images exhibit self-occluded poses. Our method *i)* performs better on pose

Figure 38: Qualitative comparison of cross-imitation with existing methods on iPER dataset.

accuracy by our pose constraints; *ii)* alleviates the self-occlusion problems compared with our method without using informative 3D information (Ours w/o 3D).

In conclusion, the proposed feature fusion strategy and pose constraints positively contribute to the overall performance for both image quality and pose accuracy.

6.5.3 *Discussion*

The main limitation of our method is that it assumes that images contain a full human body. In fact, this is a limitation for existing methods for 3D human body reconstruction. However, our approach uses 3D human poses in the second branch to enable 3D warping and to steer the animation process. Consequently, we can only apply the second branch to deal with images containing parts of human bodies to perform human motion transfer. On the other hand, another reason to use 3D human poses from 3D pose estimators instead of from 3D human models is that the former tends to be more accurate.

6.6 CONCLUSIONS

A novel method is presented for human motion transfer focusing on both the reconstruction quality as well as pose accuracy. Our method combines the warped features in both 2D- and 3D-space to guide the generation process. To maintain pose consistency, a strategy is proposed to retain semantic consistency at arm/leg local regions between synthesized and target images. Experiments and an ablation study demonstrated the effectiveness of our method in terms of the quality of synthesized images and pose accuracy.

6.7 APPENDIX

In this section, we provide additional details on our framework in Section 6.7.1 and qualitative comparison on FashionVideo [211] and iPER [117] datasets in Section 6.7.2.

6.7.1 *Additional Framework Details*

Our framework consists of four components: optical flow calculation, feature fusion, generation and discrimination. As described in Section 6.4.1 of the main manuscript, the fusion module uses the warped source image I_{syn}, and the difference between heatmaps ΔHM of source poses $\mathbf{p}_s$ and target poses $\mathbf{p}_t$, as inputs to estimate feature fusion map m_F :

$$\begin{aligned} m_F &= G_{map}\left(I_{syn}, \Delta HM\right), \\ I_{syn} &= w_{2D}(I_t, T'_{2d}), \\ \Delta HM(\mathbf{p}) &= \exp\left(-\frac{(\mathbf{p}-\mathbf{p}_t)^2}{2\sigma^2}\right) - \exp\left(-\frac{(\mathbf{p}-\mathbf{p}_s)^2}{2\sigma^2}\right). \end{aligned} \tag{6.8}$$

Note that $\mathbf{p}_t$ is detected from the target images when source and target images are from the same video, while $\mathbf{p}_t$ needs to be scaled and translated when humans in the source and target images are different. Specifically, the scale and translation parameters are calculated by the ratios of body sizes between projected $\hat{M}_t(\theta_t, \beta_s)$ and $M_t(\theta_t, \beta_t)$ and the difference of the root (hip) joints $\mathbf{p}^{root}$ in the source and target poses:

$$\mathbf{p'}_t = \frac{\hat{s}_t^{body}}{s_t^{body}} \times (\mathbf{p}_t - \mathbf{p}_t^{root}) + \hat{\mathbf{p}}_t^{root}, \tag{6.9}$$

where human body sizes s^{body} are represented by the distances between neck and hip joints, and $\hat{}$ indicates the information extracted from $\hat{M}_t(\theta_t, \beta_s)$. Therefore, ΔHM is defined by:

$$\Delta HM(\mathbf{p}) = \exp\left(-\frac{(\mathbf{p}-\mathbf{p'}_t)^2}{2\sigma^2}\right) - \exp\left(-\frac{(\mathbf{p}-\mathbf{p}_s)^2}{2\sigma^2}\right). \tag{6.10}$$

The reason why we use the modified 2D target poses $\mathbf{p'}_t$, instead of the 2D poses extracted from $\hat{M}_t(\theta_t, \beta_s)$, is that the former tends to be more precise. In this way, the dependency on the errors is reduced which may be caused by erroneous 3D estimations.

6.7.2 *Additional Qualitative Comparison*

Besides the iPER dataset [117], we also evaluate our methods on the FashionVideo dataset [211]. Figure 39 shows a qualitative comparison of existing methods on the FashionVideo dataset; Figures 40 - 41 illustrate additional qualitative results on the iPER dataset.

From Figures 39 - 41, it is shown that: *i)* compared with the 2D-based method MRAA [167], our method generates more realistic animated humans in terms of image quality and the preservation of facial information; *ii)* the animated poses generated by our method are more accurate than MRAA and LWG [117] (e.g., animated arm positions), a method which is similar to our method as it also uses 3D human models as an intermediate representation.

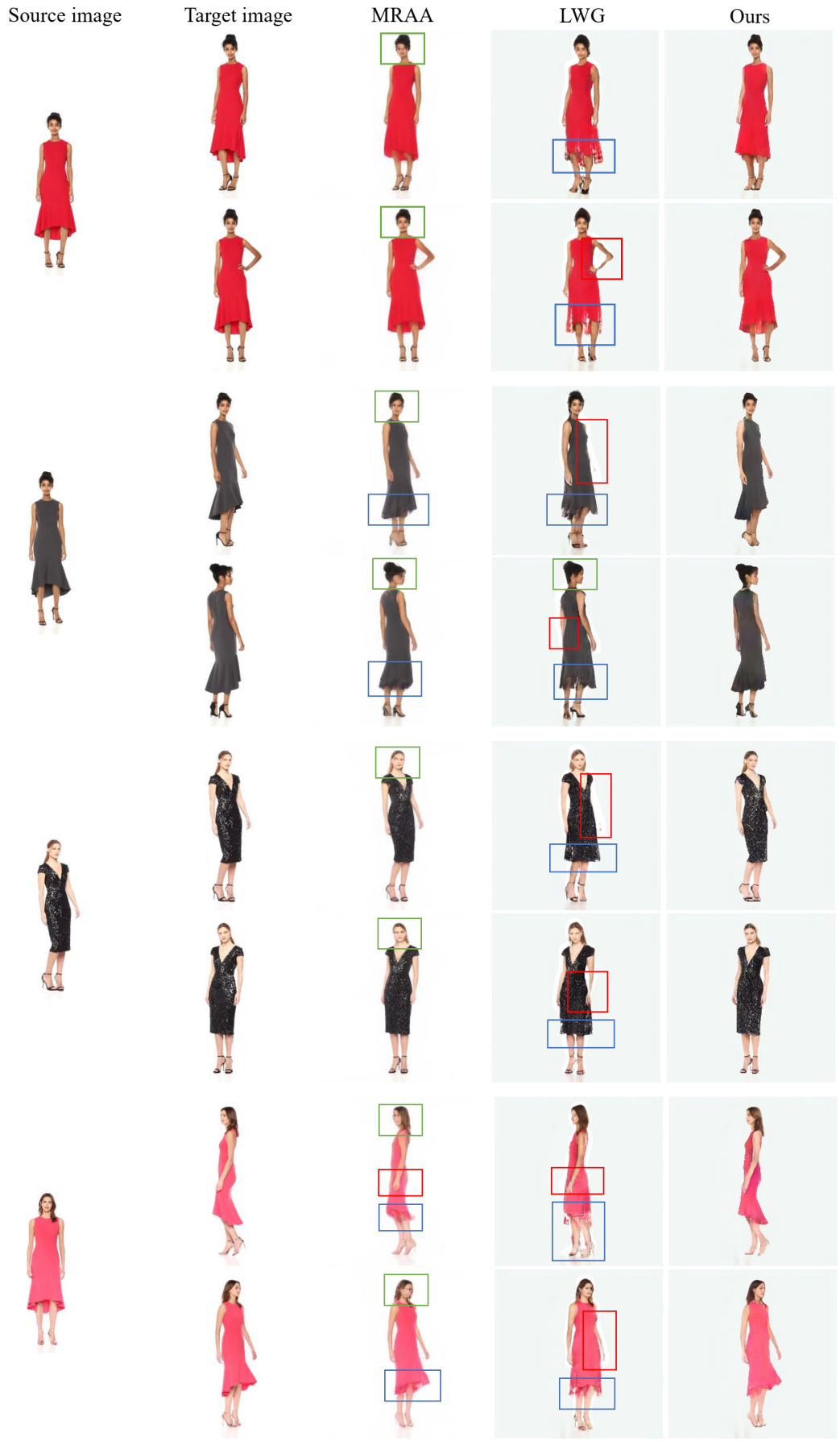


Figure 39: Qualitative comparison of existing methods on FashionVideo dataset. Red/blue/green rectangles focus on face, clothing and pose synthesis respectively.

Figure 40: Additional qualitative comparison of existing methods on iPER dataset. Rectangles focus on pose accuracy.

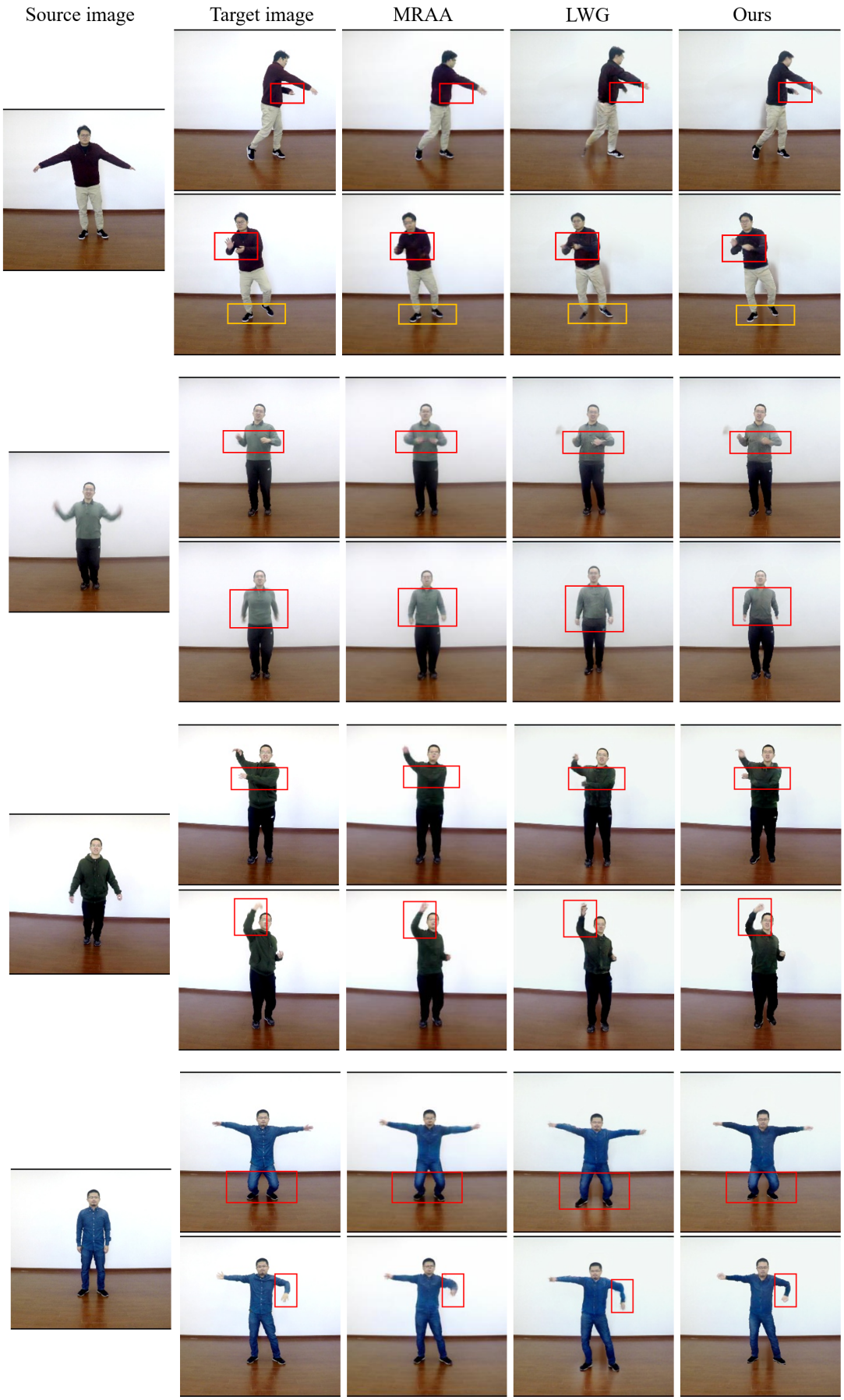


Figure 41: Additional qualitative comparison of existing methods on iPER dataset. Rectangles focus on pose accuracy.

7

SUMMARY AND CONCLUSIONS

7.1 SUMMARY

This thesis explores geometric modeling for 3D human pose estimation and motion transfer. The focus of thesis is how to *i)* utilize the geometric modeling to deal with 3D human pose estimation for both pinhole and fisheye cameras and *ii)* perform human motion transfer with pose consistency by means of 2D and 3D information of humans. We first aim to improve the generalization of single-person 3D HPE for in-the-wild images. Due to the widespread use of fisheye cameras, we then focus on egocentric 3D HPE from a single image captured by a fisheye camera. Then, the 3D HPE problem is addressed in a more general setting, *i.e.,* multi-person 3D pose estimation from a single image captured by a fisheye camera. After that, we propose a benchmark for multi-person 3D pose estimation and action recognition for a fisheye camera. Finally, we combine robust 2D and structural 3D human information to steer and guide the animation process for human motion transfer with pose consistency.

A brief summary of each chapter is as follows:

Chapter 2: Scaled Orthographic Projection for 3D Human Pose Estimation. This chapter aims to improve the generalization of methods for 3D human pose estimation from a single image. Public datasets for 3D human poses are collected in indoor environments due to the limitations of motion capture systems. In contrast to indoor datasets, 2D in-the-wild images may include strongly varying image conditions. Therefore, models trained on such datasets may have limiting generalization capabilities to in-the-wild images. To this end, we propose a re-projection based method to connect 3D poses and 2D poses. To avoid suffering from a small angle problem resulting in overfitting in the depth dimension, we propose an orthographic projection linear regression method to constrain 3D predictions, 2D poses and 2D appearances. Experiments demonstrate the effectiveness and generalization ability of the proposed method qualitatively and quantitatively.

Chapter 3: Egocentric 3D Human Pose Estimation from the Fisheye Camera. The goal of this chapter is to estimate egocentric 3D human poses from a single image captured by a fisheye camera. Due to the fisheye lens, image distortions may negatively influence 3D poses when 2D poses on the image plane pass through the line of sight of the fisheye lens. To mitigate the effect of distortions on the 3D human pose estimation, we propose an automatic calibration module with self-correction to regularize 3D predictions. In contrast to existing methods, the proposed calibration module automatically estimates the intrinsic and distortion camera parameters to perform the estimation pro-

cess. Experimental results show that our method obtains state-of-the-art performance on the modified *x*R-EgoPose containing different levels of image distortions compared to existing methods.

Chapter 4: Multi-person 3D Pose Estimation from the Fisheye Camera. In this chapter, we propose a novel top-down approach for multi-person 3D pose estimation from a single image taken by a fisheye camera. With the wide angle, fisheye cameras have been widely used in practical applications, especially video surveillance. In this chapter, we focus on this scenario *i.e.,* the top-down viewpoint. There are three challenges: *i)* humans at different positions may suffer from varying distortion strengths; *ii)* the distance between humans and cameras is not fixed; *iii)* predicting 3D human joint locations with absolute depths.

To this end, the proposed framework consists of two branches: HPoseNet for root-relative 3D human pose estimation and HRootNet for absolute depth estimation. Finally, we propose 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. In this way, the impact of image distortion is alleviated, and absolute depths of root joints are regularized. The proposed method achieves state-of-the-art performance on both synthesized and real-world datasets.

Chapter 5: A Benchmark for 3D Human Pose Estimation & Action Recognition. This chapter aims to provide a real-world dataset collected by a fisheye camera for 3D human pose estimation and skeleton-based action recognition. Experimental results on the proposed dataset demonstrate that the current methods for pinhole cameras cannot achieve superior performance with respect to fisheye cameras. But, the method considering image distortions shows to be promising and outperforms other methods. Further, to present a complete picture of the above two tasks, a comprehensive survey is provided on the recent advances of 3D human pose estimation and action recognition for both perspective and fisheye cameras.

Chapter 6: Human Motion Transfer with Pose Consistency. The goal of this chapter is to animate a human in the source image based on the pose exhibited in the target images with high reconstruction quality as well as pose consistency. Existing methods either use extracted 2D or 3D information to build the relation of humans in source and target images. The main challenge for methods using 3D information is to maintain pose consistency between synthesized images and target images, since the estimated 3D information (3D human poses or models) may not be accurate. On the other hand, for 2D-based methods, the ambiguity in source poses makes it difficult to obtain high quality of synthesised images due to ambiguous human co-parts. Therefore, we propose a novel method for human motion transfer focusing on both the reconstruction quality as well as pose consistency. Our method combines the warped features in both 2D- and 3D-space using the proposed fusion map to alleviate the self-occlusion problem. In this way, our method benefits from 2D (robustness) and 3D (steering) information to guide the generation process. To maintain pose consistency, a strategy is proposed to retain semantic consistency at arm/leg local regions between synthesized and target images. Experiments and an ablation study demonstrate the effectiveness of our method in terms of the quality of synthesized images and pose consistency.

7.2 CONCLUSIONS

This thesis has studied the effect of geometric modelling on 3D human pose estimation from both pinhole and fisheye cameras. Also, the benefit of robust 2D and structural 3D human information on human motion transfer with pose consistency have been demonstrated. We hope that our research can contribute to the development of the more advanced computer vision algorithms.

An interesting direction is how to deal with the influence of strong perspective effects caused by fisheye lenses on 3D HPE. It is expected that the absolute depth estimation of humans from a singe image captured by a fisheye camera will be more challenging, but at the same time more valuable for practical applications. Another promising direction is how to obtain and use prior information to perform motion transfer on arbitrary objects.

BIBLIOGRAPHY

[1] T. Ahmad, L. Jin, L. Lin, and G. Tang. Skeleton-based action recognition using sparse spatio-temporal gcn with edge effective resistance. *Neurocomputing*, 423:389–398, 2021.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014.

[3] D. Avola, M. Cascio, L. Cinque, G. L. Foresti, C. Massaroni, and E. Rodola. 2-d skeleton-based action recognition via two-branch stacked lstm-rnns. *IEEE Transactions on Multimedia*, 22(10):2481–2496, 2019.

[4] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, pages 8340–8348, 2018.

[5] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, pages 1669–1676, 2014.

[6] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578, 2016.

[7] I. Borg and P. J. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[8] E. Brau and H. Jiang. 3d human pose estimation via deep learning from 2d annotations. In *3DV*, pages 582–591. IEEE, 2016.

[9] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[10] C. Caetano, J. Sena, F. Brémond, J. A. Dos Santos, and W. R. Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[11] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017.

[13] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[14] J. Carreira, E. Noland, C. Hillier, and A. Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[15] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. In *ICCV*, pages 5933–5942, 2019.

[16] C.-H. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *CVPR*, pages 7035–7043, 2017.

[17] C.-H. Chen, A. Tyagi, A. Agrawal, D. Drover, S. Stojanov, and J. M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*, pages 5714–5724, 2019.

[18] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3DV*, pages 479–488. IEEE, 2016.

[19] Y. Chen, G. Ma, C. Yuan, B. Li, H. Zhang, F. Wang, and W. Hu. Graph convolutional network with structure pooling and joint-wise channel attention for action recognition. *Pattern Recognition*, 103:107321, 2020.

[20] Y. Chen, Y. Tian, and M. He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020.

[21] Y. Chen, L. Wang, C. Li, Y. Hou, and W. Li. Convnets-based action recognition from skeleton motion maps. *Multimedia Tools and Applications*, 79(3):1707–1725, 2020.

[22] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, pages 7103–7112, 2018.

[23] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, pages 13359–13368, 2021.

[24] Z. Chen, Y. Huang, H. Yu, and L. Wang. Learning a robust part-aware monocular 3d human pose estimator via neural architecture search. *International Journal of Computer Vision*, 130(1):56–75, 2022.

[25] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *ECCV*, pages 536–553. Springer, 2020.

[26] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu. Skeleton-based action recognition with shift graph convolutional network. In *CVPR*, pages 183–192, 2020.

[27] Y. Cheng, B. Wang, B. Yang, and R. T. Tan. Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. In *AAAI*, pages 1157–1165, 2021.

[28] Y. Cheng, B. Wang, B. Yang, and R. T. Tan. Monocular 3d multi-person pose estimation by integrating top-down and bottom-up networks. In *CVPR*, pages 7649–7659, 2021.

[29] H. Cho, Y. Cho, J. Yu, and J. Kim. Camera distortion-aware 3d human pose estimation in video with optimization-based meta-learning. In *ICCV*, pages 11169–11178, 2021.

[30] H. Ci, C. Wang, X. Ma, and Y. Wang. Optimizing network structure for 3d human pose estimation. In *ICCV*, pages 2262–2271, 2019.

[31] R. Cucchiara and M. Fabbri. Fine-grained human analysis under occlusions and perspective constraints in multimedia surveillance. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s):1–23, 2022.

[32] R. Dabral, N. B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan, and A. Jain. Multi-person 3d human pose estimation from monocular images. In *3DV*, pages 405–414. IEEE, 2019.

[33] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain. Learning 3d human pose from structure and motion. In *ECCV*, pages 668–683, 2018.

[34] C. Ding, K. Liu, J. Korhonen, and E. Belyaev. Spatio-temporal difference descriptor for skeleton-based action recognition. In *AAAI*, volume 35, pages 1227–1235, 2021.

[35] J. Dong, Y. Gao, H. J. Lee, H. Zhou, Y. Yao, Z. Fang, and B. Huang. Action recognition based on the fusion of graph convolutional networks with high order features. *Applied Sciences*, 10(4):1482, 2020.

[36] Z. Dong, J. Song, X. Chen, C. Guo, and O. Hilliges. Shape-aware multi-person pose estimation from multi-view images. In *ICCV*, pages 11158–11168, 2021.

[37] E. S. dos Reis, L. A. Seewald, R. S. Antunes, V. F. Rodrigues, R. da Rosa Righi, C. A. da Costa, L. G. da Silveira Jr, B. Eskofier, A. Maier, T. Horz, et al. Monocular multi-person pose estimation: A survey. *Pattern Recognition*, 118:108046, 2021.

[38] D. Drover, C.-H. Chen, A. Agrawal, A. Tyagi, and C. Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *ECCV Workshops*, pages 0–0, 2018.

[39] Y. Du, Y. Fu, and L. Wang. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 579–583. IEEE, 2015.

[40] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, pages 1110–1118, 2015.

[41] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, pages 2334–2343, 2017.

[42] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI*, 2018.

[43] R. Friji, H. Drira, F. Chaieb, H. Kchok, and S. Kurtek. Geometric deep neural network using rigid and non-rigid transformations for human action recognition. In *ICCV*, pages 12611–12620, 2021.

[44] J. Gao, T. Zhang, and C. Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, volume 33, pages 8303–8311, 2019.

[45] J. Gao, T. Zhang, and C. Xu. Learning to model relationships for zero-shot video classification. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3476–3491, 2020.

[46] X. Gao, W. Hu, J. Tang, J. Liu, and Z. Guo. Optimized skeleton-based action recognition via sparsified graph regression. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 601–610, 2019.

[47] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019.

[48] P. Ghosh, Y. Yao, L. Davis, and A. Divakaran. Stacked spatio-temporal graph convolutional networks for action segmentation. In *WACV*, pages 576–585, 2020.

[49] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[50] Y. Guo, L. Ma, Z. Li, X. Wang, and F. Wang. Monocular 3d multi-person pose estimation via predicting factorized correction factors. *Computer Vision and Image Understanding*, 213:103278, 2021.

[51] I. Habibie, W. Xu, D. Mehta, G. Pons-Moll, and C. Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *CVPR*, pages 10905–10914, 2019.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[53] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[54] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.

[55] G. Hidalgo, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh. Single-network whole-body pose estimation. In *ICCV*, pages 6982–6991, 2019.

[56] Y. Hou, Z. Li, P. Wang, and W. Li. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(3):807–811, 2016.

[57] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[58] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *CVPR*, pages 5344–5352, 2015.

[59] J.-F. Hu, W.-S. Zheng, J. Pan, J. Lai, and J. Zhang. Deep bilinear learning for rgb-d action recognition. In *ECCV*, pages 335–351, 2018.

[60] J. Huang, X. Xiang, X. Gong, B. Zhang, et al. Long-short graph memory network for skeleton-based action recognition. In *WACV*, pages 645–652, 2020.

[61] Z. Huang, X. Han, J. Xu, and T. Zhang. Few-shot human motion transfer by personalized geometry and texture modeling. In *CVPR*, pages 2297–2306, 2021.

[62] C. Hughes, M. Glavin, E. Jones, and P. Denny. Wide-angle camera technology for automotive applications: a review. *IET Intelligent Transport Systems*, 3(1):19–31, 2009.

[63] D.-H. Hwang, K. Aso, Y. Yuan, K. Kitani, and H. Koike. 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*, pages 98–111, 2020.

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

[65] E. Jahangiri and A. L. Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *ICCV Workshops*, pages 805–814, 2017.

[66] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *NeurIPS*, 31, 2018.

[67] H. Jiang and K. Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *CVPR*, pages 3501–3509. IEEE, 2017.

[68] S. Jin, W. Liu, W. Ouyang, and C. Qian. Multi-person articulated tracking with spatial and temporal embeddings. In *CVPR*, pages 5664–5673, 2019.

[69] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[70] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[71] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, pages 3334–3342, 2015.

[72] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018.

[73] J. Kannala and S. Brandt. A generic camera calibration method for fish-eye lenses. In *ICPR*, volume 1, pages 10–13. IEEE, 2004.

[74] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[75] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[76] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, pages 3288–3297, 2017.

[77] H. Kim, J. Jung, and J. Paik. Fisheye lens camera based surveillance system for wide field of view monitoring. *Optik*, 127(14):5636–5646, 2016.

[78] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[79] M. Knoche, I. Sárándi, and B. Leibe. Reposing humans by warping 3d features. In *CVPR Workshops*, pages 1044–1045, 2020.

[80] M. Kocabas, S. Karagoz, and E. Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *ECCV*, pages 417–433, 2018.

[81] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *CVPR*, pages 1077–1086, 2019.

[82] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019.

[83] Y. Kong and Y. Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[84] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[85] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, 2011.

[86] J. Y. Lee, J. DeGol, V. Fragoso, and S. N. Sinha. Patchmatch-based neighborhood consensus for semantic correspondence. In *CVPR*, pages 13153–13163, 2021.

[87] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 601–604. IEEE, 2017.

[88] C. Li, Z. Cui, W. Zheng, C. Xu, R. Ji, and J. Yang. Action-attending graphic neural network. *IEEE Transactions on Image Processing*, 27(7):3657–3670, 2018.

[89] C. Li, Y. Hou, P. Wang, and W. Li. Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Letters*, 24(5):624–628, 2017.

[90] C. Li and G. H. Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *CVPR*, pages 9887–9895, 2019.

[91] C. Li and G. H. Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *CVPR*, pages 9887–9895, 2019.

[92] F. Li, A. Zhu, Y. Xu, R. Cui, and G. Hua. Multi-stream and enhanced spatial-temporal graph convolution network for skeleton-based action recognition. *IEEE Access*, 8:97757–97770, 2020.

[93] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, pages 11025–11034, 2021.

[94] J. Li, C. Wang, W. Liu, C. Qian, and C. Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, pages 242–259, 2020.

[95] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, pages 10863–10872, 2019.

[96] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 3595–3603, 2019.

[97] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[98] S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, pages 332–347. Springer, 2014.

[99] S. Li, K. Han, T. W. Costain, H. Howard-Jenkins, and V. Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *CVPR*, pages 10196–10205, 2020.

[100] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *CVPR*, pages 5457–5466, 2018.

[101] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *ICCV*, pages 2848–2856, 2015.

[102] T. Li, Q. Ke, H. Rahmani, R. E. Ho, H. Ding, and J. Liu. Else-net: Elastic semantic network for continual action recognition from skeleton data. In *ICCV*, pages 13434–13443, 2021.

[103] W. Li, X. Liu, Z. Liu, F. Du, and Q. Zou. Skeleton-based action recognition using multi-scale and multi-stream improved graph convolutional network. *IEEE Access*, 8:144529–144542, 2020.

[104] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 9–14. IEEE, 2010.

[105] Y. Li, C. Huang, and C. C. Loy. Dense intrinsic appearance flow for human pose transfer. In *CVPR*, pages 3693–3702, 2019.

[106] J. Liang and M. C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *ICCV*, pages 4352–4362, 2019.

[107] J. Lin and G. H. Lee. Hdnet: Human depth estimation for multi-person camera-space localization. In *ECCV*, pages 633–648, 2020.

[108] J. Lin and G. H. Lee. Multi-view multi-person 3d pose estimation with plane sweep stereo. In *CVPR*, pages 11886–11895, 2021.

[109] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[110] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

[111] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, pages 816–833. Springer, 2016.

[112] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2017.

[113] K. Liu, R. Ding, Z. Zou, L. Wang, and W. Tang. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *ECCV*, pages 318–334. Springer, 2020.

[114] K. Liu, L. Gao, N. M. Khan, L. Qi, and L. Guan. A multi-stream graph convolutional networks-hidden conditional random field model for skeleton-based action recognition. *IEEE Transactions on Multimedia*, 23:64–76, 2020.

[115] M. Liu, H. Liu, and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[116] M. Liu and J. Yuan. Recognizing human actions as the evolution of pose estimation maps. In *CVPR*, pages 1159–1168, 2018.

[117] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, pages 5904–5913, 2019.

[118] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017.

[119] X.-C. Liu, Y.-L. Yang, and P. Hall. Learning to warp for style transfer. In *CVPR*, pages 3702–3711, 2021.

[120] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020.

[121] L. Lu, Y. Lu, R. Yu, H. Di, L. Zhang, and S. Wang. Gaim: Graph attention interaction model for collective activity recognition. *IEEE Transactions on Multimedia*, 22(2):524–539, 2019.

[122] C. Luo, X. Chu, and A. Yuille. Orinet: A fully convolutional network for 3d human pose estimation. In *BMVC*, 2018.

[123] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, pages 5137–5146, 2018.

[124] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. *NeurIPS*, 30, 2017.

[125] D. Mallis, E. Sanchez, M. Bell, and G. Tzimiropoulos. Unsupervised learning of object landmarks via self-training correspondence. *NeurIPS*, 33:4709–4720, 2020.

[126] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pages 2640–2649, 2017.

[127] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516. IEEE, 2017.

[128] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 39(4):82–1, 2020.

[129] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, pages 120–130. IEEE, 2018.

[130] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[131] G. Moon, J. Y. Chang, and K. M. Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, pages 10133–10142, 2019.

[132] G. Moon, J. Y. Chang, and K. M. Lee. Posefix: Model-agnostic general human pose refinement network. In *CVPR*, pages 7773–7781, 2019.

[133] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR*, pages 2823–2832, 2017.

[134] O. Moskvyak, F. Maire, F. Dayoub, and M. Baktashmotlagh. Semi-supervised keypoint localization. *arXiv preprint arXiv:2101.07988*, 2021.

[135] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, pages 2277–2287, 2017.

[136] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

[137] W. Ng, M. Zhang, and T. Wang. Multi-localized sensitive autoencoder-attention-lstm for skeleton-based action recognition. *IEEE Transactions on Multimedia*, 24:1678–1690, 2021.

[138] X. Nie, J. Feng, J. Zhang, and S. Yan. Single-stage multi-person pose machines. In *ICCV*, pages 6951–6960, 2019.

[139] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, pages 484–494, 2018.

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

[141] B. Parsa, B. Dariush, et al. Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment. In *WACV*, pages 1080–1090, 2020.

[142] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *CVPR*, pages 7307–7316, 2018.

[143] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, pages 7025–7034, 2017.

[144] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018.

[145] W. Peng, X. Hong, H. Chen, and G. Zhao. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *AAAI*, volume 34, pages 2669–2676, 2020.

[146] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *ECCV*, pages 101–117, 2018.

[147] J. Ren, M. Chai, O. J. Woodford, K. Olszewski, and S. Tulyakov. Flow guided transformable bottleneck networks for motion retargeting. In *CVPR*, pages 10795–10805, 2021.

[148] J. Ren, N. Reyes, A. Barczak, C. Scogings, and M. Liu. An investigation of skeleton-based optical flow-guided features for 3d action recognition using a multi-stream cnn model. In *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pages 199–203. IEEE, 2018.

[149] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016.

[150] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *ECCV*, pages 750–767, 2018.

[151] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, pages 6148–6157, 2017.

[152] G. Rochette, C. Russell, and R. Bowden. Weakly-supervised 3d pose estimation from a single image using multi-view consistency. In *BMVC*, 2019.

[153] G. Rogez, J. S. Supancic, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *CVPR*, pages 4325–4333, 2015.

[154] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, pages 3433–3441, 2017.

[155] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1146–1161, 2019.

[156] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe. MeTRAbs: metric-scale truncation-robust heatmaps for absolute 3D human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):16–30, 2021.

[157] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016.

[158] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *ICCV*, pages 2325–2334, 2019.

[159] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Skeleton-based action recognition with directed graph neural networks. In *CVPR*, pages 7912–7921, 2019.

[160] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019.

[161] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins. Motion capture from body-mounted cameras. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011.

[162] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *CVPR*, pages 1227–1236, 2019.

[163] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *ECCV*, pages 103–118, 2018.

[164] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, pages 2377–2386, 2019.

[165] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. *NeurIPS*, 32:7137–7147, 2019.

[166] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable gans for pose-based human image generation. In *CVPR*, pages 3408–3416, 2018.

[167] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov. Motion representations for articulated animation. In *CVPR*, pages 13653–13662, 2021.

[168] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010.

[169] Y.-F. Song, Z. Zhang, and L. Wang. Richly activated graph convolutional network for action recognition with incomplete skeletons. In *ICIP*, pages 1–5. IEEE, 2019.

[170] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[171] K. Su, X. Liu, and E. Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *CVPR*, pages 9631–9640, 2020.

[172] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019.

[173] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *ICCV*, pages 2602–2611, 2017.

[174] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. In *ECCV*, pages 529–545, 2018.

[175] Y. Tang, Y. Wei, X. Yu, J. Lu, and J. Zhou. Graph interaction networks for relation transfer in human activity videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2872–2886, 2020.

[176] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3d human pose with deep neural networks. *BMVC*, 2016.

[177] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *ICCV*, pages 3941–3950, 2017.

[178] J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *ICCV*, pages 6361–6371, 2019.

[179] D. Tome, T. Alldieck, P. Peluse, G. Pons-Moll, L. Agapito, H. Badino, and F. De la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[180] D. Tome, P. Peluse, L. Agapito, and H. Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *ICCV*, pages 7728–7738, 2019.

[181] Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan. Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition. *IEEE Transactions on Multimedia*, 2022.

[182] F. Van den Heuvel, R. Verwaal, and B. Beers. Automated calibration of fisheye camera systems and the reduction of chromatic aberration. *PHOTOGRAMMETRIE FERNERKUNDUNG GEOINFORMATION*, 2007(3):157, 2007.

[183] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117, 2017.

[184] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, volume 36, pages 349–360. Wiley Online Library, 2017.

[185] B. Wandt and B. Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *CVPR*, pages 7782–7791, 2019.

[186] C. Wang, J. Li, W. Liu, C. Qian, and C. Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, pages 242–259. Springer, 2020.

[187] J. Wang, L. Liu, W. Xu, K. Sarkar, and C. Theobalt. Estimating egocentric 3d human pose in global space. In *ICCV*, pages 11500–11509, 2021.

[188] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, pages 1290–1297. IEEE, 2012.

[189] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning and recognition. In *CVPR*, pages 2649–2656, 2014.

[190] M. Wang, X. Chen, L. Liu, C. Qian, L. Lin, and L. Ma. Drpose3d: Depth ranking in 3d human pose estimation. In *Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 978–984, 2018.

[191] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro. Few-shot video-to-video synthesis. In *NeurIPS*, 2019.

[192] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018.

[193] X. Wang and A. Gupta. Videos as space-time region graphs. In *ECCV*, pages 399–417, 2018.

[194] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[195] P. Wei, H. Sun, and N. Zheng. Learning composite latent structures for 3d human action representation and recognition. *IEEE Transactions on Multimedia*, 21(9):2195–2208, 2019.

[196] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016.

[197] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia. Graph cnns with motif and variable temporal block for skeleton-based action recognition. In *AAAI*, volume 33, pages 8989–8996, 2019.

[198] C. Wu, X.-J. Wu, and J. Kittler. Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition. In *ICCV Workshops*, 2019.

[199] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *ECCV*, pages 365–382. Springer, 2016.

[200] S. Wu, S. Jin, W. Liu, L. Bai, C. Qian, D. Liu, and W. Ouyang. Graph-based 3d multi-person pose estimation using multi-view images. In *ICCV*, pages 11148–11157, 2021.

[201] L. Xia, C.-C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 20–27. IEEE, 2012.

[202] R. Xia, Y. Li, and W. Luo. Laga-net: Local-and-global attention network for skeleton based action recognition. *IEEE Transactions on Multimedia*, 24:2648–2661, 2021.

[203] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018.

[204] T. Xu and W. Takano. Graph stacked hourglass networks for 3d human pose estimation. In *CVPR*, pages 16105–16114, 2021.

[205] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt. Mo2cap2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019.

[206] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223:106970, 2021.

[207] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[208] H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li, and S. J. Maybank. Feedback graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 31:164–175, 2021.

[209] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, pages 5255–5264, 2018.

[210] H. Yonemoto, K. Murasaki, T. Osawa, K. Sudo, J. Shimamura, and Y. Taniguchi. Egocentric articulated pose tracking for action recognition. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pages 98–101. IEEE, 2015.

[211] P. Zablotskaia, A. Siarohin, B. Zhao, and L. Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019.

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

[213] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, and C. Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *NeurIPS*, pages 8410–8419, 2018.

[214] A. Zeng, X. Sun, F. Huang, M. Liu, Q. Xu, and S. Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *ECCV*, pages 507–523. Springer, 2020.

[215] H. Zhang, Y. Song, and Y. Zhang. Graph convolutional lstm model for skeleton-based action recognition. In *ICME*, pages 412–417. IEEE, 2019.

[216] J. Zhang, F. Shen, X. Xu, and H. T. Shen. Temporal reasoning graph for activity recognition. *IEEE Transactions on Image Processing*, 29:5491–5506, 2020.

[217] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1963–1978, 2019.

[218] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, pages 1112–1121, 2020.

[219] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng. Adding attentiveness to the neurons in recurrent neural networks. In *ECCV*, pages 135–151, 2018.

[220] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[221] S. Zhang, X. Liu, and J. Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *WACV*, pages 148–157. IEEE, 2017.

[222] Y. Zhang, S. You, and T. Gevers. Automatic calibration of the fisheye camera for egocentric 3d human pose estimation from a single image. In *WACV*, pages 1772–1781, 2021.

[223] Y. Zhang, S. You, and T. Gevers. Orthographic projection linear regression for single image 3d human pose estimation. In *ICPR*, pages 8109–8116. IEEE, 2021.

[224] Y. Zhang, S. You, S. Karaoglu, and T. Gevers. Monocular 3d human pose estimation and action recognition using fisheye cameras: A survey and benchmark. 2022.

[225] Y. Zhang, S. You, S. Karaoglu, and T. Gevers. Multi-person 3d pose estimation from a single image captured by a fisheye camera. *Computer Vision and Image Understanding*, page 103505, 2022.

[226] Y. Zhang, S. You, S. Karaoglu, and T. Gevers. Pose guided human motion transfer by exploiting 2d and 3d information. In *3DV*, 2022.

[227] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, pages 3425–3435, 2019.

[228] R. Zhao, K. Wang, H. Su, and Q. Ji. Bayesian graph convolution lstm for skeleton based action recognition. In *ICCV*, pages 6882–6892, 2019.

[229] T. Zhao, S. Li, K. N. Ngan, and F. Wu. 3-d reconstruction of human body shape from a single commodity depth camera. *IEEE Transactions on Multimedia*, 21(1):114–123, 2018.

[230] J. Zhen, Q. Fang, J. Sun, W. Liu, W. Jiang, H. Bao, and X. Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *ECCV*, pages 550–566, 2020.

[231] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *AAAI*, volume 32, 2018.

[232] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *ICCV*, pages 2344–2353, 2019.

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

[234] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *CVPR*, pages 4447–4455, 2015.

[235] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, pages 4966–4975, 2016.

[236] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):901–914, 2018.

[237] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *ICCV*, pages 2904–2913, 2017.

7.3 SAMENVATTING

Dit proefschrift onderzoekt geometrische modellering voor menselijke 3D-houdingsschatting en bewegingsoverdracht. De focus van het proefschrift is *(i)* hoe de geometrische modellering gebruikt zou kunnen worden voor menselijke 3D-houdingsschatting voor zowel pinhole- als fisheye-camera's en *(ii)* hoe menselijke bewegingsoverdracht uitgevoerd kan worden door middel van 2D en 3D houdingsconsistente informatie van mensen. We willen eerst de veralgemening van enkel-persoons 3D HPE (Engels: Human Pose Estimation, Nederlands: Menselijke Houdingsschatting) voor in-het-wild afbeeldingen verbeteren. Vanwege het wijdverbreide gebruik van fisheye-camera's, richten we ons vervolgens op egocentrische 3D HPE van een enkel beeld dat is vastgelegd met een fisheye-camera. Vervolgens wordt het 3D HPE-probleem aangepakt in een meer algemene setting, *d.w.z.* 3D-houdingschatting voor meerdere personen op basis van een enkel beeld vastgelegd door een fisheye-camera. Daarna stellen we een benchmark voor voor het schatten van 3D-houdingen voor meerdere personen en actieherkenning voor een fisheye-camera. Ten slotte combineren we robuuste 2D- en structurele 3D-menselijke informatie om het animatieproces voor menselijke bewegingsoverdracht met houdingsconsistentie te sturen en te begeleiden.

Een korte samenvatting van elk hoofdstuk is als volgt:

Hoofdstuk 2: Geschaalde Orthografische Projectie voor 3D Menselijke Houdingsschatting. Dit hoofdstuk heeft tot doel de veralgemening van methoden voor het schatten van menselijke poses in 3D op basis van een enkel beeld te verbeteren. Openbare datasets voor 3D-menselijke poses worden verzameld in binnenomgevingen vanwege de beperkingen van motion capture-systemen. In tegenstelling tot indoor datasets kunnen 2D in-het-wild beelden sterk wisselende beeldomstandigheden bevatten. Daarom kunnen modellen die op dergelijke datasets zijn getraind, beperkende generalisatiemogelijkheden hebben voor in-het-wild afbeeldingen. Hiertoe stellen we een op herprojectie gebaseerde methode voor om 3D-houdingen en 2D-houdingen met elkaar te verbinden. Om het probleem van een te-kleine-hoek te voorkomen – wat zou resulteren in overfitting in de dieptedimensie – stellen we een orthografische projectie lineaire regressiemethode voor om 3D-voorspellingen, 2D-poses en 2D-verschijningen te beperken. Experimenten tonen de effectiviteit en het generalisatievermogen van de voorgestelde methode kwalitatief en kwantitatief aan.

Hoofdstuk 3: Egocentrische 3D Menselijke Houdingsschatting met een Fisheye Camera. Het doel van dit hoofdstuk is om egocentrische 3D-menselijke houdingen te schatten op basis van een enkel beeld dat is vastgelegd met een fisheye-camera. Vanwege de fisheye-lens kunnen beeldvervormingen een negatieve invloed hebben op 3D-houdingen wanneer 2D-houdingen op het beeldvlak door de zichtlijn van de fisheye-lens gaan. Om het effect van vervormingen op de schatting van de menselijke pose in 3D te verminderen, stellen we een automatische kalibratiemodule voor met zelfcorrectie om 3D-voorspellingen te regulariseren. In tegenstelling tot bestaande methoden schat de voorgestelde kalibratiemodule automatisch de intrinsieke en vervormingscameraparameters om het schattingsproces uit te voeren. Experimentele resultaten laten zien dat onze methode state-of-the-art prestaties verkrijgt op de gewijzigde *x*R-EgoPose die verschillende niveaus van beeldvervormingen bevat in vergelijking met bestaande methoden.

Hoofdstuk 4: Meer-persoons 3D Houdingsschatting met een Fisheye Camera. In dit hoofdstuk stellen we een nieuwe top-down benadering voor voor het schatten van 3D-poses van meerdere personen op basis van een enkel beeld gemaakt met een fisheye-camera. Door het grote waarneembare beeld worden fisheye-camera's op grote schaal gebruikt in praktische toepassingen, met name videobewaking. In dit hoofdstuk concentreren we ons op dit scenario *d.w.z.* het top-down gezichtspunt. Er zijn drie uitdagingen: *(i)* Mensen op verschillende posities kunnen last hebben van verschillende vervormingssterkten; *(ii)* de afstand tussen mens en camera staat niet vast; *(iii)* voorspellen van de 3D menselijke gewrichtslocaties en absolute diepten is lastig.

Hiertoe bestaat het voorgestelde raamwerk uit twee takken: HPoseNet voor wortel-relatieve 3D menselijke houdingsschatting en HRootNet voor absolute diepte-schatting. Ten slotte stellen we een herprojectiemodule voor om de twee takken te verbinden, waardoor de geschatte 3D-menselijke houdingen consistent zijn met de 2D-houdingen onder vervormingen door de herprojectiefout te minimaliseren. Op deze manier wordt de impact van beeldvervorming verminderd en worden de absolute diepten van wortelvoegen geregulariseerd. De voorgestelde methode bereikt state-of-the-art prestaties op zowel gesynthetiseerde als real-world datasets.

Hoofdstuk 5: Een Benchmark voor 3D Menselijke Houdingsschatting & Actie Herkenning. Dit hoofdstuk is bedoeld om een real-world dataset aan te bieden die is verzameld door een fisheye-camera voor 3D-schatting van menselijke houdingen en skeletgebaseerde actieherkenning. Experimentele resultaten op de voorgestelde dataset tonen aan dat de huidige methoden voor pinhole-camera's geen superieure prestaties kunnen bereiken met betrekking tot fisheye-camera's. Maar een methode die rekening houdt met beeldvervormingen, blijkt veelbelovend te zijn en presteert beter dan andere methoden. Om een compleet beeld van de bovenstaande twee taken te geven, wordt verder een uitgebreid overzicht gegeven van de recente ontwikkelingen op het gebied van 3D-schatting van menselijke poses en actieherkenning voor zowel perspectief- als fisheye-camera's.

Hoofdstuk 6: Menselijke Bewegingsoverdracht met Houding Consistentie. Het doel van dit hoofdstuk is om een mens in de bronafbeelding te animeren op basis van de pose die wordt vertoond in de doelafbeeldingen met een hoge reconstructiekwaliteit en consistentie van de houding. Bestaande methoden gebruiken ofwel geëxtraheerde 2D- of 3D-informatie om de relatie van mens in bron- en doelbeelden op te bouwen. De belangrijkste uitdaging voor methoden die 3D-informatie gebruiken, is om de poseconsistentie tussen gesynthetiseerde beelden en doelbeelden te behouden, aangezien de geschatte 3D-informatie (3D menselijke poses of modellen) mogelijk niet nauwkeurig is. Aan de andere kant – voor op 2D gebaseerde methoden – maakt de ambiguïteit in bronhoudingen het moeilijk om gesynthetiseerde afbeeldingen van hoge kwaliteit te verkrijgen vanwege dubbelzinnige menselijke co-parts. Daarom stellen we een nieuwe methode voor die de overdracht van menselijke bewegingen mogelijk maakt, waarbij op zowel de kwaliteit van de reconstructie als op de consistentie van de houding de nadruk ligt. Onze methode combineert de kromgetrokken functies in zowel 2D- als 3D-ruimte met behulp van de voorgestelde fusiekaart om het zelfocclusieprobleem te verlichten. Op deze manier profiteert onze methode van 2D (robuustheid) en 3D (stuur) informatie om het generatieproces te begeleiden. Om de poseconsistentie te behouden, wordt een strategie voorgesteld om de semantische consistentie te behouden in lokale arm/beenregio's tussen gesynthetiseerde en doelbeelden. Experimenten en een ablatiestudie tonen de effectiviteit van onze methode aan in termen van de kwaliteit van gesynthetiseerde beelden en poseconsistentie.

7.4 CONCLUSIE

Dit proefschrift heeft het effect bestudeerd van geometrische modellering op 3D-schatting van menselijke houdingen van zowel pinhole- als fisheye-camera's. Ook is het voordeel aangetoond van robuuste 2D en structurele 3D menselijke informatie over menselijke bewegingsoverdracht met houdingsconsistentie. We hopen dat ons onderzoek kan bijdragen aan de ontwikkeling van de meer geavanceerde computer vision-algoritmen.

Een interessante richting is hoe om te gaan met de invloed van sterke perspectiefeffecten veroorzaakt door fisheye-lenzen op 3D HPE. Verwacht wordt dat het schatten van de absolute diepte van mensen op basis van een enkel beeld vastgelegd door een fisheye-camera uitdagender zal zijn, maar tegelijkertijd waardevoller voor praktische toepassingen. Een andere veelbelovende richting is het verkrijgen en gebruiken van voorafgaande informatie om bewegingsoverdracht op willekeurige objecten uit te voeren.

ACKNOWLEDGMENTS

I decided to start this journey because my mother persuaded me. I still remember what my mother said — 'Now that you have a master's degree now, why not consider pursing a PhD degree? Then, you won't have a regret.' Indeed, I am very grateful that I pursue a PhD degree here. This is a challenging but amazing journey. I would like to thank all people who have supported and helped me. Because of you, everything is possible for me.

I would like to thank my promoter, Prof. dr. Theo Gevers. Thanks for giving me the opportunity to join this journey. Thanks for your guidance, trust and support to me. I have learned a lot from you in both academia and personal life in this journey. Under your supervision, I have improved myself in many aspects when I look back including the ability to do research independently, efficient communication, and writing. My appreciation is more than words that I can say.

I would also like to thank my co-promoters, dr. Shaodi You and dr. Sezer Karaoglu. Shaodi has inspired me a lot for my research and the importance of making plans. With his guidance and help, I know how to analyze the research questions and the reason behind the phenomenon. Sezer, thanks for your questions during the meeting and providing the deep insight to my research. With his guidance, I learned how to focus on the target especially when I prepare the rebuttal for the paper submissions.

Next, I would like to thank my committee members, prof. dr. Albert Salah, prof. dr. ir. Ben Krose, prof. dr. Marcel Worring, dr. Hamdi Dibeklioglu, dr. Dimitrios Tzionas. It is a great pleasure and honor to have you in my committee.

Many thanks to my friends: Yuanqing, Longfu, Yixuan, Wenxi, Hui, Yan, Qingkang, Dongwei, Yanhua, Zenglin, Yunlu, Zeshun, Yixian, Yunhua, Jianghua, Yangjun, Yue, Pengjie, Teng, Shuo, Tao, Qi, Yigong, Jin, Xiaotian, Weijia, Hongyun, Zhenghui, Pengwan, Jie, Yongtuo, Jiayi, Zehao, Yingjun, Luyuan, Taichi, Longjiao, Wenyang, Xinyu, Shijia. Thanks to Shiqi and our cat Leon for being with me through the difficult time and sharing the happy time with me in this journey.

Furthermore, I would like to thank all my great colleagues in Computer Vision Lab. Thanks Wei Wang and Rick Groenendijk for being my paranymphs. Wei, we joined our group at the same time, and we always share the experience and opinions on both research and daily life, which enriches the life here. Rick, thanks for organizing 'Gambe' and the card game nights, and it was really fun. Jian Han, thanks for helping me in many aspects. You are always patient and passing on your experience to me. Partha Das, thanks for sharing Indian food and your experience with me. Ozzy Ülger, thanks for organizing BBQs, and it was tasty. Weijie Wei, thanks for the organization of the hotpot and card games. Ruihong Yin, thanks for organizing activities of the volleyball and kart racing, and I really enjoyed it. Wei Zeng, thanks for introducing me and show me around when I arrived the main building of Science Park 904. Thanks to Leo Dorst, Dennis Koelma, Arnoud Visser, Martin Oswald, Yang Liu, Finde Xumara, Anil Baslamisli, Hoang-An Le, Minh Ngo, Hanan ElNaghy, Qi Bi, Xiaoyan Xing, Melis Öcal.

Special thanks to my supervisors when I did my master degree, prof. Xinyu Shao and prof. Ping Jiang for their support and help for my study abroad. Thanks to dr. Qi Zhou for his help and advice on my research and planning.

Finally, I sincerely thank my parents and family for their selfless support, care and encouragement. Due to Covid-19, I haven't been back during this period. I feel very happy every time when I video call you, and you always have a big smile on your faces and it seems you are still by my side. It is you make me better.