Computer Vision and Image Understanding 210 (2021) 103225

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Deep 3D human pose estimation: A review

Jinbao Wang [a,1], Shujie Tan [a,1], Xiantong Zhen [b,e], Shuo Xu [c], Feng Zheng [a,*], Zhenyu He [d], Ling Shao [b]

[a] Department of Computer Science and Engineering, Southern University of Science and Technology, 518055, China
[b] Inception Institute of Artificial Intelligence, Abu Dhabi, The United Arab Emirates
[c] Department of Electronics and Information Engineering, Anhui University, 230601, China
[d] Harbin Institute of Technology (Shenzhen), China
[e] AIM Lab, University of Amsterdam, The Netherlands

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
68-02
68T45
68U10

Keywords:
3D Human Pose Estimation
Deep Learning

ABSTRACT

Three-dimensional (3D) human pose estimation involves estimating the articulated 3D joint locations of a human body from an image or video. Due to its widespread applications in a great variety of areas, such as human motion analysis, human–computer interaction, robots, 3D human pose estimation has recently attracted increasing attention in the computer vision community, however, it is a challenging task due to depth ambiguities and the lack of in-the-wild datasets. A large number of approaches, with many based on deep learning, have been developed over the past decade, largely advancing the performance on existing benchmarks. To guide future development, a comprehensive literature review is highly desired in this area. However, existing surveys on 3D human pose estimation mainly focus on traditional methods and a comprehensive review on deep learning based methods remains lacking in the literature. In this paper, we provide a thorough review of existing deep learning based works for 3D pose estimation, summarize the advantages and disadvantages of these methods and provide an in-depth understanding of this area. Furthermore, we also explore the commonly-used benchmark datasets on which we conduct a comprehensive study for comparison and analysis. Our study sheds light on the state of research development in 3D human pose estimation and provides insights that can facilitate the future design of models and algorithms.

1. Introduction

Human pose estimation is generally regarded as the task of predicting the articulated joint locations of a human body from an image or a sequence of images of that person. Due to its wide range of potential applications, human pose estimation is a fundamental and active research direction in the area of computer vision. Driven by powerful deep learning techniques and recently collected large-scale datasets, human pose estimation has continued making great progress, especially on 2D images. However, the performance of 3D human pose estimation remains barely satisfactory, which could be largely due to the lack of sufficient 3D in-the-wild datasets. Recently, some methods (Trumble et al., 2017; von Marcard et al., 2018) have been proposed to solve this problem, and to a certain extent, these methods have made some progress. However, there is still significant room for improvement.

In this section, we will first introduce the vast number of potential applications of 3D pose estimation to highlight the significance of research in this topic, then discuss the main challenges, and finally describe the scope of this survey in comparison to related work.

1.1. Applications

Since 3D pose representation provides additional depth information compared with 2D pose representation, 3D human pose estimation enables more widespread applications. To better understand the use of 3D human pose estimation, we provide a brief description of some of its interesting real-world applications:

- *Human–Computer Interaction.* A robot can better serve and help users if it can understand 3D poses, actions and emotions of people. For example, a robot can take timely actions when it detects the 3D pose of a person who is prone to fall. In addition, assistant robots can better socially interact with human users, provided they can perceive 3D human poses. Meanwhile, it is also very useful for computer control, *i.e.* as input for productive software packages. Moreover, people can play games using their poses and gestures through Microsoft Kinect sensors (Zhang, 2012).

* Corresponding author.
E-mail address: zhengf@sustech.edu.cn (F. Zheng).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cviu.2021.103225
Received 12 August 2020; Received in revised form 13 May 2021; Accepted 14 May 2021
Available online 24 May 2021

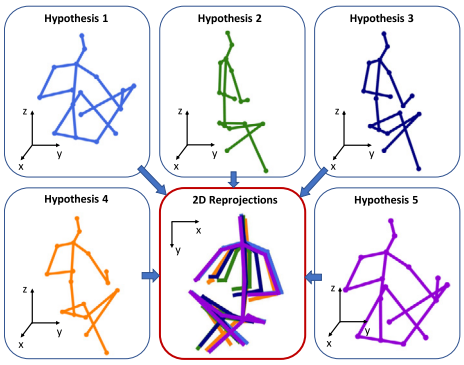

Fig. 1. Illustration of the depth ambiguity (Li and Lee, 2019).

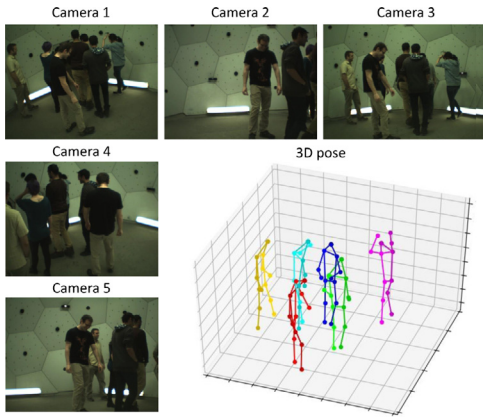

Fig. 2. Illustration of the correspondence of people in different views (Dong et al., 2019).

- *Autonomous Driving.* Self-driving cars are required to make decisions to avoid collision with pedestrians, and thus understanding a pedestrian's pose, movement and intention is very important (Kim et al., 2019; Du et al., 2019).
- *Video Surveillance.* Nowadays, video surveillance is of great significance for public safety. In this area, 3D pose estimation techniques could be used to assist the re-identification task (Su et al., 2017; Xu et al., 2018; Zheng et al., 2019), which helps video surveillance and enables supervisors to quickly find the targets of interest.
- *Biomechanics and Medication.* Human pose and movement can indicate the health status of humans. Thus, 3D pose estimation techniques could be used to construct a sitting posture correction system to monitor the status of users. For exercise, the system can be used to avoid injury by providing timely feedback of correct movement poses to users. Moreover, pose estimation systems are also able to assist doctors for remote diagnose and tele-rehabilitation of patients (Airò Farulla et al., 2016).
- *Sports Performance Analysis and Education.* The automated extraction of 3D poses from videos can help further analysis of the performance of athletes and provide immediate feedback for their improvement (Hwang et al., 2017). Thus, human pose estimation can be used to evaluate and educate people in various forms of sports such as swimming (Zecha et al., 2018), Tai Chi (Scott et al., 2017), soccer (Rematas et al., 2018).
- *Psychology.* 3D human body poses can also reveal the mental states of people and the emotion can even be recognized from poses (Noroozi et al., 2018). Scientists can utilize pose estimation related techniques to quantify behavior for further research (Joo et al., 2017). As a result, human pose estimation can be used for psychology therapy of certain mental diseases such as children autism (Marinoiu et al., 2018).
- *Try-on and Fashion.* Online shopping has become more and more popular in recent years, especially for fashion clothes. Users can see how they look like when wearing a certain piece of clothing on the Internet in a virtual try-on system based on 3D pose estimation (Pons-Moll et al., 2017; Han et al., 2018).
- *Others.* 3D pose estimation can also be used to assist other computer vision tasks such as pose transfer (Li et al., 2019a), action recognition (Luvizon et al., 2018), human parsing (Xia et al., 2016), person image generation (Siarohin et al., 2018), animation (Weng et al., 2019), pose search (Ferrari et al., 2009).

1.2. Challenges

Recently, 3D human pose estimation has become an increasingly popular research topic due to its widespread application. However, it is far from being solved because of its unique challenges in contrast to 2D human pose estimation, in which the main challenges include variations of body poses, complicated backgrounds, diverse clothing appearance and occlusions. 3D human pose estimation faces further challenges, including a lack of in-the-wild 3D datasets, depth ambiguities, a huge demand for rich posture information (such as translations and rotations), a large searching state space for each joint (representing a discretized 3D space), etc. We will discuss the challenges of single 3D human pose estimation from different inputs, multi-person 3D human pose estimation and in-the-wild datasets.

(1) Different Inputs. Generally speaking, based on different considerations, various types of inputs are used to estimate 3D pose and thus the corresponding challenges are varied as well. Visual cues, such as shadows and objects of known size, can be used to address ambiguities in images. However, it is very difficult to directly capture such information from images. When ignored, using 2D joints to recover a 3D pose becomes an ill-defined problem. For instance, as shown in Fig. 1, one 2D skeleton may correspond to many varied 3D poses. Actually, the depth ambiguity could be considerably reduced by using temporal information, multi-view images, etc. First, for recovering 3D human pose from a sequence of images, temporal information could be exploited to reduce the depth ambiguity. At the same time, there are many additional challenges such as background variation, camera movement, fast motion, changes of clothing, illumination changes, which may cause the shape and appearance of people that alter dramatically over time. Second, when utilizing multi-view images, researchers face the problem how to fuse information from multiple cameras. In fact, due to the occlusion and inaccuracy estimation of 2D poses, this is not a trivial problem that could be simply solved by triangularization from estimated 2D poses, especially when there are few cameras in practical scenes.

(2) Multiple Persons. Compared with single human pose estimation, estimating 3D poses of multiple persons is more challenging. When estimating multi-person from a monocular image, the additional challenge is the occlusion caused by nearby individuals. When estimating 3D poses of multiple persons from multiple views, the main challenges include the larger state space, occlusions and cross-view ambiguities, as shown in Fig. 2. Besides, most existing methods are based on two-stage frameworks which suffer from problems in efficiency, while single-stage methods (Nie et al., 2019) have been proposed to solve this problem, they are far from mature.

(3) In-the-Wild Scenario. In addition, the lack of in-the-wild datasets is a bottleneck for research on 3D pose estimation. For 2D human pose estimation, it is feasible to construct large in-the-wild datasets (Andriluka et al., 2014a; Lin et al., 2014a) by manually labeling the 2D poses of humans in the image. However, since 3D annotations are generally acquired by marker-based vision systems, collecting a large-scale in-the-wild dataset with 3D annotations is very resource-intensive.

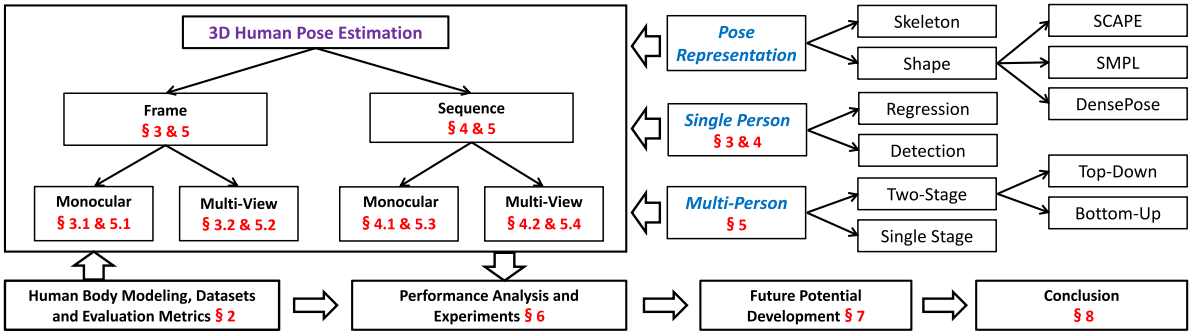

Fig. 3. Framework of this review.

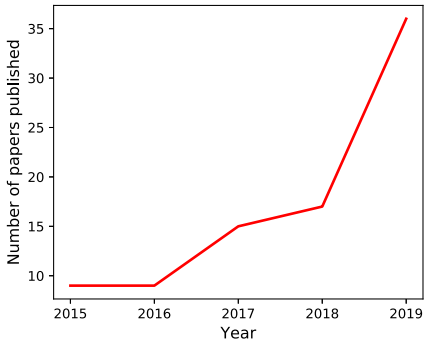

Fig. 4. The numbers of 3D human pose estimation papers published in top conferences (CVPR, ICCV and ECCV).

As is well-known, the most popularly used datasets such as HumanEva and Human3.6M, are captured by motion capture systems under an indoor environment. Thus, the algorithms trained on such datasets inevitably confront a generalization challenge when they are used for in-the-wild applications. To mitigate the problem, many methods have explored, such as lifting 2D pose to 3D pose (Tome et al., 2017), transferring knowledge (Zhou et al., 2017), utilizing weak supervision signal (Chen et al., 2019a) and synthesizing in-the-wild images (Varol et al., 2017). However, the in-the-wild performance of these methods are still unsatisfactory compared with 2D pose estimation.

1.3. Scope of this survey

Previous surveys generally focus on traditional methods, such as pictorial models and exemplar-based approaches. Readers are encouraged to read these review articles, in which more details have been provided. A recent survey (Sarafianos et al., 2016) mainly focuses on the review of work from 2008 to 2015. In that survey, the authors proposed a rather complete taxonomy for 3D pose estimation and introduced a new synthetic dataset as well. However, they mainly summarized classical methods and only a few deep learning based methods were mentioned. Furthermore, the rapid progress of deep learning in recent years has greatly promoted the development of 3D human pose estimation. While recent surveys do not cover these methods comprehensively or give a summary from a specific perspective. For example, Chen et al. (2020) merely provide a review of deep learning-based methods for monocular human pose estimation.

Therefore, we follow the same reasonable taxonomy but instead focus on deep learning based methods to reveal the current research state of this field. Moreover, we observe that, in recent years, 3D human pose estimation has gained increasing attention in the area of computer vision community according to the numbers of published papers in top computer vision conferences (CVPR,[2] ICCV,[3] and ECCV[4]), as shown in Fig. 4 In addition, the representation of the 3D pose and datasets are very important for human pose estimation. According to the types of models, we classify the representations of poses to skeleton and shape based approaches, as shown in Fig. 3. In recent years, many new datasets have been proposed. We will discuss human pose modeling and datasets in Section 2.

In summary, the framework of our review is shown in Fig. 3. We cover deep learning based algorithms for estimating 3D human pose, where the inputs ranging from a single image to a sequence of images, from a single view to multiple views, and from a single person to multiple persons. From the perspective of pose representation, the input data can be divided into two types: skeleton and shape (contour). Also, many parametric models are used to supplement the body shape, such as SCAPE (Anguelov et al., 2005), SMPL (Loper et al., 2015), and DensePose (Alp Güler et al., 2018). As for 3D pose estimation of multiple people, the approaches can be classified into single-stage methods and two-stage methods. The two-stage methods can be further divided into top-down and bottom-up methods as shown in Fig. 3. Specifically, the top-down methods detect each person first and then locate their joints individually, whilst the bottom-up methods locate all the body joints first and then assign them to the corresponding person. In contrast, the one-stage methods (Nie et al., 2019) normally estimate the locations of root position and joint displacements, simultaneously.

2. Human body modeling, datasets and evaluation metrics

2.1. Human body modeling

Generally, the human body structure is very complex, and different methods adopt different models based on their specific considerations. Nevertheless, the most commonly used models are the skeleton and shape models. Besides, a new pose estimation is a surface-based representation called DensePose (Alp Güler et al., 2018), which is worth mentioning due to the extension of the existing pose representation. Next, we will introduce them in detail.

Skeleton-Based Model: First and foremost, the skeleton model is commonly used in 2D human pose estimation (Cao et al., 2018) and is naturally extended to 3D. The human skeleton model is treated as a tree structure, which contains many keypoints of the human body and connects natural adjacent joints using edges between key joints, as shown in Fig. 5.

SMPL-Based Model: For the shape model, recent works use the skinned multi-person linear (SMPL) model (Loper et al., 2015), as shown in Fig. 6, to estimate 3D human body joints (Bogo et al., 2016). The human skin is represented as a triangulated mesh with 6890 vertices, which is parameterized by shape and pose parameters. The

[2] IEEE conference on Computer Vision and Pattern Recognition
[3] IEEE International Conference on Computer Vision
[4] European Conference on Computer Vision

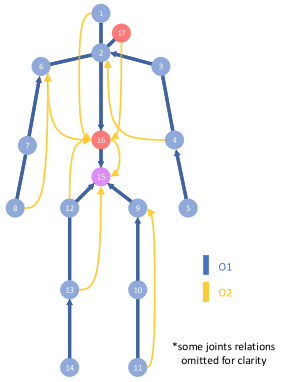

Fig. 5. Human body skeleton from the MPI-INF-3DHP dataset, with the root joint 15, O1 (blue): relative to first order and O2 (orange): relative to second order parents in the kinematic skeleton hierarchy. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 6. The SMPL model (Loper et al., 2015). The white points are pre-defined keypoints.

shape parameters are used to model the body proportions, height and weight, while the pose parameters are used to model the determined deformation of the body. The 3D pose positions can be estimated by learning the shape and body parameters.

Surface-Based Model: Recent, a new model of the human body: DensePose (Alp Güler et al., 2018) is recently proposed, considering the fact that sparse correspondence of the image and keypoints is not enough to capture the status of the human body. To address the issue, a new dataset named DensePose-COCO is constructed, which establishes the dense correspondences between image pixels and a surface-based representation of the human body. This work further promotes the development of human understanding in images and can be understood as the next step in the line of works on extending the standard for humans in 2D or 3D human estimation datasets, such as MPII Human Pose (Andriluka et al., 2014b), Microsoft COCO (Lin et al., 2014b), HumanEva (Sigal et al., 2009), Human3.6M (Ionescu et al., 2013).

2.2. Datasets

The 3D pose estimation datasets are often gathered by a motion capture system. A previous review has analyzed the datasets from 2009 to 2015 (Sarafianos et al., 2016). The HumanEva dataset (Sigal et al., 2009) and Human3.6M dataset (Ionescu et al., 2013) are still the standard for 3D human pose estimation. Moreover, since there have been many new datasets proposed recently, we will introduce these dataset in detail in the following sections and sum up the main points in Table 1.

HumanEva-I Sigal et al. (2010) contains 7 calibrated video sequences (4 grayscale and 3 color) that are synchronized with 3D body poses obtained from a motion capture system. The database contains 4 subjects performing a 6 common actions, e.g. walking, jogging, gesturing. The dataset contains training, validation and testing sets.

Human3.6M Ionescu et al. (2013) is one of the largest motion capture datasets, which consists of 3.6 million human poses and corresponding images. The dataset provides accurate 3D human joint positions and synchronized high-resolution videos acquired by a motion capture system at 50 Hz. The dataset contains activities by 11 professional actors in 17 scenarios: discussion, smoking, taking photo, talking on the phone, etc., from 4 different camera views.

MARCOnI (MARker-Less Motion Capture in Outdoor and Indoor Scenes, Elhayek et al. (2016) is a comprehensive dataset that can be used for versatile testing. The dataset is composed of 12 sequences with different conditions, such as sensor modalities, numbers and types of cameras, identities of actors, scene and motion complexities. All cameras are synchronized, even the cell phone and the GoPro cameras. This dataset provides 3D joint positions calculated by three reference methods as follows. (1) MP: some sequences are recorded by a synchronized Phasespace active-LED marker-based motion capture system and the 3D joint locations could be captured by markers. (2) A3D: the 2D poses of other sequences are annotated manually to calculate ground truth 3D joint locations. (3) DMC: for sequences with enough cameras, the dataset also provides 3D joint positions using a baseline approach (Stoll et al., 2011).

MPI-INF-3DHP Mehta et al. (2017a) uses a commercial marker-less motion capture system to collect data, which does not require special suits or markers, and thus actors could wear everyday clothes including loose clothes. There are 8 actors (4 females + 4 males), each performing 8 action sets, each of which lasts about 1 min. The test set consists of 2929 valid frames from 6 subjects performing 7 actions. The actions range from walking, sitting, and complex exercise actions to dynamic actions. The number of action classes is more than that of Human3.6M dataset. To increase the diversity of data, each actor performs activities of both daily apparel and plain-colored clothing sets. Moreover, the dataset increases the scope of foreground and background augmentation by providing chroma-key masks for the background.

Total Capture Trumble et al. (2017) is the first dataset that provides both multi-viewpoint video (MVV), inertial measurement unit (IMU), and skeleton annotations obtained by a commercial motion capture system (Vicon). The dataset does not use any markers, so actors could wear very loose clothes to increase the variation of appearance. The XSens IMUS system (Roetenberg et al., 2009) uses 13 IMU sensors on key body parts including head, upper/lower back, upper/lower limbs, and feet. The dataset provides accurate background subtraction for each pixel. It contains five kinds of actions, each of which is repeated three times by actors. Finally, the dataset is split into several subsets according to the subjects and action sequences, allowing for testing both unseen subjects and seen subjects with unseen actions.

SURREAL (Synthetic hUmans foR REAL tasks, Varol et al. (2017)) is a large-scale synthetic dataset with randomly generated 3D poses, shapes, textures, illustrations and backgrounds. The shape information of the dataset was from the CMU motion capture (MoCap) dataset. Next, the MoSh (Loper et al., 2014) method is explored to fit the SMPL parameters using the raw data of the 3D MoCap markers. Then, given the fitted parameters, the synthetic body is generated by the SMPL model, and the real appearance image is mapped into the body shape. Further, the texture information is obtained from 3D scans of the subjects wearing normal clothing, largely increasing the authenticity of the synthetic data. The background images are from a subset of the LSUN dataset (Song and Xiao, 2015), which includes a total of 400 K images from the kitchen, living room, bedroom, and dining room. The illumination variation uses the model of Spherical Harmonics with 9 coefficients (Green, 2003). The SURREAL dataset is the first to provide 3D pose annotation, part segmentation, and flow ground truth, which can be used for multi-task training. The authors also generate

Table 1
3D human pose datasets.

Year	Dataset	No. of images	No. of subjects	Characteristics
2010	HumanEva-I	12 sequences	4 subjects, 6 actions	Indoor multi-view video, markerless motion capture
2013	Human3.6M	3.6M	11 (5 female + 6 male)	One of the largest motion capture dataset; Multiple views
2014	Shelf		4	Indoor multi-view video; multiple persons; each view suffers from heavy occlusion
2014	Campus		3	Outdoor multi-view video; multiple persons
2016	MARCOnI	12 sequences	1 or 2 per sequence	Comprehensive dataset for versatile testing
2016	CMU Panoptic	1.5M	Up to 8 subjects	Captured in a studio with hundreds of cameras; large social interaction
2016	MPI-INF-3DHP	1.3M frames	8 (4 female + 4 male)	Indoor multi-view video, markerless motion capture, data augmentation
2017	MuCo-3DHP MuPoTS-3D		8 (4 female + 4 male)	Build upon segmentation masks in MPI-INF-3DHP
2017	Total Capture	1.892M frames	5 (4 male + 1 female)	Indoor multi-view video, IMU, and vicon mocap
2017	SURREAL	6M frames	145	Rendered from 3D sequences of motion capture data (Human3.6M)
2017	Unite the people	8515 images		Improve SMPLify to semi-automated annotate dataset, annotate 31 segments on the body and 91 landmarks
2018	JTA	500K frames	> 21	Massive simulated dataset, 500K frames with almost 10 million pose
2018	3DPW	60 sequences	7	The only promising 3D pose in the wild dataset; 24 train, 24 test, 12 validation

the predicted body part segmentation and depth maps for samples in the Human 3.6M dataset. Finally, the dataset is divided according to subjects: 115 subjects are used as the training sets and 30 of them are used as the test set.

Unite the People Lassner et al. (2017) contains 5569 training images and 1208 test images. This dataset is collected based on the observations that the CNNs are often applied in isolated and separated datasets, such as MPII (Andriluka et al., 2014a), LSP (Johnson and Everingham, 2010), and are independent of 3D body estimation. To *unite the people* of multiple human datasets, the authors improve the SMPLify method to obtain high-quality 3D human body models, and then manually sort these body models based on the quality. This semi-automated approach makes annotations more efficient and enables consistent labeling by reprojecting the body model to the original image. The denser set of annotations that predict 31 segments on the body and 91 landmark positions enable eliminating the ambiguity of poses and shapes in a single view. Furthermore, a regression tree model is proposed to predict poses or shapes, which is one to two orders of magnitude faster than SMPLify. Finally, experiments show that using 91 landmarks the pose estimators can be trained with fewer data without requiring gender or pose assumptions.

JTA (Joint Track Auto, Fabbri et al. (2018)) is a synthetic people tracking dataset in urban scenarios with ground-truth annotations of 3D poses, of which 256 videos are used for training and 256 videos are used for testing. These collected videos with varying illumination conditions and viewpoints are from the highly photorealistic video games *Grand Theft Auto V* developed by *Rockstar North*. The distance from the camera varies from 0.1 to 100 m, resulting in heights of subjects varying from 20 to 1100 pixels. By accessing the game rendering module, 14 body parts are automatically annotated in Andriluka et al. (2014a, 2018). Besides that, some simulated challenges including *occlusion* and *self-occlusion* are provided as well. *Occlusion* denotes that the joint is occluded by objects or other pedestrians, while *self-occlusion* denotes that the joint is occluded by the owner of the joint. Besides, each person is assigned an identifier so that the dataset can also be used for person re-identification research.

3DPW (3D Poses in the Wild, von Marcard et al. (2018)) is the first dataset in the wild with accurate 3D poses for evaluation. It is created by utilizing information from IMUs and a hand-held phone camera. A 3D pose estimation method named video inertial poser (VIP) is used to integrate the images and IMU readings of all frames in video sequences. The VIP has been validated on the Total Capture dataset, which has an accuracy of 26 mm and is accurate enough to create the dataset for image-based 3D pose estimation. For tracking single subjects, 17 IMUs would be used, while 9–10 IMUs would be used to simultaneously track up to 2 subjects. Then, the video and IMUs data are synchronized by a clapping motion as in Pons-Moll et al. (2011). In total, the dataset contains up to 18 clothing styles and actions such as walking in cities, going up-stairs, having coffee, or taking the bus. Compared with Total Capture, there are more subjects in a scene.

Shelf and Campus (Belagiannis et al., 2014) The shelf dataset has annotated the body joints of four actors interacting with each other using cameras 2, 3, and 4. Triangulation is performed using the three camera views for deriving the 3D ground-truth. The actor 4 (Vasilis) is occluded in most of the camera views and thus excluded from the evaluation. The Campus dataset has annotated the body joints of the main three actors performing different actions for the frames that are observed from the first two cameras. The ground-truth for the third camera view is the result of the triangulation (between cameras 1 and 2), and then projected to camera 3.

CMU Panoptic Joo et al. (2017) provides some examples with large social interaction. It used 480 synchronized VGA cameras, 31 synchronized HD cameras (temporally aligned with VGA cameras), and 10 RGB-D sensors for motion capture. All of the 521 cameras are calibrated by structure from the motion approach.

MuCo-3DHP (Multiperson Composited 3D Human Pose) is created by leveraging segmentation masks provided in MPI-INF-3DHP dataset (Mehta et al., 2017a). To collect this dataset, per-camera composites with 1 to 4 subjects are first generated in the images randomly selected from the MPI-INF-3DHP dataset, in which each camera has 16 sequences. The composited dataset covers many kinds of inter-person overlaps and activities. Using a commercial multi-view marker-less motion capture system, a new filmed multi-person test set named **MuPoTS-3D** (Multiperson Pose Test Set in 3D) is collected as well. In total, this dataset comprises 20 general real-world scenes (5 indoor and 15 outdoor) for up to three subjects with challenging elements such as drastic illuminations and lens flares for outdoor settings.

In summary, for indoor 3D human pose estimation datasets, the Human3.6m dataset is the most common one used in recent years, although the HumanEva dataset is still frequently employed. Besides, the MPI-INF-3DHP is also widely used, since it has more action classes than Human3.6m and provides chroma-key masks for foreground and background augmentation. As for the other three indoor datasets, the CMU Panoptic dataset is created for large social interaction capture; the MARCOnI dataset can be used for versatile testing since it contains sequences with different conditions; the Total Capture dataset provides MVV, IMU, and Vicon annotations in constrained environments. However, these three datasets are less used than the first two. To evaluate the generalization ability of 3D human pose estimation algorithms, several in-the-wild datasets have been proposed including SURREAL, JTA, Unite the People, MuCo-3DHP, and 3DPW. The first two are seldom used recently while the third is widely used by SMPL based 3D pose estimation methods. The fourth dataset can generally be used for multi-person pose estimation. To some extent, the last dataset is a promising in-the-wild dataset, since the annotations with high accuracy of 26 mm are obtained from the Total Capture dataset.

2.3. Evaluation metrics

We list some of the most frequently used metrics below for reference and detailed settings based on datasets.

MPJPE (Mean Per Joint Position Error): This metric is calculated by

$$E_{MPJPE}(f,S)=\frac{1}{N_S}\sum_{i=1}^{N_S}\|P_{\mathbf{f},S}^{(f)}(i)-P_{\mathbf{gt},S}^{(f)}(i)\|_2, \tag{1}$$

where f denotes a frame and S denotes the corresponding skeleton. $P_{\mathbf{f},S}^{(f)}(i)$ is the estimated position of joint i and $P_{\mathbf{gt},S}^{(f)}(i)$ is the corresponding ground truth position. All joints are considered, $N_S = 17$. Finally, the MPJPEs are averaged over all frames. Besides, we refer to the resulting normalized metrics as **NMPJPE**. Since orientation is left unchanged, this is a less constraining transformation than the more commonly used procrustes alignment, to which we refer as **PA-MPJPE**.

PCP (Percentage of Correctly estimated Parts): The PCP metric measures the percentage of correctly predicted parts (Ferrari et al., 2008). As mentioned in Sarafianos et al. (2016), a body part is considered correct by the algorithm if:

$$\frac{\|s_n-\hat{s}_n\|+\|e_n-\hat{e}_n\|}{2}\le\alpha\|s_n-e_n\|, \tag{2}$$

where s_n and e_n are the ground truth start and end location of part n, $\hat{s}_n$ and $\hat{e}_n$ are the corresponding estimated locations, and α is a threshold parameter.

PCK (Percentage of Correct Keypoints): It is first used in 2D pose estimation (Yang and Ramanan, 2012). Mehta et al. (2017a) extend PCK to the 3D space and calculate the area under the curve (**AUC**) when varying the PCK threshold. A estimated joint is considered correct if its distance to the corresponding ground truth is less than a threshold (e.g., 150 mm). This metric is often used in the new MPI-INF-3DHP dataset. The normalized version of PCK (**NPCK**) is used in Rhodin et al. (2018b), Kocabas et al. (2019).

Bone Error, Bone Std, Illegal Angle: Sun et al. (2017) propose corresponding metrics for their bone representation of the human body because they argue that absolute joint location based metrics such as MPJPE and PCK do not consider the pose's internal structures. The mean per bone position error (**Bone Error**) measures the relative joint location accuracy. The bone length standard deviation (**Bone Std**) measures the stability of bone length by computing the standard deviation over a subject's all testing samples. The percentage of illegal joint angle (**Illegal Angle**) measures the feasibility of a joint's rotation angles (Akhter and Black, 2015).

MRPE (Mean of the Root Position Error): Moon et al. (2019) propose this metric to evaluate the accuracy of the absolute location of an estimated 3D human root:

$$MRPE=\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{R}^{(i)}-\mathbf{R}^{(i)*}\right\|_2, \tag{3}$$

where $\mathbf{R}$ and $\mathbf{R}^{(i)*}$ are the ground truth and estimated locations of the ith sample respectively, and N is the number of testing samples.

HumanEva-I: Sigal et al. (2010) use 3D error (**3D Error**) metric to evaluate performance on their HumanEva dataset. The 3D error is the mean squared distance between coordinates of estimated and ground truth pose.

Human3.6M: There are three main protocols for evaluating the performance of 3D human pose estimation algorithms in terms with MPJPE.

P1 (protocol #1, the standard protocol) uses 5 subjects (S1, S5, S6, S7, S8) for training and 2 subjects (S9, S11) for testing.

P2 (protocol #2) differs from Protocol #1 in that it uses S11 for testing while using 6 subjects (S1, S5, S6, S7, S8 and S9) for training. The pose error is calculated after a similarity transformation (Procrustes analysis) between the estimated pose and ground truth. The original video is down-sampled to every 64th frame and evaluation is performed on sequences from all 4 cameras and all trials. The error is averaged over 14 joints.

P3 (protocol #3) splits the dataset in the same way as protocol #1 (Bogo et al., 2016). However, the evaluation is only conducted on sequences captured by the frontal camera ("cam 3") in trial 1 and the original video is not sub-sampled. The error is averaged over a subset of 14 joints.

3. 3D human pose estimation based on a frame

This section will detailedly introduce 3D human pose estimation methods which do not use temporal information, that is, only uses a monocular image or multi-view images at a single time. Thanks to its great advantages, e.g. suitable for indoor and outdoor use, it has been widely studied recently.

3.1. 3D human pose estimation from a monocular image

Recovering a 3D human pose from a single image is appealing due to the low requirement of the image, but it suffers from an ill-defined problem that different 3D poses may correspond to the same 2D images. Besides, based on the setting, using temporal or multi-view information to reduce the ambiguity cannot be achieved during the recovering process. Therefore, significant research has been done and several methods have been developed to solve these problems. In this section, we will first introduce the methods and then illustrate some representative works. Specifically, we will review methods from three parts, namely directly predicting 3D poses from images, lifting from 2D poses, and SMPL-based methods.

3.1.1. Direct 3D pose estimation

The most straightforward way to estimate 3D human poses is to design an end-to-end network to predict the 3D coordinates of joints for the poses. Methods that directly map input images to 3D body joint positions can be categorized into two classes: detection-based methods (Pavlakos et al., 2017a; Luvizon et al., 2018) and regression-based methods (Li and Chan, 2014; Zhou et al., 2016a; Sun et al., 2017; Tekin et al., 2017; Zhou et al., 2017; Luvizon et al., 2019). It is worth noting that attempts have also been made to unify the heatmap representation and joint regression (Sun et al., 2018). All of these methods are summarized in Table 2–1).

Detection-based methods predict a likelihood heatmap for each joint, and the joint's location is determined by taking the maximum likelihood of the heatmap. Pavlakos et al. (2017a) use a volume to represent a 3D pose and then train a CNN to predict the voxel-wise likelihood for each joint in the volume, which greatly improves the direct regression of joint coordinates. They adopt a coarse-to-fine prediction scheme, which employs intermediate supervision and an iterative estimation module to gradually increase the resolution of the supervision volume. Luvizon et al. (2018) propose a multi-task framework to jointly estimate 2D/3D poses and recognize actions, where 2D and 3D pose estimation is unified using volumetric heatmaps. However, such methods rely on additional steps to convert heatmaps to joint positions, usually by applying the argmax function, which is not differentiable. This interfaces with the learning mechanism of neural networks. Besides, the precision of predicted keypoints is proportional to that of the heat map resolution, which lacks inherent spatial generalization. To achieve high precision, the predicted heatmaps usually require a reasonable spatial resolution, which quadratically increases the computational cost and memory consumption.

Human pose estimation is essentially a regression problem that directly estimates the locations of joints relative to the root joint location. Li and Chan (2014) design a simple but effective neural network with two branches that simultaneously detect the root location and regress the relative locations of other joints. To incorporate prior knowledge of the geometric structure of the human body, Zhou et al. (2016a) introduce a kinematic object model consisting of several joints and bones,

Table 2
Estimating 3D human pose from a single monocular image.

(1) Direct 3D pose estimation	Highlight	Dataset	Metric	Code
Li and Chan (2014)	Network with two branches, one detects the root location and one regresses the relative location of other joints	Human3.6M	MPJPE	No
Zhou et al. (2016a)	Geometric constraints and direct regression	Human3.6M	MPJPE	No
Sun et al. (2017)	Regress bones	Human3.6M	P2, P3	Compositional
Luvizon et al. (2018)	Jointly 2D/3D pose estimation and action recognition; use Soft-argmax on heatmaps	Human3.6M	MPJPE	Deephar
Pavlakos et al. (2017a)	Coarse-to-fine prediction scheme based on intermediate supervision	Human3.6M; HumanEva-I; KTH Football II	MPJPE; reconstruction error; PCP	C2f
Sun et al. (2018)	Modify the taking-maximum operation to taking-expectation	Human3.6M	P2, P3	Integral
(2) Lifting 2D pose to 3D pose	**Highlights**	**Dataset**	**Metric**	**Code**
Park et al. (2016)	Concatenate 2D pose estimation results and image features, estimate the 3D position relative to multiple root joints	Human3.6M	P1	No
Tome et al. (2017)	Multi-stage; jointly 2D/3D pose estimation; lifting by a unimodal Gaussian 3D pose model	Human3.6M	P1, P2, P3	Lifting
Moreno-Noguer (2017)	2D to 3D distance matrix regression	Human3.6M; HumanEva-I	P1, P2, P3; PCP	No
Fish Tung et al. (2017)	Utilize feedback from 3D re-projection and use a discriminator to judge feasibility of the generated 3D pose	Human3.6M	P1	No
Martinez et al. (2017)	Lift 2D pose to 3D by a simple neural network	Human3.6M; HumanEva	P1, P3; 3D Error	Baseline
Nie et al. (2017)	Exploit body part images to predict the depth to reduce lifting ambiguity; skeleton-LSTM utilizes the global 2D pose features; patch-LSTM utilizes the body part images	Human3.6M; HHOI	P2, P1; MPJPE	No
Tekin et al. (2017)	Fuse 3D image cues with 2D joint heatmaps in a trainable scheme	Human3.6m; HumanEva-I; KTH Football II	P1, P3; 3D Error; PCP	Fuse
Yang et al. (2018)	Multi-source discriminator; train on images with 2D only	Human3.6M;	P1, P3	No
(Pavlakos et al., 2018a)	Ordinal depth as supervision	Human3.6M; HumanEva-I; MPI-INF-3DHP	P1, P3; 3D Error; AUC, 3DPCK	Ordinal
Alp Güler et al. (2018)	Estimate the pixel-wise 3D surface correspondence of the human body	DensePose-COCO	AP	DensePose
Zhou et al. (2019)	The part-centric heatmap triplet (negative, zero and positive polarity heatmaps)	Human3.6M; HumanEva-I; MPI-INF-3DHP	P1, P2; 3D Error; 3DPCK, AUC	No
Wang et al. (2019a)	Predict 3D poses from low-DOF to high-DOF	Human3.6M; MPI-INF-3DHP	P1, P3; 3DPCK, AUC	No
Ci et al. (2019)	Use different filters for feature extraction	Human3.6M	P1, P3	LCN
Sharma et al. (2019)	Deep conditional variational autoencoder generates 3D pose samples to reduce lifting ambiguity	Human3.6M; HumanEva-I	P1, P2; PCP	Generative
Li et al. (2019b)	Generate multiple corresponding feasible 3D pose solutions for 2D joint points	Human3.6M; MPI-INF-3DHP	P1; 3DPCK	Multi-hypo
Chen et al. (2019a)	Jointly understand holistic scene and estimate 3D human pose (holistic++ scene understanding)	PiGraphs; SUN RGB-D; WnP	Average Euclidean distance	Coming
Zhao et al. (2019)	Lifting by semantic graph convolutional network	Human3.6M	P1, P2	SemGCN
Habibie et al. (2019)	Explicitly represent the 2D pose with heatmap and implicitly represent the depth information	MPI-INF-3DHP; Human3.6M	3DPCK, MPJPE, AUC; P1	No
Chen et al. (2019b)	Unsupervised, lift 2D joint to 3D pose, generate 2D pose after rotation (this stage has discriminator to judge whether the image is realistic), then lift the generated 2D to 3D again	Human3.6M; MPI-INF-3DHP	MPJPE; 3DPCK, AUC	No
Wandt and Rosenhahn (2019)	Use a GAN to discriminate whether the 3D pose generated by the network is realistic. Estimate the camera parameters	Human3.6M; MPI-INF-3DHP	P1, P3; MJPE, AUC, 3DPCK	RepNet
Jack et al. (2019)	Learn a consistency measure between 2D observations and a proposed world model by a neural network	Human3.6M;	P1, P3	Ige-net
(3) SMPL model based estimation	**Highlights**	**Dataset**	**Metric**	**Code**
Bogo et al. (2016)	Matching the 2D keypoints projected from the SMPL model with detected 2D keypoint; optimization based	HumanEva-I; Human3.6M	Average Euclidean distance (PA); P3	SMPLify
Lassner et al. (2017)	Improved SMPLify by additionally matching the image silhouette and the silhouette projected from the SMPL model	HumanEva; Human3.6M; UP-3D	Average error over all joints	UP
Tan et al. (2018)	Encoder–decoder (decoder takes the SMPL parameters as input and output corresponding silhouette)	UP-3D		No
Kanazawa et al. (2018)	End-to-end deep learning scheme mapping from image to SMPL model parameters; use a discriminator	Human3.6M; MPI-INF-3DHP	P1, P2; P1, PCK, AUC	HMR
Omran et al. (2018)	Use a semantic segmentation CNN to the image into 12 semantic parts, then encode the semantic part probability maps into SMPL parameters	HumanEva-I; Human3.6M	3D Error; P3	NBF
Pavlakos et al. (2018b)	Predict 2D heatmaps and masks first, two networks predict pose and shape parameters individually; 3D per-vertex loss	UP-3D; SURREAL; Human3.6M	Mean per vertex errors; mean per vertex errors; P3	No

where the bones have a fixed length and can be rotated around the combined joint. However, the fixed length of bones does not reflect the variability of the human skeleton, limiting the model's generalization ability. Sun et al. (2017) believe that it is more reasonable to regress bones rather than joints for pose estimation, because bone representations are easier to learn and better to reflect geometric constraints, as well as being more stable. Furthermore, to overcome the issue of the $L2$ loss of bones being local and independent, a compositional loss function is employed in their work, which encodes long range interactions between the bones. However, this method requires pose data to be converted to the relative bone-based format. More recently, Luvizon et al. (2019) propose the soft-argmax function to convert feature maps to joint coordinates, resulting in a fully differentiable framework. Similar to the soft-argmax operation, Nibali et al. (2018) introduce a new layer, called differentiable spatial to numerical transform (DSNT), to preserve the end-to-end differentiability and the spatial generalization of the model.

In summary, heatmap representations suffer from a few drawbacks in practice. The "taking-maximum" operation is not differentiable and prevents training from being end-to-end. In contrast, regression approaches achieve end-to-end learning and produce continuous outputs by replacing "taking-maximum" with "taking-expectation". However, they are not as effective as detection-based methods. To incorporate the merits of both, Sun et al. (2018) propose a simple and effective integral regression approach in which the joint position is estimated as the probability-weighted average of all positions in the heatmap. This method allows end-to-end training and requires low computation and storage.

3.1.2. Lifting from 2D to 3D pose

Inspired by the rapid development of 2D human pose estimation algorithms, many works have tried to utilize 2D pose estimation results for 3D human pose estimation to improve in-the-wild generalization performance. For example, Martinez et al. (2017) propose a simple baseline focusing on lifting 2D poses to 3D with a simple yet very effective neural network, which popularizes the research on lifting 2D pose to 3D space. Other methods focus on how to fuse 2D joint heatmaps with 3D image cues to reduce the ambiguity (Park et al., 2016; Tekin et al., 2017; Habibie et al., 2019; Zhou et al., 2019). The relationships between joints have been exploited by long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997; Nie et al., 2017; Wang et al., 2019a) and Euclidean distance matrix (EDM) (Moreno-Noguer, 2017), and graph neural networks (Zhao et al., 2019; Ci et al., 2019). The reprojection of the generated 2D pose is often used as supervision (Tome et al., 2017; Habibie et al., 2019; Jack et al., 2019; Chen et al., 2019a). To produce a more realistic 3D human pose, generative adversarial networks (GANs) are often used (Fish Tung et al., 2017; Yang et al., 2018; Wandt and Rosenhahn, 2019). Since the full 3D in-the-wild annotation is too difficult, some authors have considered providing more weak supervision of depth information such as (Pavlakos et al., 2018a; Alp Güler et al., 2018). These methods are summarized in Table 2–2.

We firstly review the methods that focus on fusing 2D pose results with image features containing depth cues to reduce the ambiguity. In order to help the model to discard unnatural human 3D joint positions, Park et al. (2016) come up with an new solution which is concatenating 2D pose estimation results and image features. To avoid manually designing the fusing, Tekin et al. (2017) use a fusion stream to utilize 3D image cues and 2D joint heatmaps in a trainable scheme. There are also some methods that focus on feature representation. For example, Habibie et al. (2019) explicitly represent the 2D pose with heatmap and implicitly represent the depth information. Both 2D pose heatmap and depth feature are used to estimate 3D pose and viewpoint parameters, and the generated 3D pose is projected to the 2D plane by viewpoint parameters. It has been confirmed that part-centric heatmap triplets (HEMlets) as an intermediate representation can divide the full-body skeleton into 14 parts and model the local order information of these subparts (Zhou et al., 2019). The part-centric heatmap triplet consists of three heatmaps named negative polarity heatmap, zero polarity heatmap, and positive polarity heatmap, respectively. In this way, the relative depth information can bridge the gap between 2D poses and 3D poses.

The geometric relationships between human pose joints can be utilized in designing algorithms. Early, Nie et al. (2017) process 2D poses and body part images using LSTMs, which exploit the tree structure of the human skeleton to propagate contextual information similar to Tai et al. (2015). Besides, Moreno-Noguer (2017) argue that the ambiguity can be reduced by calculating pairwise distances of body joints formulated by Euclidean distance matrix (EDM), which can encode structural information and capture pairwise correlations and dependencies. Another work directly predicts 3D pose joints from low degree of freedom (DOF) to high DOF (Wang et al., 2019a). With the development of graph neural networks (GNNs), there are many attempts in recent works. Noted that the human pose skeleton naturally forms a graph, Zhao et al. (2019) propose the semantic graph convolution network (GCN), which learns the weight of the edges in the skeleton graph channel-by-channel, to extract a 3D pose from the 2D joint points. To overcome the representation limits of GCN, the locally connected network (LCN) is proposed (Ci et al., 2019), which uses different filters for feature extraction rather than a shared filter as in GCN. Importantly, they also note the limitations of the natural skeleton adjacency matrix and make it learnable, similar to Zhao et al. (2019).

An estimated 3D pose is often reprojected to the 2D space to ensure consistency. As a common fashion, a multi-stage approach is used to reason jointly about the 2D and 3D pose, such as (Tome et al., 2017), where the generated 3D pose is projected to 2D to produce 2D belief maps, which are fused with belief maps produced by an off-the-shelf 2D estimator. As mentioned before, Habibie et al. (2019) project the generated 3D pose to the 2D plane using estimated viewpoint parameters. Another way is to learn the loss function rather than manually design it, like (Jack et al., 2019). The energy function (loss) is the sum of reprojection energy and feasibility energy. The reprojection energy measures the consistency between a proposed 3D pose and the ground truth by a two-layer dense network, while the feasibility energy measures how much feasible the proposed pose is in the real world. In an unsupervised manner, Chen et al. (2019a) utilize the rotation invariant of the generated 3D pose. The 2D pose is first lifted to a 3D pose, which generates a 2D pose after rotation. Then, the generated 2D pose is lifted to 3D again. The loss between the two 3D poses is calculated, and then the 3D pose is restored to the 2D pose to calculate the loss between the restored 2D pose and the input 2D pose. Similarly, Novotny et al. (2019) propose to learn a canonicalization network that maps equivalent class members to the canonical reconstruction to regularize the results. They reconstruct the 3D shape by a factorization network that can factorize viewpoint information and object deformations. The reconstruction branch is trained by minimizing the re-projection error.

To generate realistic 3D poses, adversarial learning has been used. For example, Fish Tung et al. (2017) propose adversarial inverse graphics networks (AIGNs) composed of a generator, a renderer, and a discriminator. The generator first predicts 2D pose heatmaps and then predicts the camera and shape parameters from heatmaps. The renderer is simply the reprojection function. The discriminator is used to judge the feasibility of the generated 3D pose. Subsequently, a multi-source discriminator is trained in Yang et al. (2018), which takes the original image, the geometric descriptor on image-pose correspondence, the 2D pose, and depth heatmaps as input. Recently, it has achieved good performance. Wandt and Rosenhahn (2019) adopt a GAN to discriminate whether a 3D pose generated by the network is realistic rather than the 2D pose. In addition, they also propose a network to estimate the camera parameters, and which they use to project the generated 3D pose to the 2D space, as well as calculate the loss with the original 2D pose.

The 3D human pose estimation from a single monocular image suffers from inherent ambiguity, and thus more supervision signals are required. Pavlakos et al. (2018a) propose to use the ordinal depth (closer–farther relationship) annotation as weak supervision, which seems to be a promising substitute for full in-the-wild 3D annotation. They augment the 2D pose datasets MPII (Andriluka et al., 2014a) and LSP (Johnson and Everingham, 2010) with ordinal annotations, which can be used in a variety of settings, including the volumetric representation. Following the work of Alp Guler et al. (2017), DensePose (Alp Güler et al., 2018) is proposed and its dataset named DensePose-COCO is constructed, where it estimates the pixel-wise 3D surface correspondence of the human body in an RGB image and the COCO dataset with dense correspondences is annotated.

There is a lot of work worth mentioning from different points as follows. For instance, Zhou et al. (2017) regress the 2D joint heatmaps and the combination of intermediate feature representations of the 2D pose estimator, where these multiple levels of semantic features provide additional clues for 3D pose estimation. Mehta et al. (2017a) address the generalization problem of 3D pose estimation by transfer learning. Their 3D prediction network (3DPoseNet) is the same as the 2D pose estimation network (2DPoseNet) in several layers and adds the 2D heatmaps prediction task as an auxiliary task. They balance the transferred features preservation and new pertinent features learning through a learning rate discrepancy between the transferred layers and the new layers. Sharma et al. (2019) propose a deep conditional variational autoencoder (CAVE) based model to generate 3D pose samples to reduce the depth ambiguity when lifting from 2D to 3D. The CAVE dataset is used to generate a set of 3D pose samples according to estimated 2D pose and latent code samples. These samples are scored using the ordinal relationships predicted from the image by a CNN. Finally, the estimated 3D pose is computed according to these scores and the corresponding 3D pose samples. Wang et al. (2019b) design a knowledge distilling model for one type of non-rigid structure from motion (NRSFM) methods. They use a 3D shape dictionary to recover camera matrices and codes, which can be used to reconstruct the depth information.

3.1.3. SMPL model based methods

Early work used the SCAPE body model (Anguelov et al., 2005) and fitted it to images using manually annotated keypoints and silhouettes. More recent works use the SMPL model (Loper et al., 2015) and fit it automatically. This is done by either solving an optimization problem to fit the model to the data (Bogo et al., 2016) or regressing the model parameters directly using a neural network (Kanazawa et al., 2018). Since the SMPL model incorporates prior knowledge about human shape, it can thus be fitted with very little data. Several optimization methods have been proposed for 3D human pose estimation. As a well-known method, the SMPLify (Bogo et al., 2016) first estimates 2D keypoints using DeepCut (Pishchulin et al., 2016) and then fits the SMPL model to these keypoints. The fitting procedure is guided by matching the 2D keypoints projected from the SMPL model and detected 2D keypoints. Lately, the SMPLify is improved in Lassner et al. (2017) by additionally matching the image silhouette and the silhouette projected from the SMPL model, where the silhouette is defined to be all pixels of a body's projection.

Recent methods regress the SMPL parameters directly by a variety of networks in different tasks. For example, an end-to-end deep learning scheme proposed by Kanazawa et al. (2018) to learn the mapping from image pixels to SMPL model parameters, as shown in Fig. 7. They also minimize the reprojection loss of keypoints, and train a discriminator by a large mesh database to determine if the generated shape and pose parameters are real. Pavlakos et al. (2018b) train a network called PosePrior that takes heatmaps as input and outputs pose parameters of the SMPL model and another network to estimate the shape parameters from the silhouette. Finally, the projected 3D pose is matched with annotated keypoints and masks. At the same time,

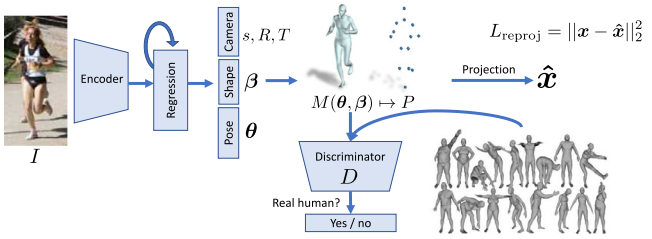

Fig. 7. Framework of human mesh recovery (HMR) (Kanazawa et al., 2018).

an encoder–decoder architecture is designed by Tan et al. (2018) to reduce the dependence on 3D human body shape and pose ground truth dataset. The decoder takes as input the SMPL parameters and outputs the corresponding silhouette. The decoder is then fixed and the whole network is trained end-to-end on the real image and the corresponding silhouette. By leveraging the task of semantic segmentation for body parts and the body constraints of SMPL, Omran et al. (2018) use a semantic segmentation CNN to segment the image into 12 semantic parts, and they then encode the semantic part probability maps into SMPL parameters.

For regression-based methods, they commonly suffer from mediocre image-model alignment due to the one-shot prediction and need for a huge amount of data. Kolotouros et al. (2019a) propose the SPIN (SMPL oPtimization IN the loop) to combine the merits of optimization-based and regression-based methods, where the regression results are used as the initialization for optimization and pixel accurate optimization stage could further exploit the supervision signal.

In addition, some methods have been developed to utilize more constraints and supervision signals. For example, Hassan et al. (2019) try to exploit the context information of a 3D scene by using the constraint that the human body model and the scene model cannot share the same 3D space. The 3D scene is constructed by an off-the-shelf solution, Kinect, and the signed distance field is used to penalize the body–scene inter-penetrations. Besides, the self-penetrations of a model are also considered by using the bounding volume hierarchies. Xu et al. (2019) propose to exploit dense pose information to get stronger supervision, where DensePose is used to produce the IUV maps, which represent the body pixel-to-surface correspondence. Finally, we summarize these methods in Table 2–3.

3.2. 3D human pose estimation from multi-view images

Multi-view images can reduce the ambiguity significantly. However, it is challenging to fuse information from multiple views. Typical methods include fusing multi-view 2D heatmaps (Pavlakos et al., 2017b; Tome et al., 2018; Qiu et al., 2019), enforcing multiple view consistency (Rhodin et al., 2018a,b), triangulation (Kocabas et al., 2019; Iskakov et al., 2019), and utilizing the SMPL model (Liang and Lin, 2019). We summarized these methods in Table 3.

To fuse multi-view information, different strategies have been designed. For example, Pavlakos et al. (2017b) combine the 2D joint heatmaps of each view using a 3D pictorial structures model. These heatmaps are back projected to a common discretized 3D space and the prior distribution is modeled by constraining the lengths of the limbs and the data likelihood by the heatmaps. Then, a pose is estimated by computing the mean of the joints' marginal distribution. Commonly, a multi-stage framework is widely used, such as (Tome et al., 2018), to iteratively refine the 3D pose estimation from multi-view images with 3D priors (Tome et al., 2017). In each stage, the inputs of the CNN are multi-view images and 2D pose heatmaps from the previous stage. Finally, the 3D poses are estimated by optimizing the latent 3D pose prior space consistent with 2D poses inferred from 2D heatmaps of all views. Simultaneously, this 3D pose is reprojected onto the 2D

Table 3
3D human pose estimation from multi-view images.

Multi-view images	Highlights	Dataset	Metric	Code
Pavlakos et al. (2017b)	Combine the 2D joint heatmaps of each view using a 3D pictorial structures model	KTH Football II; Human3.6M	PCP; MPJPE	Harvesting
Tome et al. (2018)	Extent "lifting" to multi-view and multi-stage, in each stage, the input are multi-view images and 2D pose heatmaps from previous stage; the 3D pose are estimated by optimizing the latent 3D pose prior space	Human3.6M	MPJPE, P3	No
Rhodin et al. (2018b)	Use multi-view consistency by forcing the system to predict the same pose for all views	Human3.6M; MPI-INF-3DHP; Ski Dataset	MPJPE, NMPJPE, PA-MPJPE; NMPJPE, PCK, NPCK	No
Rhodin et al. (2018a)	Learn a geometry-aware body representation by novel view synthesis	Human3.6M	MPJPE, NMPJPE, PA-MPJPE;	Unsupervised
Kocabas et al. (2019)	Use epipolar geometry method to recover the 3D pose from the 2D poses and use it as supervision	Human3.6M; MPI-INF-3DHP	MPJPE, NMPJPE, PA-MPJPE, mPSS; MPJPE, NMPJPE, PCK, NPCK, mPSS	EpipolarPose
Chen et al. (2019a)	Learn a latent geometry representation of the 3D pose with representational consistency (by multiplying rotation matrix) constraint	Human3.6M; MPI-INF-3DHP	P1, P2, P3; PCK, AUC	No
Iskakov et al. (2019)	Learn how to triangulate (the features maps are unprojected into 3D volumes, then the volumes from multiple views are aggregated and processed by a 3D convolutional neural network to output 3D heatmaps.)	Human3.6M; Panoptic	MPJPE; MPJPE	Triangulation
Liang and Lin (2019)	SMPL; synthesize a large dataset with multiple views, different shapes and clothes to train the model; multi-stage, where each stage estimates the parameters view by view; each regression block takes as input the images features and previous human body and camera estimates, and outputs corrective values	Human3.6M; MPI-INF-3DHP; Synthetic	MPJPE, P3; PCK, AUC, MPJPE; MPJPE/Hausdorff Distance	Shape_aware

image for each camera view to form 2D heatmaps, which are fused with 2D heatmaps regressed by the 2D estimator. Recently, Qiu et al. (2019) feed multi-view images into a CNN model to merge information from other views to the current one. Furthermore, they propose a recursive pictorial structure model to optimize the 3D poses, which can progressively reduce the quantization error to obtain better results.

The multi-view consistency of the same pose can also be utilized to design algorithms. For instance, the multi-view consistency is used in Rhodin et al. (2018b) as weak supervision, by forcing the system to predict the same pose from all views only during training. This approach greatly reduces the need for labeled data and can be applied to the environment that 3D human pose annotations are hard to obtain as sports. By employing a semi-supervision, an encoder–decoder network in Rhodin et al. (2018a) first learns a geometry-aware body representation using unlabeled multi-view images and then uses a small amount of supervision to learn a mapping from the our representation to actual 3D poses. Another approach to encoding geometry representation is to encode the 2D pose to a latent one of the 3D pose with a representation consistency constraint (Chen et al., 2019a). The encoder is trained using multi-view image pairs and the latent geometry representation of one image is multiplied by the relative rotation matrix from this image to the other. Then a decoder takes the rotated representation as input and tries to output the pose in the other image.

Triangulation is another fundamental method for reconstruction in computer vision. EpipolarPose (Kocabas et al., 2019) uses the epipolar geometry method to recover the 3D pose from the 2D poses and uses it as a supervision signal to train the 3D pose estimation model, as shown in Fig. 8. Iskakov et al. (2019) first propose a baseline method that feeds the 2D joint confidences and 2D positions of all views produced by the 2D pose detector to the algebraic triangulation module to obtain the 3D pose. The drawback of this method is that images from different cameras are processed independently. Therefore, a more powerful triangulation procedure is proposed by them. During processing, the feature maps are not projected into 3D volumes and the volumes from multiple views are aggregated and processed by a 3D CNN to output 3D heatmaps.

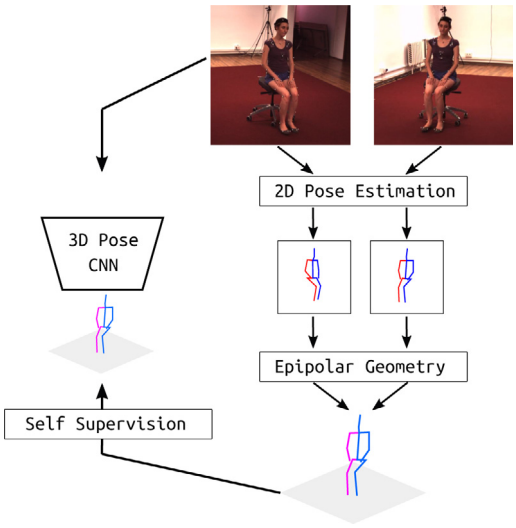

Fig. 8. Main framework of EpipolarPose (Kocabas et al., 2019).

Previous pose and shape estimation methods using silhouette as supervision cannot be directly applied to subjects with loose garments. To address this problem, Liang and Lin (2019) synthesize a large dataset with different views, shapes, and clothes, and design a model to be shape-aware. The model architecture consisting of multiple stages, where each stage estimates the parameters view by view. In general, it takes the encoded image features and previous human body as input, and camera estimates corrective values, e.g. camera parameters that are view-specific and human body parameters that are shared from all views.

Table 4
3D human pose estimation from a sequence of monocular images.

Methods	Highlights	Dataset	Metric	Code
Du et al. (2016)	Utilize height-map by a dual-stream network	HumanEva; Human3.6M	Average error; MPJPE, PCK	No
Tekin et al. (2016)	Directly regress a 3D pose from a spatio-temporal volume of bounding boxes centered on the target pose's frame	Human3.6M; HumanEva; KTH Multiview Football	MPJPE; 3D Error; PCP	No
Zhou et al. (2016)	Sparse representation of 3D poses; impose temporal smoothness on both pose coefficients and rotations; estimate the 3D pose sequence by penalized maximum likelihood estimation from 2D poses	Human3.6M	MPJPE	Sparseness
Mehta et al. (2017b)	Real-time; extend the 2D heatmap formulation to 3D using three additional location-maps	Human3.6m; MPI-INF-3DHP	MPJPE; PCK, MPJPE, AUC	Vnect-caffe Vnect-tf
Lin et al. (2017)	Multi-stage refinement with LSTM to enforce temporal consistency	Human3.6M; HumanEva-I	MPJPE; 3D Error	RPSM
Coskun et al. (2017)	Learns motion and noise models of the Kalman filter by LSTM to utilize temporal information	Human3.6M	MPJPE	Lstmkf
(Katircioglu et al., 2018)	Autoencoder with large latent representation to encode joints' dependencies; use LSTM to impose temporal constraint on the early features	Human3.6M; HumanEva-I; KTH Multiview Football II	P1, P3; 3D Error; PCP	No
Rayat Imtiaz Hossain and Little (2018)	Sequence-to-sequence model (2D pose sequence to 3D pose sequence); impose temporal smoothness constraint on the predicted 3D pose encouraging to be close to the pose of previous frame,	Human3.6M; HumanEva	MPJPE, P3; Mean Reconstruction Error	Pose_3D
Lee et al. (2018)	Use propagating LSTM networks (p-LSTMs) to infer pose in a central-to-peripheral order; the multi-stage architecture can capture temporal correlations	Human3.6M; HumanEva	P1, P2; 3D Error	No
Dabral et al. (2018)	Structure-Aware PoseNet with illegal joint-angle loss and left–right symmetry loss; Temporal PoseNet with a simple two-layer FCN to learn complex structural and motion cues.	Human3.6M; MPI-INF-3DHP	MPJPE, PAMPJE; MPJPE, PCK, AUC	No
Pavllo et al. (2019)	Capture long-term information by dilated temporal convolutions; semi-supervised	Human3.6M; HumanEva-I	MPJPE, NMPJPE, PA-MPJPE; MPJVE, MPJPE	VideoPose3D
Arnab et al. (2019a)	Specialize the multi-frame bundle adjustment to human pose estimation and apply it to unlabeled but real-world Youtube videos and generate a dataset as a weak supervision signal; Huber penalty function	Human 3.6M; 3DPW; HumanEva; Ordinal Depth	MPJPE, PA-MPJPE; PA-MPJPE; PA-MPJPE	Temporal-Kinectics

Table 5
3D human pose estimation from a sequence of multi-view images.

Methods	Highlights	Dataset	Metric	Code
Rhodin et al. (2016)	Sum-of-Gaussians representation; refined the model by fitting the contours of the person	HumanEva-I	3D Error	No
Joo et al. (2017)	Volume representation; 480 synchronized VGA views; project the center of the voxel to all 2D views to calculate the 3D joint likelihood; refine the 3D pose temporally by computing dense 3D point tracks (depth)	CMU Panoptic	Accuracy	No
Huang et al. (2017)	Multi-view SMPLify; SMPL model fits all views independently at each moment first; regularize the motion in time to estimate a consistent 3D shape in the entire sequence	Human3.6M; HumanEva	MPJPE	MuVS
Trumble et al. (2018)	Volumetric representation; use probabilistic visual hull (PVH) on views to form volumetric reconstruction; LSTM enforce temporal consistency on latent representation	Human3.6M; Total Capture	MPJPE; Average per joint error	No

4. 3D human pose estimation from image sequences

Recovering 3D human pose from a sequence of images is often the same as marker-free human performance capturing. With the development of technology, the number of videos is growing drastically, so it has become appealing to extract poses from a sequence of images. However, there are still several challenges to this. For example, the background variation, occlusion, camera movement, fast motion, loose clothing, illumination, may cause the shape and appearance of people to change dramatically over time. Some methods attempt to process video sequences in real-time systems, while others take the entire video as input and output a sequence of poses.

4.1. 3D human pose estimation from a sequence of monocular images

3D human pose estimation from a sequence of monocular images suffers from inherent depth ambiguity and is thus a ill-defined problem. To reduce the ambiguities, the image sequence is adopted as input by many works. The continue frames of the sequence often give multiple shots of the same person, while the bone length and shape of the

same person are invariant, and the movement of a person is often regular. As the task of videos, temporal relationship need to be more importantly learned using networks, such as LSTM (Coskun et al., 2017; Katircioglu et al., 2018; Lin et al., 2017; Lee et al., 2018; Rayat Imtiaz Hossain and Little, 2018), CNNs (Tekin et al., 2016; Pavllo et al., 2019), MLPs (Dabral et al., 2018), TCNs (Cheng et al., 2019), and GCNs (Cai et al., 2019). During training, some works penalize pose related parameters to generate temporally smooth poses, such as (Zhou et al., 2016; Du et al., 2016; Mehta et al., 2017b; Xu et al., 2018a; Tung et al., 2017). Alternately, some works try to optimize the trajectory, such as (Li et al., 2019b; Arnab et al., 2019a). We summarize the above-mentioned methods in Table 4.

From the perspective of network architectures, some works are described as follows. (1) An LSTM and the sequence-to-sequence model (Sutskever et al., 2014) are widely used in modeling temporal relationships of sequences. For example, the LSTM Kalman filter (LSTM-KF) is proposed in Coskun et al. (2017), which learns motion and noise models of the Kalman filter by LSTM in order to utilize temporal information. In order to capture abundant temporal information, recurrent 3D pose sequence machine (RPSM) is designed in Lin et al. (2017) to perform multi-stage refinement and capture long-range dependencies among multiple body-parts for 3D pose estimation. RPSM also enforces the temporal consistency of the predicted pose sequence. Similarly, the reconstructed 3D pose is refined in a multi-stage manner. Lee et al. (2018) first extract the 2D pose using a CNN and then employ the proposed propagating LSTM networks (p-LSTMs) to reconstruct a 3D pose. Additionally, Katircioglu et al. (2018) indicate that imposing the temporal constraint on the features earlier in the network is more effective than applying it to 3D pose predictions. Rayat Imtiaz Hossain and Little (2018) propose a sequence-to-sequence network that takes previous 2D poses as input and predicts a sequence of 3D poses relative to the root node (central hip). The network is composed of LSTM units with shortcut connections on the decoder side. The encoder of the network encodes the 2D pose sequence in a fixed size vector. (2) To enforce temporal consistency, CNNs-based structures have also been explored to process temporal sequences. Tekin et al. (2016) propose to directly regress the 3D pose from a spatio-temporal volume of bounding boxes centered on the target frame. The authors also note that extracting spatio-temporal features using 3D CNNs directly on the volume does not improve the performance much compared to using spatial CNNs. Recently, Pavllo et al. (2019) propose a temporal dilated convolutional model taking 2D keypoint sequences as input to estimate 3D poses. They capture long-term information by dilated temporal convolutions, and overall they first predict 2D poses with the 2D keypoint detector and then lift them to 3D poses. (3) Other network architectures have also been explored to solve the problem of 3D human pose estimation. For example, Dabral et al. (2018) propose the Temporal PoseNet, which is a simple two-layer fully connected network with rectified linear units (ReLUs). The model takes a fixed number of adjacent poses as input and outputs the required pose, and can learn complex structural and motion cues. To address the occlusion problem, Cheng et al. (2019) propose an occlusion-aware 2D temporal CNN that takes incomplete keypoints sequence of 2D poses as input to reduce the effect of the error-prone estimation of occluded joints. Concretely, they propose to use a *Cylinder Man Model* to generate 2D-3D pose pairs with occlusion labels to train the 3D TCN and regularize the occluded joints. Nowadays, Cai et al. (2019) represent a 2D pose sequence as a graph and design a local-to-global network to estimate the corresponding 3D pose sequence, where the network can capture multi-scale features and learn a temporal constraint for the pose sequence.

Seen from implementation of these approaches, several considerations are addressed, such as the temporal consistency of pose and shape, a variety of pose reconstruction procedure. More importantly, compared of single-time based methods, the pose or shape of a human should be temporally consistent, which can be enforced by penalizing the corresponding parameters. For example, Zhou et al. (2016) represent a 3D pose as the linear combination of pre-defined basis poses, and impose temporal smoothness on both the pose coefficients and rotations. Du et al. (2016) impose limb-length constraints and enforce the temporal constraint on 3D poses when estimate 3D motion from the estimated 2D pose sequence. Xu et al. (2018a) employ an actor-specific template mesh, and the human motion is parameterized with a kinematic skeleton and a medium-scale deformation field. They estimate the skeleton deformations in a batch manner by using both 2D and 3D pose and forcing the trajectory of each skeleton parameter to rely on a low dimensional linear subspace. That leads to temporal smoothness and solves the flipping ambiguities, as well as refine the surface by fitting automatically extracted monocular silhouettes. Besides, dynamic motion information under the video sequences is used by the structure-from-motion methods, *i.e* the motion reprojections are forced to match the 2D optical flow vectors. Some works use a parameterized model such as SMPL. For instance, Tung et al. (2017) exploit the video sequence and 2D joint heatmaps to predict the SMPL parameters with reprojection supervision. And some works consider that the root initialization is very important for the 3D pose reconstruction. For this purpose, Mehta et al. (2017b) first estimate both 2D and root (pelvis) relative 3D pose using a CNN. In detail, they exploit the predicted 2D and 3D pose, combined with the temporal history to estimate temporally consistent global 3D pose. At present, more works are focusing on weakly supervised and unsupervised manners. By employing a two-stage framework, Li et al. (2019b) try to utilize unannotated monocular videos. In the first stage, the initial predictions are obtained from a pose estimation network with only a few annotated videos. Then the initial predictions are used to supervise the further training of the pose estimation network by matrix completion methods applied to 3D trajectories. Arnab et al. (2019a) specialize the multi-frame bundle adjustment to human pose estimation. They generate a dataset as a weak supervision signal by applying bundle adjustment to unlabeled but real-world Youtube videos, and propose a 3D pose and mesh reconstruction algorithm to eliminate the estimation ambiguity.

4.2. 3D human pose estimation from a sequence of multi-view images

A few recent works fall into this category, which are introduced separately as follows. Conventional multi-view 3D human pose estimation methods have been well studied and their performance is better than single-view methods. However, they require expensive dense cameras and controlled studios. We summarize these methods in Table 5.

Many methods have been proved to be highly effective in estimating 2D joint points to guide model initialization, such as (Stoll et al., 2011; Rhodin et al., 2015). These methods are based on sum-of-Gaussians representations (Rhodin et al., 2016) and a 2D pose estimator is used (Tompson et al., 2014). Besides, the model is refined by fitting the contours of the person in this method. In addition, Trumble et al. (2018) illustrate that the probabilistic visual hull (PVH, (Grauman et al., 2003)) on views of several calibrated cameras can be used in forming a volumetric reconstruction. And the resulting volumetric representation with the coarse resolution is first upscaled via tricubic interpolation and then the upscaled volume is used as input of a convolutional autoencoder to learn a deep representation. Then the latent representation of pose sequences is processed by an LSTM to estimate 3D pose and enforce temporal consistency. More recently, Pavlakos et al. (2019) introduce a strategy to utilize appearance consistency in a video with different views, which is helpful for model-based pose estimation. The key idea is assuming that the changes in the texture of the person are not dramatic between frames, so the texture consistency can be used to help reconstruct the body model.

There are also some datasets designed to address this case. We first take the CMU Panoptic dataset as example provided by Joo et al. (2017). It first produces 2D keypoint locations and heatmaps of all synchronized views for all subjects using an off-the-shelf 2D pose estimator (Wei et al., 2016). In general, the basic framework is to use

Table 6
Methods for multi-person 3D pose estimation.

Method	Highlights	Input	Type	Dataset	Metric	Code
Belagiannis et al. (2014,b)	3D pictorial structure (3DPS) model; detect the 2D pose from all views and then create a reduced state space by triangulation; temporal term which encourages temporal consistency	Multi-view video	Top-Down	Campus; Shelf	PCP; PCP	No
Zanfir et al. (2018)	Fit the SMPL model to the predicted 3D pose by penalizing the cosine distance between limbs that are shared in both SMPL and DMHS representations; dense, pixel-wise semantic error function; Hungarian algorithm based on body shape, appearance and motion cues to solve the person assignment problem over time; optimize the trajectory guided by constant velocity priors	Monocular video	Top-Down	Human3.6M; CMU Panoptic	MPJPE; MPJPE	No
Mehta et al. (2018)	Use a fixed number of maps to encode 3D poses; exploit the body part association to enable the inference of an arbitrary number of people; MuCo-3DHP dataset	Monocular single	Bottom-Up	Human3.6M; MPI-INF-3DHP	MPJPE; PCK, AUC, MPJPE	SShot
Rogez et al. (2019)	Generate pose candidates at different locations and then classify and regress them	Monocular single	Top-Down	Human3.6M	P1, P2, P3	LCR
Moon et al. (2019)	Use RootNet to predict the coordinates of human root and PoseNet to predict 3D pose relative to the root	Monocolar single	Top-Down	Human3.6M; MuPoTS-3D	MPJPE, P2, MRPE; AUC, 3DPCK	RootNet PoseNet
Nie et al. (2019)	Structured pose representation (SPR), which comprises the root positions of subjects and corresponding body joint displacements	Monocular single	Bottom-Up, Single Stage	CMU Panoptic	3DPCK	SPM
Dong et al. (2019)	Incorporate appearance information and geometric information to solve the cross-view correspondence as a convex optimization problem with a cycle-consistency constraint; 3DPS model	Multi-view single	Top-Down	Campus; Shelf	PCP	Mvpose
Rhodin et al. (2019)	Learn a high-level scene representation (neural scene decomposition) to reduce the annotation labor; extend novel view synthesis for multiple persons by exploiting appearance similarity clues and geometry constraints	Multi-view single	Top-Down	Human3.6M	MPJPE, NMPJPE	NSD

the volume representation and project the center of the voxel to all 2D views to calculate the 3D joint likelihood. Specifically, the node proposals are calculated by non-maxima suppression (NMS) at each time instance; then part proposals are generated using the limb connectivity information; at last, the skeletal proposals are generated by using a dynamic programming method on previous part proposals. The second one is the MuVS (Multi-View SMPLify, Huang et al. (2017)) dataset that extends SMPLify (Bogo et al., 2016) to multi-view sequence data. In the first phase, a separate SMPL model fits all views independently at each moment, which is less ambiguous compared to the single view. In the second phase, the pose parameters are first initialized to the median of all the shape parameters obtained in the first phase, and then the motion is regularized in time to estimate a consistent 3D shape for the entire sequence. Unlike the research in Rhodin et al. (2015), Huang et al. (2017) explicitly use a deep CNN to segment people in the image from the background, eliminating the need for background images. They exploit temporal information based on discrete cosine transform (DCT, Akhter et al. (2012)) to solve possible left and right body parts confusion in the 2D joint estimator.

5. Multi-person 3D human pose estimation

Multi-person 3D pose estimation is more challenging than single human 3D pose estimation, due to the problems of much larger state space (all possible translations and rotations of the human body parts in 3D space), occlusions, and cross-view ambiguities when not knowing the identity of the humans in advance. In this section, we focus on multi-person 3D pose estimation. Similar to the introduction for single person case, these methods are reviewed and summarized in Table 6.

Compared with the single-person case, multi-person pose estimation is very different and more complex. First of all, the difficulty lies on how to distinguish different human joint points and body parts due to close distance and occlusion from each other. Second, the root joint localization is based on different assumptions. Unlike the single-person methods that predict root joint-relative 3D poses, the multi-person methods predict absolute 3D poses to differentiate people in the global 3D space. Third, the number of persons leads to the decrease of calculation efficiency and the increase of errors when detecting human body boxing boxes or joints. Existing multi-person methods typically adopt two-stage solutions, namely the **top-down** strategy that employs off-the-shelf detectors to localize person instances at first and then locates their joints individually, and the **bottom-up** strategy that locates all the body joints at first and then assigns them to the corresponding person.

5.1. 3D human pose estimation from a monocular single image

For multi-person estimation from a single monocular image, we introduce three two-stage methods (Rogez et al., 2019; Moon et al., 2019; Mehta et al., 2018) and a new single-stage method (Nie et al., 2019). It has been confirmed by Rogez et al. (2019) that enabling the network to predict the 2D and 3D poses of multiple people simultaneously by generating and scoring some pose proposals for each image is a promising way to promote both the accuracy and efficiency of pose estimation. Because this method benefits from the bypassing the requirement of human initial localization. They first employ a pose generator to generate pose candidates at different locations in the image, and then use the classifier to score the proposed poses and use a regression head to refine both the 2D and 3D pose proposals. As a result, a location-classification-regression network (LCR-Net) is trained in an end-to-end manner. Furthermore, recent studies on solving the problem reveal that using new designed frameworks and pose-maps are effective ways to figure out 3D poses among monocular images. For example, Moon et al. (2019) propose a framework to estimate 3D multi-person poses. They first adopt Mask R-CNN (He et al., 2017) to detect human bounding boxes and calculate image features. And then a RootNet and a PoseNet are employed to predict the coordinates

of human root and 3D poses relative to the roots, respectively. It is elucidated by Mehta et al. (2018) that introducing the novel occlusion-robust pose-maps (ORPM) can outputs a fixed number of maps to encode the 3D poses of everyone in the picture. Inspired by the 2D pose estimation method called part affinity fields (Cao et al., 2017), the authors exploit the body part association to enable the inference of an arbitrary number of people without the need for a detection bounding box. Since two-stage methods suffer low efficiency, a single-stage multi-person pose machine (SPM) is proposed in Nie et al. (2019) to overcome this problem. Though it follows a one-stage pipeline, strictly speaking, in this method, the joint points are first detected and then the root node is used to find or distinguish the different people. Therefore, we also consider this a type of bottom-up method. Concretely, a structured pose representation (SPR) is explored, which comprises the root positions of subjects and corresponding body joint displacements. Both root positions and joint displacements are estimated by two CNN branches based on Newell et al. (2016). Finally, the poses are recovered from the results of these two branches. Satisfactorily, this method can achieve high efficiency (20 fps) and accuracy on the CMU Panoptic dataset.

5.2. 3D human pose estimation from multiple views

For multi-person 3D pose estimation from multi-view images, inferring the cross-view correspondences among 2D pose predictions is the major bottleneck due to the possible incompleteness and low confidence of 2D poses. Thus, developing new strategies to solve this problem is critical and tremendous efforts has been made. Dong et al. (2019) incorporate appearance and geometric information to calculate the affinity between the detected 2D poses of two persons. They take all affinity matrices between two views as input and infer correspondence matrix. This matching problem is formulated as a convex optimization problem with a cycle-consistency constraint and is solved by using the result of Huang and Guibas (2013). Finally, the 3D poses are reconstructed using a 3D pictorial structures (3DPS) model (Belagiannis et al., 2014) and the 3D pose proposals are reconstructed from all pairs of 2D poses by triangulation for faster speed. Furthermore, a report from Rhodin et al. (2019) has demonstrated that a high-level scene representation for 3D human pose estimation from a single image can further reduce the annotation labor and can help to overcome the lack of a large dataset for pretraining the 3D pose estimator as image classification and object detection (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). They call this representation neural scene decomposition (NSD), which is formed by three aspects including the spatial layout, the 2D shape representation, and subject-specific appearance and 3D pose information. Specifically, the spatial layout is represented by bounding boxes and relative depth of subjects, and is instantiated by utilizing multi-view data for training. In a self-supervised manner, NSD can be trained using the proposed novel view synthesis (NVS) that exploits multi-view information by enforcing consistency when reconstructing results from one scene to a novel view. Moreover, the authors extend NVS for multiple persons by exploiting appearance similarity clues and geometry constraints. In their approach, the number of people in the scene and camera calibration are needed.

5.3. 3D human pose estimation from a sequence of monocular images

For multi-person 3D pose estimation from a monocular video, there are a fully automatic monocular visual sensing system is designed in Zanfir et al. (2018) for multiple people from a monocular image. They infer the 2D pose, 3D pose, and semantic body parts of multiple people by the deep multitask human sensing network (DMHS, Popa et al. (2017)). Then the SMPL model is fitted to the predicted 3D pose by penalizing the cosine distance between limbs which are shared in both the SMPL and DMHS representations. A new error function is also adopt in their strategy. This function measures the dense, pixel-wise semantic error between the semantic segmentation from DMHS and the projection of the SMPL model, combined with Euclidean distance, to refine the parameters of the SMPL model. In addition, a loss is defined as well to avoid simultaneously occupying the same 3D space volume. Furthermore, a ground plane is been estimated based on which most people stand but leave room for outliers who do not get in touch with the plane. Finally, to solve the person assignment problem over time, and then optimize the trajectory guided by constant velocity priors on pose angles and translation variables for all people throughout the video, they use a Hungarian algorithm based on body shape, appearance, and motion cues.

5.4. 3D human pose estimation from a sequence of multi-view images

Compared with single human pose estimation, multi-person 3D pose estimation from multiple views is more difficult due to the larger state space, occlusion, and cross-view ambiguities. Belagiannis et al. (2014) extend the pictorial structure model (PSM) used in 2D human pose estimation to solve this task. They first detect the 2D pose from all views and then create a reduced state space by triangulation of corresponding body joints to overcome the high-dimensional state space. To resolve ambiguities after triangulation, they propose the 3D pictorial structures (3DPS) model. As an extension, Belagiannis et al. (2014b) make the 3DPS model temporally consistent by adding a temporal term, which encourages temporal consistency of the human poses over time.

6. Performance analysis and experiments

In this section, we will give a detailed summary of the performance of state-of-the-art methods for the 3D human pose estimation task on the popular datasets, e.g. HumanEva, Human3.6M, MPI-INF-3DHP, 3DPW.

6.1. Summary of performance on HumanEva

The HumanEva dataset is still widely used in the 3D pose estimation community, therefore we summarize the performance (3D Error, mm) of 3D pose estimation methods on this dataset in Table 7. We could observe that the 3D error reduces significantly from 77.2 mm to 13.5 mm on the Walking sequence of HumanEva.

Only a small number of multi-view methods (Rhodin et al., 2016; Huang et al., 2017) have reported results on this dataset in recent years. (Sarafianos et al., 2016) note that the temporal information may not be well utilized. The method of Pavllo et al. (2019) achieves state-of-the-art results on almost all subjects of the three sequences, which uses a temporal dilated convolution to extract temporal information. The works of Tekin et al. (2016), Katircioglu et al. (2018), Rayat Imtiaz Hossain and Little (2018), Lee et al. (2018), Pavllo et al. (2019) explore how to utilize temporal information from the video. The performance of these methods validate the effectiveness and importance of the temporal constraint inherent in videos.

The SMPL model based methods (Bogo et al., 2016) rarely report performances on this sequence. They do not perform well at least on the HumanEva dataset. This may be due to the limitations of the SMPL model, and more advanced models may be required such as the Adam model (Joo et al., 2018).

6.2. Summary of performance on Human3.6M

The conventional methods are mainly evaluated on HumanEva, but most recent methods also report results on the Human3.6M dataset. Therefore, we mainly analyze the performance of the proposed 3D human pose estimation methods on Human3.6M. The MPJPEs of these methods are summarized in Table 8. The MPJPE reduces by about half from 117.3 mm to 39.9 mm in Table 8–(1).

Table 7
3D Error (mm) on the HumanEva dataset.

Method	*Walking*				*Jogging*				*Boxing*			
	S1	S2	S3	Avg.	S1	S2	S3	Avg.	S1	S2	S3	Avg.
Bogo et al. (2016)	73.3	59.0	99.4	77.2	–	–	–	–	82.1	79.2	87.2	82.8
Tekin et al. (2016)	37.5	25.1	49.2	37.3	–	–	–	–	50.5	61.7	57.5	56.5
Moreno-Noguer (2017)	19.7	13.0	24.9	19.2	39.7	20.0	21.0	26.9	–	–	–	–
Pavlakos et al. (2017a)	22.1	21.9	29.0	24.3	29.8	23.6	26.0	26.5	–	–	–	–
Martinez et al. (2017)	19.7	17.4	46.8	38.0	26.9	18.2	18.6	21.2	–	–		–
Pavlakos et al. (2018a)	18.8	12.7	29.2	20.2	23.5	15.4	14.5	17.8	–	–	–	–
Katircioglu et al. (2018)	29.3	17.3	62.6	36.4	–	–	–	–	–	–	–	–
Rayat Imtiaz Hossain and Little (2018)	19.1	13.6	43.9	25.5	23.2	16.9	15.5	18.5	–	–	–	–
Lee et al. (2018)	18.6	19.9	30.5	23.0	25.7	16.8	17.7	20.1	–	–	–	–
Sharma et al. (2019)	19.3	12.5	41.8	24.5	40.9	22.1	18.6	27.2	–	–	–	–
Pavllo et al. (2019)	13.9	10.2	46.6	23.6	**20.9**	**13.1**	**13.8**	**15.9**	**23.8**	**33.7**	**32.0**	**29.8**
Zhou et al. (2019)	**13.5**	**9.9**	**17.1**	**13.5**	24.5	14.8	14.4	17.9	–	–	–	–

The effectiveness of utilizing multi-view and sequence images is verified by methods in Table 8–(2) and (-3), respectively. Although the number of methods utilizing a multi-view video is small, these methods work well, see Table 8–(4). The SMPL model based methods (Arnab et al., 2019a; Liang and Lin, 2019; Huang et al., 2017; Bogo et al., 2016; Kanazawa et al., 2018; Lassner et al., 2017) are comparable to other methods. Liang and Lin (2019) achieve 45.1 mm MPJPE. The weakly supervised methods (Zhou et al., 2017; Fish Tung et al., 2017; Pavlakos et al., 2018a; Wandt and Rosenhahn, 2019) achieve impressive performance. Ordinal depth, multi-view information, appearance consistency of video, and 3D pose geometry structure are shown to be effective supervision for weakly-supervised 3D human pose estimation. Although the performance of weakly supervised methods is lower than fully supervised approaches, they require far fewer data and can significantly reduce annotation labor.

6.3. Performance analysis on MPI-INF-3DHP

The MPI-INF-3DHP dataset has more action classes than Human3.6M. While, MPI-INF-3DHP has undergone multiple changes to the test set annotations, which makes comparison across papers difficult. Therefore, we additionally investigate the training and testing protocols for reference. The most used metrics for this dataset are PCK, AUC, and MPJPE, which are summarized in Table 9.

Large improvements have been achieved in recent years. Liang and Lin (2019) even achieved 95.0 in PCK, demonstrating the effectiveness of utilizing the SMPL model and multi-view information. A few methods (Mehta et al., 2017b; Dabral et al., 2018) utilize temporal constraints on this dataset, which obtain better performance than other works. This is because video sequences provide continuous information, which helps reduce ambiguity. Similar methods that exploit multi-view images use stronger supervision and also outperform most methods using only a single monocular image. Besides, many works, such as Wandt and Rosenhahn (2019), Xu et al. (2019), use adversarial learning to obtain more accurate predictions. These works also achieve state-of-the-art performance, proving that adversarial learning is helpful.

6.4. Summary of performance on 3DPW

The 3DPW dataset is the first in-the-wild dataset created by von Marcard et al. (2018). Since 3DPW is a relatively new benchmark, most literature report results on Human3.6M, but not 3DPW. Some work is to solve the occlusion problem and the viewpoint problem, such as (Cheng et al., 2020; Wang et al., 2020). Because of the small amount of information in a single image, most works adopt a parametric model like SMPL and learn to predict the shape and pose coefficients, such as (Arnab et al., 2019b; Sun et al., 2019; Sengupta et al., 2020; Choutas et al., 2020; Kolotouros et al., 2019b; Moon and Lee, 2020; Choi et al., 2020; Kocabas et al., 2020; Lin et al., 2020). Some other methods explore the temporal and shape consistency of time series to improve the accuracy of modeling, such as (Arnab et al., 2019b; Sengupta et al., 2020; Luo et al., 2020; Lin et al., 2020). The most popularly used metrics for this dataset are MPJPE (mm) and PA-MPJPE (mm), which are summarized in Table 10.

Specifically, to deal with occlusion, Cheng et al. (2020) apply data augmentation and multi-scale spatial features for 2D keypoints prediction in each frame, and multi-stride temporal convolutional networks (TCNs) to estimate 3D keypoints. To reduce camera parametric bias, Wang et al. (2020) predict the camera viewpoint as an auxiliary task to significantly reduce the 3D joint prediction error and improve generalization in cross-dataset 3D human pose evaluation.

To resolve ambiguities and address the lack of real-world data in monocular 3D pose estimation, Arnab et al. (2019b) exploit temporal consistencies across a video with bundle adjustment. They leverage predictions on real-world videos as a source of weak supervision to improve existing 3D pose estimation models and retrain the single-frame 3D pose estimator to improve performance on the real-world dataset. Sengupta et al. (2020) use multi-frame optimization, with forced shape consistency between frames, to obtain pseudo-ground-truth SMPL shape and pose parameters for the sports-person in each image. Luo et al. (2020) generate smooth and accurate 3D human pose and motion estimates from RGB video sequences using the autoencoder architecture. Lin et al. (2020) use a transformer encoder to jointly model vertex–vertex and vertex–joint interactions, and outputs 3D joint coordinates and mesh vertices simultaneously.

There are many other works based on the parametric model to regress 3D poses or meshes. For example, Choutas et al. (2020) propose

Table 8
Estimating 3D human pose on the Human3.6M dataset using different inputs in terms with MPJPE (mm)

(1) A single monocular image	Dir	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.	Settings
Park et al. (2016)	100.3	116.2	90.0	116.5	115.3	150.6	117.6	106.9	137.2	190.8	105.8	125.1	131.9	62.6	96.2	117.3	P1; 17 joints
(Zhou et al., 2016a)	91.8	102.4	97.0	98.8	113.4	125.2	90.0	93.8	132.2	159.0	106.9	94.4	126.0	79.0	99.0	107.3	P1; 17 joints; pre-trained on ImageNet; randomly sample 800k frames for training
Bogo et al. (2016)	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	79.7	86.8	81.7	82.3	P3; 14 joints
Fish Tung et al. (2017)	77.6	91.4	89.9	88.0	107.3	110.1	75.9	107.5	124.2	137.8	102.2	90.3	–	78.6	–	97.2	P1; 50/3fps; detected 2D keypoints
Tome et al. (2017)	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8	110.2	173.9	85.0	85.8	86.3	71.4	73.1	88.4	P1; 17 joints; 10fps; evaluated on all trials
Lassner et al. (2017)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	80.7	P3; 14 joints; 10fps
Moreno-Noguer (2017)	69.5	80.1	78.2	87.0	100.7	76.0	69.6	104.7	113.9	89.8	102.7	98.4	79.1	82.4	77.1	87.3	P1; testing in all images
Mehta et al. (2017a)	57.5	68.6	59.6	67.3	78.0	56.9	69.1	100.0	117.5	69.4	82.4	68.0	55.2	76.5	61.4	72.9	P1; 17 joints; initialized from ImageNet; extra trained on MPI-INF-3DHP
Nie et al. (2017)	62.8	69.2	79.6	78.8	80.8	72.5	73.9	96.1	106.9	88.0	86.9	70.7	71.9	76.5	73.2	79.5	P2; removed some poses without synchronized images
Pavlakos et al. (2017a)	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	P1; 10fps; a single model for all actions from all views
(Tekin et al., 2017)	54.2	61.4	60.2	61.2	79.4	63.1	81.6	70.1	107.3	69.3	78.3	70.3	51.8	74.3	63.2	69.7	P1; 17 joints; monocular in all views for training and testing
Zhou et al. (2017)	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	P1; 10fps for training and testing
Martinez et al. (2017)	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9	P1; 17 joints; all views; single action model
Sun et al. (2017)	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	P1; 17 joints
Kanazawa et al. (2018)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	88.0	P1; 10fps
Yang et al. (2018)	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	P1; all views and joints after aligning the depth of the root joints
Pavlakos et al. (2018a)	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2	P1; 10fps; single action model for all actions
Luvizon et al. (2018)	49.2	51.6	47.6	50.5	51.8	60.3	48.5	51.7	61.5	70.9	53.7	48.9	57.9	44.4	48.9	53.2	P1; 17 joints; MPII and Human3.6M for training
Lee et al. (2018)	40.2	49.2	47.8	52.6	50.1	75.0	50.2	43.0	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8	P1; 10fps; all views
Zhao et al. (2019)	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	P1; 10fps for training and testing
Habibie et al. (2019)	54.0	65.1	58.5	62.9	67.9	54.0	60.6	82.7	8.2	63.3	75.0	61.2	50.0	66.9	56.5	65.7	P1; 5fps; using 2D labeled datasets during training
Li and Lee (2019)	43.8	48.6	49.1	49.8	57.6	61.5	45.9	48.3	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7	P1; 17 joints; single action model for all views
Sharma et al. (2019)	42.9	48.1	47.8	50.2	56.1	65.0	44.9	48.6	61.8	69.9	52.6	50.4	56.0	42.1	45.1	52.1	P1; 17joints; 10fps; evaluated on all views and trials
Wang et al. (2019a)	44.7	48.9	47.0	49.0	56.4	67.7	48.7	47.0	63.0	78.1	51.1	50.1	54.5	40.1	43.0	52.6	P1; all views and joints after aligning the depth of the root joints
Wandt and Rosenhahn (2019)	50.0	53.5	44.7	51.6	49.0	58.7	48.8	51.3	51.1	66.0	46.6	50.6	42.5	38.8	60.4	50.9	P1; 17 joints
Zhou et al. (2019)	34.4	42.4	36.6	42.1	38.2	39.8	34.7	40.2	45.6	60.8	39.0	42.6	42.0	29.8	31.7	39.9	P1; 17 joints; single action model for all views
(2) Multi-view images	Dir	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.	Settings
Martinez et al. (2017)	46.5	48.6	54.0	51.5	67.5	70.7	48.5	49.1	69.8	79.4	57.8	53.1	56.7	42.2	45.4	57.0	P1; reported from Tome et al. (2018)
Pavlakos et al. (2017b)	41.2	49.2	42.8	43.4	55.6	46.9	40.3	63.7	97.6	119.0	52.1	42.7	51.9	41.8	39.4	56.9	P1; 17 joints
Tome et al. (2018)	43.3	49.6	42.0	48.8	51.1	64.3	40.3	43.3	66.0	95.2	50.2	52.2	51.1	43.9	45.3	52.8	P1; 17 joints; every 5th frame for evaluation
Rhodin et al. (2018a)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	131.7	P1; cropped images; semi-supervised S1(full 3D ground truth)
Rhodin et al. (2018b)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	66.8	P1; 16 joints; 10fps
Kocabas et al. (2019)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	51.8	P1; fully-supervised; every 64th frame for evaluation
Pavlakos et al. (2019)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	110.7	P1; semi-supervised S1(full 3D ground truth)
Chen et al. (2019a)	41.1	44.2	44.9	45.9	46.5	39.3	41.6	54.8	73.2	46.2	48.7	42.1	35.8	46.6	38.5	46.3	P1; add 3D structure prior
Liang and Lin (2019)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	45.1	P1; 14 joints; Procrustes Aligned results
Qiu et al. (2019)	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6	32.1	26.9	31.0	25.6	25.0	28.1	24.4	26.2	P1; single action model for all views; extra training on MPII
Iskakov et al. (2019)	18.8	20.0	19.3	18.7	20.2	19.3	18.7	22.3	23.3	29.1	21.2	20.3	19.3	21.6	19.8	20.8	P1; 17 joints; every 5th frame for the evaluation
(3) A sequence of monocular images	Dir	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.	Settings
Du et al. (2016)	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5	P1; 17 joints; 1 out of 50 frames from all 4 cameras for training and every 5th frame from camera 2 for testing
Tekin et al. (2016)	102.4	147.7	88.8	125.3	118.0	112.4	129.2	138.9	224.9	118.4	182.7	138.8	55.1	126.3	65.8	125.0	P1; 17 joints; all camera views for each separate action
Zhou et al. (2016)	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	P1; 10fps; evaluated on the bounding box crops
Mehta et al. (2017b)	61.7	77.8	64.6	70.3	90.5	61.9	79.8	113.2	153.1	80.9	94.4	75.1	54.9	83.5	61.0	82.5	P1; 17 joints; evaluated on all actions
Lin et al. (2017)	58.0	68.2	63.3	65.8	75.3	61.2	65.7	98.7	127.7	70.4	93.1	68.2	50.6	72.9	57.7	73.1	P1; 2fps; trained on training samples from all 15 actions
Coskun et al. (2017)	57.8	64.6	59.4	62.8	71.5	57.5	60.4	80.2	104.1	66.3	80.5	61.2	52.6	70.0	60.1	67.3	P1; all joint positions relative to a root joint
Katircioglu et al. (2018)	69.6	93.8	69.0	96.5	103.4	83.4	85.2	116.6	147.6	87.2	120.5	95.3	55.9	85.7	64.7	91.6	P1; 17 joints; input images crops
Lee et al. (2018)	40.2	49.2	47.8	52.6	50.1	75.0	50.2	43.0	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8	P1; 10fps; all views
Dabral et al. (2018)	44.8	50.4	44.7	49.0	52.9	43.5	45.5	63.1	87.3	51.7	61.4	48.5	37.6	52.2	41.9	52.1	P1; evaluated on the bounding box crops; extra trained on MPII
Rayat Imtiaz Hossain and Little (2018)	44.2	46.7	52.3	49.3	59.9	59.4	47.5	46.2	59.9	65.6	55.8	50.4	52.3	43.5	45.1	51.9	P1; a single model for all actions
Arnab et al. (2019a)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	77.8	P1; 10fps
Pavllo et al. (2019)	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8	P1; 17 joints; a single model for all actions
(4) A sequence of multi-view images	Dir	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.	Settings
Trumble et al. (2017)	92.7	85.9	72.3	93.2	86.2	101.2	75.1	78.0	83.5	94.8	85.8	82.0	114.6	94.9	79.7	87.3	P1; 17 joints
Huang et al. (2017)	44.3	47.0	51.8	45.0	67.7	54.6	49.3	48.9	72.8	76.5	63.7	116.2	55.4	42.9	37.2	58.2	P1; 17 joints; evaluated on all views

a body-driven attention to quickly and accurately capture the poses of 3D bodies, faces, and hands together from an RGB image. In Kolotouros et al. (2019b), a topology of the SMPL template mesh is retained, but instead of predicting model parameters, directly regressing the 3D location of mesh vertices. Moon and Lee (2020) propose I2L-MeshNet, an image-to-lixel (line+pixel) prediction network. A novel graph convolutional neural network-based system (Choi et al., 2020) is explored to estimate the 3D coordinates of human mesh vertices directly from the 2D human pose. To make use of existing large-scale motion capture dataset (AMASS, Mahmood et al. (2019)) together with unpaired, in-the-wild, 2D keypoint annotations, a video inference for body pose and shape estimation (VIBE) is proposed in Kocabas et al. (2020).

6.5. Summary of performance of multiple persons

Multi-person 3D pose estimation is a problem that has not yet been extensively addressed. In fact, several works provide their new datasets, such as the Shelf and Campus datasets (Belagiannis et al., 2014), the MuCo-3DHP, and MoPoTS-3D datasets (Mehta et al., 2018), the Boxing dataset (Rhodin et al., 2019), but they have not been popularly used in other literature.

The results on Shelf, Campus, and Boxing datasets are summarized in Table 11. To resolve the ambiguities of mixed body parts of multiple humans after triangulation, Belagiannis et al. (2014) introduce a novel 3D pictorial structure model and achieves high PCP performance. However, this 3DPS-based approach is computationally expensive due to the huge state space. Besides, it is not robust particularly when the number of cameras is small. Therefore, Dong et al. (2019) propose a multi-way matching algorithm to address the aforementioned challenges and achieves better performance by a large margin. For the Boxing dataset that comprises 8 sequences with sparring fights between 11 different boxers, Rhodin et al. (2019) propose to learn a neural scene decomposition (NSD) representation that is optimized for 3D human pose estimation tasks. Compared with Rogez et al. (2019), better performance has been achieved. The 3DPCK results of the state of the art on the MoPoTS-3D dataset are reported in Table 12. Note that results are sequence-wised and the accuracy is obtained for all ground truths. Notably, Moon et al. (2019) propose a fully learning-based camera distance-aware top-down approach that consists of human detection, 3D human root localization, and root-relative 3D single-person pose estimation models. This method has great potential to be further applied to 3D multi-person pose estimation.

Since previous works (Dong et al., 2019) only conduct qualitative evaluations on the CMU Panoptic dataset, there are few comparisons reported with different settings. For example, Nie et al. (2019) separate 10k images from the dataset to form the testing split and use the remaining images for training, achieving 77.8% 3DPCK. Zanfir et al. (2018) select data from 4 activities (*Haggling*, *Mafia*, *Ultimatum* and

Table 9
Performance of methods estimating 3D human pose on the MPI-INF-3DHP dataset using different inputs.

(1) A single monocular image	PCK	AUC	MPJPE	Training and testing protocols
Zhou et al. (2017)	69.2	32.5	–	Using its test set split; employing average PCK (with a threshold 150mm), after aligning the root joint (pelvis); moving the pelvis and hips towards neck in a fixed ratio (0.2)
Mehta et al. (2017a)	76.5	40.8	–	With weight transfer from 2DPoseNet by scene setting
Pavlakos et al. (2018a)	71.9	35.3	–	Following the typical protocol (Zhou et al., 2017; Mehta et al., 2017a)
Li and Lee (2019)	67.9	–	–	Only using the test split
Habibie et al. (2019)	69.6	35.5	127.0	After training on Human3.6M
Chen et al. (2019b)	64.3	31.6	–	After training on Human3.6M; 14 joints
Kanazawa et al. (2018)	86.3	47.8	89.8	After rigid alignment
Wandt and Rosenhahn (2019)	82.5	58.5	97.8	Using the training set of MPI-INF-3DHP
Xu et al. (2019)	89.0	49.1	83.5	Using all sequences from S1–S7 as training set and sequences from S8 as testing set; applying rigid transformations
Nibali et al. (2019)	85.4	47.0	91.3	After training on Human3.6M and MPI-INF-3DHP; 17 joints; using universally-scaled skeletons (fixed scale of 920 mm knee–neck); Since the scale is known, the ground truth root joint depth is not used to find the absolute depth of the predicted skeleton
(2) Multi-view images	PCK	AUC	MPJPE	Training and testing protocols
Rhodin et al. (2018b)	–	–	–	Supervised training on MPII-3DHP S1, weakly-supervised on S2 to S8; 17 joints; known rotations; NPCK: 73.1; NMPJPE: 119.8
Kocabas et al. (2019)	77.5	–	109.0	Supervised training; following the standard protocol: The five chest-height cameras and the provided 17 joints; NPCK: 78.1; NMPJPE: 106.4
Chen et al. (2019a)	75.9	36.3	–	After training on Human3.6M
Liang and Lin (2019)	95.0	65.0	59.0	Without synthetic training
(3) A sequence of monocular images	PCK	AUC	MPJPE	Training and testing protocols
Mehta et al. (2017b)	75.7	39.3	117.6	With the 3D joint position lookup in the location-maps done using the ground truth 2D locations rather the predicted 2D locations.
Dabral et al. (2018)	76.7	39.1	103.8	Skeleton fitting is done as an optional step to fit the pose into a skeleton of known bone lengths.

Table 10
Performance of methods estimating 3D human pose on the 3DPW dataset.

Method	MPJPE	PA-MPJPE	Training and testing protocols
Arnab et al. (2019b)	–	72.2	After training on its original data and 300K and 3M frames from their Kinetics dataset
Cheng et al. (2020)	–	71.8	Do not train on 3DPW and only use its testing set for quantitative evaluation
Sun et al. (2019)	–	69.5	Testing set for quantitative evaluation
Sengupta et al. (2020)	–	66.8	Testing set for quantitative evaluation
Wang et al. (2020)	89.7	–	Validation set for quantitative evaluation
Choutas et al. (2020)	93.4	60.7	Predictions for the main body area, excluding the head and hands
Kolotouros et al. (2019b)		59.2	Validation set for quantitative evaluation; no training data from 3DPW
Moon and Lee (2020)	93.2	58.6	Using MuCo-3DHP for the additional training dataset
Choi et al. (2020)	89.2	58.9	Trained on Human3.6M, COCO, and MuCo-3DHP
Kocabas et al. (2020)	82.9	51.9	Trained with 3DPW training set
Luo et al. (2020)	86.9	54.7	Trained without the Human3.6M dataset and SMPL supervision
Lin et al. (2020)	77.1	47.9	Trained with 3DPW training set

Pizza) which contain multiple people interacting with each other, and reports their pose and translation estimation errors.

7. Future potential development

Although 3D human pose estimation methods based on deep learning have achieved significant progress in recent years, challenges still exist due to the complexity of the task. We propose several next works worthy of attention and future directions for 3D human pose estimation as follows.

- *Multi-person 3D Pose Estimation.* Multi-person situations are very common in practice. However, 3D pose estimation methods still suffer under complex environments, such as human–human interactions and occlusions, while 2D multi-person pose estimation methods achieve satisfactory performance. Some existing works address these problems by using multiple views and scene information, but they can still be improved.
- *Weak Supervision.* Although many 3D pose estimators perform well on particular datasets, it remains difficult to directly generalize them to practical scenes. One way to solve this problem would

Table 11
Performance of methods estimating 3D human pose of multiple persons on the Shelf, Campus and Boxing datasets.

Method	Shelf@PCP				Campus@PCP				Boxing		
	Actor 1	Actor 2	Actor 3	Avg.	Actor 1	Actor 2	Actor 3	Avg.	MPJPE	NMPJPE	Detection rate
Belagiannis et al. (2014)	66	65	83	71.3	82	72	73	75.6	–	–	–
Belagiannis et al. (2014b)	75	67	86	76	83	73	78	78	–	–	–
Dong et al. (2019)	98.8	94.1	97.8	96.9	97.6	93.3	98.0	96.3	–	–	–
Rogez et al. (2019)	–	–	–	–	–	–	–	–	155.6	154.4	79.7
Rhodin et al. (2019)	–	–	–	–	–	–	–	–	125.4	99.7	99.8

Table 12
Performance of methods estimating 3D human pose of multiple persons on the MuPoTS-3D dataset in terms of 3DPCK.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg.
Rogez et al. (2017)	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
Mehta et al. (2018)	81.0	60.9	64.4	63.0	69.1	30.3	65.0	59.6	64.1	83.9	68.0	68.6	62.3	59.2	70.1	80.0	79.6	67.3	66.6	67.2	66.0
Rogez et al. (2019)	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
Moon et al. (2019)	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

be to leverage weak supervision, which exploits large amounts of data with only 2D pose annotations. This would be helpful because the 2D pose datasets are much larger than current 3D pose datasets and incur relatively lower annotation costs.

- *Model-based Methods.* There are many advantages of model-based methods, such as stability and scalability. For stability, the predictions of model-based methods are more robust than those of skeleton-based methods, especially under occlusions or in the wild. For scalability, model-based methods can be combined with other methods easily. For instance, (1) more keypoint information can be exploited and stronger supervision can be obtained by using DensePose; (2) errors can be reduced due to inter-penetrations by exploiting scene constraints, as well as the multi-person situation; (3) spatial–temporal information can be exploited in a straightforward way which only needs to consider the pose and the body shape parameters; (4) the gap can be bridged between human pose and texture of appearance, which is potentially useful in other tasks, such as person re-identification.
- *Interaction and Reconstruction between Scene Object and Human.* Advances in deep learning techniques have allowed recent work to reconstruct the shape of a single object given only one RBG image as input. Many works aim to capture overall object geometry, such as (Popov et al., 2020; Wei et al., 2020). As known, 3D human pose estimation can be used to recover sparse joint points (skeleton) or dense mesh points (shape). Reconstructing object geometry can provide extra information (e.g., depth and occlusion) to facilitate 3D human pose estimation. While, to our knowledge, the combination of scene and human interaction and reconstruction is still immature and has not even been paid attention to.
- *Human Pose Estimation for Scene Understanding.* In an image, the presence of a human is more attractive, so using more information captured from them can have a better understanding of the scene, such as visual question answering (VQA, Agrawal et al. (2015)) in the field of cross-modal understanding. While, existing works mainly focus on action recognition or abnormal behavior detection (Lentzas and Vrakas, 2019) rather than the scene understanding guided by human poses.
- *Performance Improved by Neural Architecture Search.* Neural architecture search (NAS) is a hot topic in the field of artificial intelligence in recent years, which is especially suitable for industry. It can greatly reduce the workload of manual parameter adjustment and find a more efficient network structure. While few works consider NAS for human activity recognition (Peng et al., 2020). Besides, multi-objective NAS (e.g., accuracy, model size) not only reveals the potential for human pose estimation in theoretical exploration but also can play a role in practical systems for pose estimation.

8. Conclusion

In this review, we summarize the recent progress of 3D human pose estimation from RGB images and videos. We observe that this problem has become increasingly popular in the computer vision community and, recently, great performance achievements have been made on the Human3.6M, HumanEva, and MPI-INF-3DHP datasets. However, the generalization to scenarios in the wild remains extremely challenging. As for multiple-person cases, single-stage methods are less developed, indicating that 3D human pose estimation in real-world scenarios is far from being established. Most recently, a comprehensive understanding of scenes and poses has drawn great attention. Furthermore, deep learning is very effective in solving this problem, so we can expect many innovations in the next few years, especially when new deep learning technologies are applied to this field. In addition, we conjecture that research on robustness, security, and federated learning for 3D human pose estimation will also be a promising direction in the future.

CRediT authorship contribution statement

Jinbao Wang: Conceptualization, Methodology, Data curation, Investigation, Visualization, Writing - review & editing. **Shujie Tan:** Conceptualization, Methodology, Data curation, Writing - original draft. **Xiantong Zhen:** Supervision, Writing - review & editing. **Shuo Xu:** Data curation, Investigation, Validation. **Feng Zheng:** Supervision, Project administration, Funding acquisition, Writing - review & editing. **Zhenyu He:** Supervision, Writing - review & editing. **Ling Shao:** Supervision, Writing - review & editing.

Declaration of competing interest

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

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61972188.

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D., 2015. VQA: Visual question answering. Int. J. Comput. Vis. 123, 4–31.

Airò Farulla, G., Pianu, D., Cempini, M., Cortese, M., Russo, L., Indaco, M., Nerino, R., Chimienti, A., Oddo, C., Vitiello, N., 2016. Vision-based pose estimation for robot-mediated hand telerehabilitation. Sensors 16 (2), 208.

Akhter, I., Black, M.J., 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1446–1455.

Akhter, I., Simon, T., Khan, S., Matthews, I., Sheikh, Y., 2012. Bilinear spatiotemporal basis models. ACM Trans. Graph. 31 (2), 17.

Alp Güler, R., Neverova, N., Kokkinos, I., 2018. Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7297–7306.

Alp Guler, R., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I., 2017. Densereg: Fully convolutional dense shape regression in-the-wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6799–6808.

Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B., 2018. Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5167–5176.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014a. 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3686–3693.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014b. 2D human pose estimation: New Benchmark and state of the art analysis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3686–3693.

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J., 2005. SCAPE: shape completion and animation of people. In: SIGGRAPH 2005.

Arnab, A., Doersch, C., Zisserman, A., 2019a. Exploiting temporal context for 3D human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3395–3404.

Arnab, A., Doersch, C., Zisserman, A., 2019b. Exploiting temporal context for 3D human pose estimation in the wild. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3390–3399.

Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S., 2014a. 3D pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1669–1676.

Belagiannis, V., Wang, X., Schiele, B., Fua, P., Ilic, S., Navab, N., 2014b. Multiple human pose estimation with temporally consistent 3D pictorial structures. In: European Conference on Computer Vision. Springer, pp. 742–754.

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J., 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: European Conference on Computer Vision. Springer, pp. 561–578.

Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.-J., Yuan, J., Thalmann, N.M., 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2272–2281.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y., 2018. Openpose: realtime multi-person 2D pose estimation using part affinity fields. ArXiv Preprint arXiv:1812.08008.

Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7291–7299.

Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.-C., 2019a. Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. arXiv preprint arXiv:1909.01507.

Chen, X., Lin, K.-Y., Liu, W., Qian, C., Lin, L., 2019b. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10895–10904.

Chen, Y., Tian, Y., He, M., 2020. Monocular human pose estimation: A survey of deep learning-based methods. Comput. Vis. Image Underst. 192, 102897.

Chen, C.-H., Tyagi, A., Agrawal, A., Drover, D., MV, R., Stojanov, S., Rehg, J.M., 2019a. Unsupervised 3D Pose Estimation with Geometric Self-Supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5714–5724.

Cheng, Y., Yang, B., Wang, B., Tan, R., 2020. 3D Human pose estimation using spatio-temporal networks with explicit occlusion training. In: AAAI.

Cheng, Y., Yang, B., Wang, B., Yan, W., Tan, R.T., 2019. Occlusion-aware networks for 3D human pose estimation in video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 723–732.

Choi, H., Moon, G., Lee, K.M., 2020. Pose2mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. ArXiv, arXiv:abs/2008.09047.

Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J., 2020. Monocular expressive body regression through body-driven attention. ArXiv arXiv:abs/2008.09062.

Ci, H., Wang, C., Ma, X., Wang, Y., 2019. Optimizing network structure for 3D human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2262–2271.

Coskun, H., Achilles, F., DiPietro, R., Navab, N., Tombari, F., 2017. Long short-term memory kalman filters: Recurrent neural estimators for pose regularization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5524–5532.

Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A., Jain, A., 2018. Learning 3d human pose from structure and motion. In: Proceedings of the European Conference on Computer Vision, ECCV. pp. 668–683.

Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X., 2019. Fast and robust multi-person 3d pose estimation from multiple views. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7792–7801.

Du, X., Vasudevan, R., Johnson-Roberson, M., 2019. Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction. IEEE Robot. Autom. Lett. 4 (2), 1501–1508.

Du, Y., Wong, Y., Liu, Y., Han, F., Gui, Y., Wang, Z., Kankanhalli, M., Geng, W., 2016. Marker-less 3d human motion capture with monocular image sequence and height-maps. In: European Conference on Computer Vision. Springer, pp. 20–36.

Elhayek, A., de Aguiar, E., Jain, A., Thompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C., 2016. MARCOnI—ConvNet-Based MARker-less motion capture in outdoor and indoor scenes. IEEE Trans. Pattern Anal. Mach. Intell. 39 (3), 501–514.

Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., Cucchiara, R., 2018. Learning to detect and track visible and occluded body joints in a virtual world. In: Proceedings of the European Conference on Computer Vision, ECCV. pp. 430–446.

Ferrari, V., Marin-Jimenez, M., Zisserman, A., 2008. Progressive search space reduction for human pose estimation. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.

Ferrari, V., Marin-Jimenez, M., Zisserman, A., 2009. Pose search: retrieving people using their pose. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.

Fish Tung, H.-Y., Harley, A.W., Seto, W., Fragkiadaki, K., 2017. Adversarial inverse graphics networks: learning 2D-to-3D lifting and image-to-image translation from unpaired supervision. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4354–4362.

Grauman, K., Shakhnarovich, G., Darrell, T., 2003. A bayesian approach to image-based visual hull reconstruction. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 1. IEEE, I–I.

Green, R., 2003. Spherical harmonic lighting: The gritty details. In: Archives of the Game Developers Conference, vol. 56, p. 4.

Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C., 2019. In the wild human pose estimation using explicit 2D features and intermediate 3D representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10905–10914.

Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S., 2018. Viton: An image-based virtual try-on network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7543–7552.

Hassan, M., Choutas, V., Tzionas, D., Black, M.J., 2019. Resolving 3D human pose ambiguities with 3D scene constrains. In: Proceedings IEEE International Conference on Computer Vision, ICCV.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J., 2017. Towards accurate marker-less human shape and pose estimation over time. In: 2017 International Conference on 3D Vision (3DV). IEEE, pp. 421–430.

Huang, Q.-X., Guibas, L., 2013. Consistent shape maps via semidefinite programming. Comput. Graph. Forum 32 (5), 177–186.

Hwang, J., Park, S., Kwak, N., 2017. Athlete pose estimation by a global-local network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 58–65.

Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. 36 (7), 1325–1339.

Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y., 2019. Learnable triangulation of human pose. arXiv preprint arXiv:1905.05754.

Jack, D., Maire, F., Shirazi, S., Eriksson, A., 2019. IGE-Net: Inverse Graphics Energy Networks for Human Pose Estimation and Single-View Reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7075–7084.

Johnson, S., Everingham, M., 2010. Clustered pose and nonlinear appearance models for human pose estimation. Bmvc 2 (4), 5.

Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I., et al., 2017. Panoptic studio: A massively multiview system for social interaction capture. IEEE Trans. Pattern Anal. Mach. Intell. 41 (1), 190–204.

Joo, H., Simon, T., Sheikh, Y., 2018. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8320–8329.

Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J., 2018. End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7122–7131.

Katircioglu, I., Tekin, B., Salzmann, M., Lepetit, V., Fua, P., 2018. Learning latent representations of 3d human pose with deep neural networks. Int. J. Comput. Vis. 126 (12), 1326–1341.

Kim, W., Ramanagopal, M.S., Barto, C., Yu, M.-Y., Rosaen, K., Goumas, N., Vasudevan, R., Johnson-Roberson, M., 2019. Pedx: Benchmark dataset for metric 3-D pose estimation of pedestrians in complex urban intersections. IEEE Robot. Autom. Lett. 4 (2), 1940–1947.

Kocabas, M., Athanasiou, N., Black, M.J., 2020. VIBE: Video inference for human body pose and shape estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5252–5262.

Kocabas, M., Karagoz, S., Akbas, E., 2019. Self-supervised learning of 3d human pose using multi-view geometry. arXiv preprint arXiv:1903.02330.

Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K., 2019a. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2252–2261.

Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K., 2019b. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2252–2261.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105.

Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V., 2017. Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6050–6059.

Lee, K., Lee, I., Lee, S., 2018. Propagating lstm: 3d pose estimation based on joint interdependency. In: Proceedings of the European Conference on Computer Vision, ECCV. pp. 119–135.

Lentzas, A., Vrakas, D., 2019. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: a review. Artif. Intell. Rev. 53, 1975–2021.

Li, S., Chan, A.B., 2014. 3d human pose estimation from monocular images with deep convolutional neural network. In: Asian Conference on Computer Vision. Springer, pp. 332–347.

Li, Y., Huang, C., Loy, C.C., 2019a. Dense intrinsic appearance flow for human pose transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3693–3702.

Li, C., Lee, G.H., 2019. Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9887–9895.

Li, Z., Wang, X., Wang, F., Jiang, P., 2019b. On boosting single-frame 3D human pose estimation via monocular videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2192–2201.

Liang, J., Lin, M., 2019. Shape-aware human pose and shape reconstruction using multi-view images. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV. pp. 4351–4361.

Lin, M., Lin, L., Liang, X., Wang, K., Cheng, H., 2017. Recurrent 3d pose sequence machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 810–819.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014a. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Lin, T.-Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014b. Microsoft COCO: Common objects in context. In: ECCV.

Lin, K., Wang, L., Liu, Z., 2020. End-to-end human pose and mesh reconstruction with transformers. ArXiv, arXiv:abs/2012.09760.

Loper, M., Mahmood, N., Black, M.J., 2014. MoSh: Motion and shape capture from sparse markers. ACM Trans. Graph. 33 (6), 220.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J., 2015. SMPL: A skinned multi-person linear model. ACM Trans. Graph. (TOG) 34 (6), 248.

Luo, Z., Golestaneh, S., Kitani, K.M., 2020. 3D Human motion estimation via motion compression and refinement. ArXiv, arXiv:abs/2008.03789.

Luvizon, D.C., Picard, D., Tabia, H., 2018. 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5137–5146.

Luvizon, D.C., Tabia, H., Picard, D., 2019. Human pose regression by combining indirect part detection and contextual information. Comput. Graph. 85, 15–22.

Mahmood, N., Ghorbani, N., Troje, N., Pons-Moll, G., Black, M.J., 2019. AMASS: Archive of motion capture as surface shapes. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, pp. 5441–5450.

von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G., 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision, ECCV. pp. 601–617.

Marinoiu, E., Zanfir, M., Olaru, V., Sminchisescu, C., 2018. 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2158–2167.

Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017. A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2640–2649.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C., 2017a. Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International Conference on 3D Vision (3DV). IEEE, pp. 506–516.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C., 2018. Single-shot multi-person 3d pose estimation from monocular rgb. In: 2018 International Conference on 3D Vision (3DV). IEEE, pp. 120–130.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., Theobalt, C., 2017b. Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Trans. Graph. 36 (4), 44.

Moon, G., Chang, J.Y., Lee, K.M., 2019. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. arXiv preprint arXiv:1907.11346.

Moon, G., Lee, K.M., 2020. I2l-meshnet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In: ECCV.

Moreno-Noguer, F., 2017. 3d human pose estimation from a single image via distance matrix regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2823–2832.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. Springer, pp. 483–499.

Nibali, A., He, Z., Morgan, S., Prendergast, L., 2018. Numerical coordinate regression with convolutional neural networks. ArXiv, arXiv:abs/1801.07372.

Nibali, A., He, Z., Morgan, S., Prendergast, L., 2019. 3D Human pose estimation with 2D marginal heatmaps. In: 2019 IEEE Winter Conference on Applications of Computer Vision, WACV. pp. 1477–1485.

Nie, B.X., Wei, P., Zhu, S.-C., 2017. Monocular 3d human pose estimation by predicting depth on joints. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 3467–3475.

Nie, X., Zhang, J., Yan, S., Feng, J., 2019. Single-stage multi-person pose machines. arXiv preprint arXiv:1908.09220.

Noroozi, F., Kaminska, D., Corneanu, C., Sapinski, T., Escalera, S., Anbarjafari, G., 2018. Survey on emotional body gesture recognition. IEEE Trans. Affect. Comput..

Novotny, D., Ravi, N., Graham, B., Neverova, N., Vedaldi, A., 2019. C3DPO: Canonical 3D pose networks for non-rigid structure from motion. In: Proceedings of the IEEE International Conference on Computer Vision.

Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B., 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 International Conference on 3D Vision (3DV). IEEE, pp. 484–494.

Park, S., Hwang, J., Kwak, N., 2016. 3D human pose estimation using convolutional neural networks with 2D pose information. In: European Conference on Computer Vision. Springer, pp. 156–169.

Pavlakos, G., Kolotouros, N., Daniilidis, K., 2019. TexturePose: Supervising Human Mesh Estimation with Texture Consistency. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 803–812.

Pavlakos, G., Zhou, X., Daniilidis, K., 2018. Ordinal depth supervision for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7307–7316.

Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017a. Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7025–7034.

Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017b. Harvesting multiple views for marker-less 3d human pose annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6988–6997.

Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K., 2018. Learning to estimate 3D human pose and shape from a single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 459–468.

Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M., 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7753–7762.

Peng, W., Hong, X., Chen, H., Zhao, G., 2020. Learning graph convolutional network for skeleton-based human action recognition by neural searching. ArXiv, arXiv:abs/1911.04131.

Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B., 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4929–4937.

Pons-Moll, G., Baak, A., Gall, J., Leal-Taixe, L., Mueller, M., Seidel, H.-P., Rosenhahn, B., 2011. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In: 2011 International Conference on Computer Vision. IEEE, pp. 1243–1250.

Pons-Moll, G., Pujades, S., Hu, S., Black, M.J., 2017. Clothcap: Seamless 4D clothing capture and retargeting. ACM Trans. Graph. 36 (4), 73.

Popa, A.-I., Zanfir, M., Sminchisescu, C., 2017. Deep multitask architecture for integrated 2d and 3d human sensing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6289–6298.

Popov, S., Bauszat, P., Ferrari, V., 2020. CoReNet: Coherent 3D scene reconstruction from a single RGB image. In: ECCV.

Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W., 2019. Cross View Fusion for 3D Human Pose Estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4342–4351.

Rayat Imtiaz Hossain, M., Little, J.J., 2018. Exploiting temporal information for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 68–84.

Rematas, K., Kemelmacher-Shlizerman, I., Curless, B., Seitz, S., 2018. Soccer on your tabletop. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4738–4747.

Rhodin, H., Constantin, V., Katircioglu, I., Salzmann, M., Fua, P., 2019. Neural scene decomposition for multi-person motion capture. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Rhodin, H., Robertini, N., Casas, D., Richardt, C., Seidel, H.-P., Theobalt, C., 2016. General automatic human shape and motion capture using volumetric contour cues. In: European Conference on Computer Vision. Springer, pp. 509–526.

Rhodin, H., Robertini, N., Richardt, C., Seidel, H.-P., Theobalt, C., 2015. A versatile scene model with differentiable visibility applied to generative pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 765–773.

Rhodin, H., Salzmann, M., Fua, P., 2018. Unsupervised geometry-aware representation for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision, ECCV. pp. 750–767.

Rhodin, H., Spörri, J., Katircioglu, I., Constantin, V., Meyer, F., Müller, E., Salzmann, M., Fua, P., 2018. Learning monocular 3D human pose estimation from multi-view images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8437–8446.

Roetenberg, D., Luinge, H., Slycke, P., 2009. Xsens MVN: Full 6dof human motion tracking using miniature inertial sensors. Tech. Rep. 1, Xsens Motion Technologies BV.

Rogez, G., Weinzaepfel, P., Schmid, C., 2017. LCR-Net: Localization-classification-regression for human pose. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 1216–1224.

Rogez, G., Weinzaepfel, P., Schmid, C., 2019. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. IEEE Trans. Pattern Anal. Mach. Intell..

Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A., 2016. 3d human pose estimation: A review of the literature and analysis of covariates. Comput. Vis. Image Underst. 152, 1–20.

Scott, J., Collins, R., Funk, C., Liu, Y., 2017. 4D model-based spatiotemporal alignment of scripted Taiji Quan sequences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 795–804.

Sengupta, A., Budvytis, I., Cipolla, R., 2020. Synthetic training for accurate 3D human pose and shape estimation in the wild. ArXiv, arXiv:abs/2009.10013.

Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A., 2019. Monocular 3d human pose estimation by generation and ordinal ranking. arXiv preprint arXiv:1904.01324.

Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N., 2018. Deformable gans for pose-based human image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3408–3416.

Sigal, L., Balan, A., Black, M.J., 2009. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. 87, 4–27.

Sigal, L., Balan, A.O., Black, M.J., 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. 87 (1–2), 4.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Song, F.Y.Y.Z.S., Xiao, A.S.J., 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. arXiv preprint arXiv:1506.03365.

Stoll, C., Hasler, N., Gall, J., Seidel, H.-P., Theobalt, C., 2011. Fast articulated motion tracking using a sums of gaussians body model. In: 2011 International Conference on Computer Vision. IEEE, pp. 951–958.

Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q., 2017. Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3960–3969.

Sun, X., Shang, J., Liang, S., Wei, Y., 2017. Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2602–2611.

Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y., 2018. Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 529–545.

Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T., 2019. Human mesh recovery from monocular images via a skeleton-disentangled representation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5348–5357.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: NIPS.

Tai, K.S., Socher, R., Manning, C.D., 2015. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.

Tan, V., Budvytis, I., Cipolla, R., 2018. Indirect deep structured learning for 3d human body shape and pose prediction.

Tekin, B., Márquez-Neila, P., Salzmann, M., Fua, P., 2017. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3941–3950.

Tekin, B., Rozantsev, A., Lepetit, V., Fua, P., 2016. Direct prediction of 3d body poses from motion compensated sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 991–1000.

Tome, D., Russell, C., Agapito, L., 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2500–2509.

Tome, D., Toso, M., Agapito, L., Russell, C., 2018. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In: 2018 International Conference on 3D Vision (3DV). IEEE, pp. 474–483.

Tompson, J.J., Jain, A., LeCun, Y., Bregler, C., 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems. pp. 1799–1807.

Trumble, M., Gilbert, A., Hilton, A., Collomosse, J., 2018. Deep autoencoder for combined human pose estimation and body model upscaling. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 784–800.

Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J., 2017. Total capture: 3D human pose estimation fusing video and inertial sensors.. In: BMVC, vol. 2. p. 3.

Tung, H.-Y.F., Tung, H.-W., Yumer, E., Fragkiadaki, K., 2017. Self-supervised learning of motion capture. In: NIPS.

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C., 2017. Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 109–117.

Wandt, B., Rosenhahn, B., 2019. RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7782–7791.

Wang, J., Huang, S., Wang, X., Tao, D., 2019a. Not all parts are created equal: 3D pose estimation by modelling bi-directional dependencies of body parts. arXiv preprint arXiv:1905.07862.

Wang, C., Kong, C., Lucey, S., 2019b. Distill Knowledge from NRSfM for Weakly Supervised 3D Pose Learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 743–752.

Wang, Z., Shin, D., Fowlkes, C.C., 2020. Predicting camera viewpoint improves cross-dataset generalization for 3D human pose estimation. In: ECCV Workshops.

Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4724–4732.

Wei, X., Zhang, Y., Li, Z., Fu, Y., Xue, X., 2020. Deepsfm: Structure from motion via deep bundle adjustment. In: ECCV.

Weng, C.-Y., Curless, B., Kemelmacher-Shlizerman, I., 2019. Photo wake-up: 3d character animation from a single photo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5908–5917.

Xia, F., Zhu, J., Wang, P., Yuille, A.L., 2016. Pose-guided human parsing by an and/or graph using pose-context features. In: Thirtieth AAAI Conference on Artificial Intelligence.

Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.-P., Theobalt, C., 2018a. Monoperfcap: Human performance capture from monocular video. ACM Trans. Graph. (ToG) 37 (2), 27.

Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W., 2018b. Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2119–2128.

Xu, Y., Zhu, S.-C., Tung, T., 2019. DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7760–7770.

Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X., 2018. 3d human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5255–5264.

Yang, Y., Ramanan, D., 2012. Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. 35 (12), 2878–2890.

Zanfir, A., Marinoiu, E., Sminchisescu, C., 2018. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2148–2157.

Zecha, D., Einfalt, M., Eggert, C., Lienhart, R., 2018. Kinematic Pose Rectification for Performance Analysis and Retrieval in Sports. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1791–1799.

Zhang, Z., 2012. Microsoft kinect sensor and its effect. IEEE multimedia 19 (2), 4–10.

Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N., 2019. Semantic graph convolutional networks for 3D human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3425–3435.

Zheng, L., Huang, Y., Lu, H., Yang, Y., 2019. Pose invariant embedding for deep person re-identification. IEEE Trans. Image Process..

Zhou, K., Han, X., Jiang, N., Jia, K., Lu, J., 2019. HEMlets pose: Learning part-centric heatmap triplets for accurate 3D human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2344–2353.

Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y., 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 398–407.

Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y., 2016a. Deep kinematic pose regression. In: European Conference on Computer Vision. Springer, pp. 186–201.

Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K., 2016b. Sparseness meets deepness: 3D human pose estimation from monocular video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4966–4975.