Computer Vision and Image Understanding 192 (2020) 102897

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Monocular human pose estimation: A survey of deep learning-based methods☆

Yucheng Chen[a], Yingli Tian[b,*], Mingyi He[a]

[a] Northwestern Polytechnical University, Xi'an, 710072, China
[b] The City College, City University of New York, NY 10031, USA

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
41A05
41A10
65D05
65D17

Keywords:
Deep learning
Human pose estimation
Survey

ABSTRACT

Vision-based monocular human pose estimation, as one of the most fundamental and challenging problems in computer vision, aims to obtain posture of the human body from input images or video sequences. The recent developments of deep learning techniques have been brought significant progress and remarkable breakthroughs in the field of human pose estimation. This survey extensively reviews the recent deep learning-based 2D and 3D human pose estimation methods published since 2014. This paper summarizes the challenges, main frameworks, benchmark datasets, evaluation metrics, performance comparison, and discusses some promising future research directions.

1. Introduction

The human pose estimation (HPE) task, which has been developed for decades, aims to obtain posture of the human body from given sensor inputs. Vision-based approaches are often used to provide such a solution by using cameras. In recent years, with deep learning shows good performance on many computer version tasks such as image classification (Krizhevsky et al., 2012), object detection (Ren et al., 2015), semantic segmentation (Long et al., 2015), etc., HPE also achieves rapid progress by employing deep learning technology. The main developments include well-designed networks with great estimation capability, richer datasets (Lin et al., 2014; Joo et al., 2017; Mehta et al., 2017a) for feeding networks and more practical exploration of body models (Loper et al., 2015; Kanazawa et al., 2018). Although there are some existing reviews for HPE, however, there still lacks a survey to summarize the most recent deep learning-based achievements. This paper extensively reviews deep learning-based 2D/3D human pose estimation methods from monocular images or video footage of humans. Algorithms relied on other sensors such as depth (Shotton et al., 2012), infrared light source (Faessler et al., 2014), radio frequency signal (Zhao et al., 2018), and multi-view inputs (Rhodin et al., 2018b) are not included in this survey.

As one of the fundamental computer vision tasks, HPE is a very important research field and can be applied to many applications such as action/activity recognition (Li et al., 2017b; Luvizon et al., 2018; Li et al., 2018b), action detection (Li et al., 2017a), human tracking (Insafutdinov et al., 2017), Movies and animation, Virtual reality, Human–computer interaction, Video surveillance, Medical assistance, Self-driving, Sports motion analysis, etc.

Movies and animation: The generation of various vivid digital characters is inseparable from the capture of human movements. Cheap and accurate human motion capture system can better promote the development of the digital entertainment industry.

Virtual reality: Virtual reality is a very promising technology that can be applied in both education and entertainment. Estimation of human posture can further clarify the relation between human and virtual reality world and enhance the interactive experience.

Human–computer interaction (HCI): HPE is very important for computers and robots to better understand the identification, location, and action of people. With the posture of human (e.g. gesture), computers and robots can execute instructions in an easy way and be more intelligent.

Video surveillance: Video surveillance is one of the early applications to adopt HPE technology in tracking, action recognition, re-identification people within a specific range.

Medical assistance: In the application of medical assistance, HPE can provide physicians with quantitative human motion information especially for rehabilitation training and physical therapy.

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.cviu.2019.102897.

* Corresponding author.
E-mail addresses: chenyucheng@mail.nwpu.edu.cn (Y. Chen), ytian@ccny.cuny.edu (Y. Tian), myhe@nwpu.edu.cn (M. He).

https://doi.org/10.1016/j.cviu.2019.102897
Received 25 August 2019; Received in revised form 21 December 2019; Accepted 28 December 2019
Available online 7 January 2020

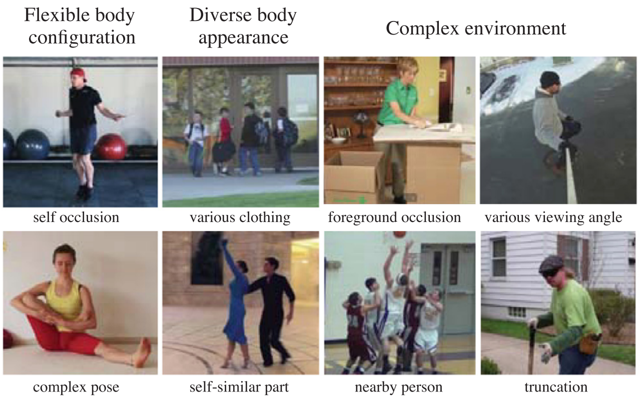

Fig. 1. Typical challenges of HPE in monocular images or videos. Example images are from Max Planck Institute for Informatics (MPII) dataset (Andriluka et al., 2014).

Self-driving: Advanced self-driving has been developed rapidly. With HPE, self-driving cars can respond more appropriately to pedestrians and offer more comprehensive interaction with traffic coordinators.

Sport motion analysis: Estimating players' posture in sport videos can further obtain the statistics of athletes' indicators (e.g. running distance, number of jumps). During training, HPE can provide a quantitative analysis of action details. In physical education, instructors can make more objective evaluations of students with HPE.

Monocular human pose estimation has some unique characteristics and challenges. As shown in Fig. 1, the challenges of human pose estimation mainly fall in three aspects:

- Flexible body configuration indicates complex interdependent joints and high degree-of-freedom limbs, which may cause self-occlusions or rare/complex poses.
- Diverse body appearance includes different clothing and self-similar parts.
- Complex environment may cause foreground occlusion, occlusion or similar parts from nearby persons, various viewing angles, and truncation in the camera view.

The papers of human pose estimation can be categorized in different ways. Based on whether to use designed human body models or not, the methods can be categorized into generative methods (model-based) and discriminative methods (model-free). According to from which level (high-level abstraction or low-level pixel evidence) to start the processing, they can be classified into top-down methods and bottom-up methods. More details of different category strategies for HPE approaches are summarized in Table 2 and described in Section 2.1.

As listed in Table 1, with the development of human pose estimation in the past decades, several notable surveys summarized the research work in this area. The surveys (Aggarwal and Cai, 1999; Gavrila, 1999; Poppe, 2007; Ji and Liu, 2010; Moeslund et al., 2011) reviewed the early work of human motion analysis in many aspects (e.g., detection and tracking, pose estimation, recognition) and described the relation between human pose estimation and other related tasks. While Hu et al. (2004) summarized the research of human motion analysis for video surveillance application, the reviews (Moeslund and Granum, 2001; Moeslund et al., 2006) focused on the human motion capture systems. More recent surveys were mainly focusing on relatively narrow directions, such as RGB-D-based action recognition (Chen et al., 2013; Wang et al., 2018b), 3D HPE (Sminchisescu, 2008; Holte et al., 2012; Sarafianos et al., 2016), model-based HPE (Holte et al., 2012; Perez-Sala et al., 2014), body parts-based HPE (Liu et al., 2015), and monocular-based HPE (Sminchisescu, 2008; Gong et al., 2016).

Different from existing review papers, this survey extensively summarizes the recent milestone work of deep learning-based human pose estimation methods, which were mainly published from 2014. In order to provide a comprehensive summary, this survey includes a few research work which has been discussed in some surveys (Liu et al., 2015; Gong et al., 2016; Sarafianos et al., 2016), but most of the recent advances are not been presented in any survey before.

The remainder of this paper is organized as follows. Section 2 introduces the existing review papers for human motion analysis and HPE, different ways to category HPE methods, and the widely used human body models. Sections 3 and 4 describe 2D HPE and 3D HPE approaches respectively. In each section, we further describe HPE approaches for both single person pose estimation and multi-person pose estimation. Since data are a very important and fundamental element for deep learning-based methods, the recent HPE datasets and the evaluation metrics are summarized in Section 5. Finally, Section 6 concludes the paper and discusses several promising future research directions.

2. Categories of HPE methods and human body models

2.1. HPE method categories

This section summarizes the different categories of deep learning-based HPE methods based on different characteristics: (1) generative (human body model-based) and discriminative (human body model-free); (2) top-down (from high-level abstraction to low-level pixel evidence) and bottom-up (from low-level pixel evidence to high-level abstraction); (3) regression-based (directly mapping from input images to body joint positions) and detection-based (generating intermediate image patches or heatmaps of joint locations); and (4) one-stage (end-to-end training) and multi-stage (stage-by-stage training).

Generative vs. Discriminative: The main difference between generative and discriminative methods is whether a method uses human body models or not. Based on the different representations of human body models, generative methods can be processed in different ways such as prior beliefs about the structure of the body model, geometrically projection from different views to 2D or 3D space, high-dimensional parametric space optimization in regression manners. More details of human body model representation can be found in Section 2.2. Discriminative methods directly learn a mapping from input sources to human pose space (learning-based) or search in existing examples (example-based) without using human body models. Discriminative methods are usually faster than generative methods but may have less robustness for poses never trained with.

Top-down vs. Bottom-up: For multi-person pose estimation, HPE methods can generally be classified as top-down and bottom-up methods according to the starting point of the prediction: high-level abstraction or low-level pixel evidence. Top-down methods start from high-level abstraction to first detect persons and generate the person locations in bounding boxes. Then pose estimation is conducted for each person. In contrast, bottom-up methods first predict all body parts of every person in the input image and then group them either by human body model fitting or other algorithms. Note that body parts could be joints, limbs, or small template patches depending on different methods. With an increased number of people in an image, the computation cost of top-down methods significantly increases, while keeps stable for bottom-up methods. However, if there are some people with a large overlap, bottom-up methods face challenges to group corresponding body parts.

Regression-based vs. Detection-based: Based on the different problem formulations, deep learning-based human pose estimation methods can be split into regression-based or detection-based methods. The regression-based methods directly map the input image to the coordinates of body joints or the parameters of human body models. The detection-based methods treat the body parts as detection targets based on two widely used representations: image patches and heatmaps of joint locations. Direct mapping from images to joint coordinates

Table 1
Summary of the related surveys of human motion analysis and HPE.

No.	Survey & Reference	Venue	Content
1	Human motion analysis: A review (Aggarwal and Cai, 1999)	CVIU	A review of human motion analysis including body structure analysis, motion tracking and action recognition.
2	The visual analysis of human movement: A survey (Gavrila, 1999)	CVIU	A survey of whole-body and hand motion analysis.
3	A survey of computer vision-based human motion capture (Moeslund and Granum, 2001)	CVIU	An overview based on motion capture system, including initialization, tracking, pose estimation, and recognition.
4	A survey on visual surveillance of object motion and behaviors (Hu et al., 2004)	TSMCS	A summary of human motion analysis based one the framework of visual surveillance in dynamic scenes.
5	A survey of advances in vision-based human motion capture and analysis (Moeslund et al., 2006)	CVIU	Further summary of human motion capture and analysis from 2000 to 2006, following (Moeslund and Granum, 2001).
6	Vision-based human motion analysis: An overview (Poppe, 2007)	CVIU	A summary of vision-based human motion analysis with markerless data.
7	3D human motion analysis in monocular video: techniques and challenges (Sminchisescu, 2008)	Book Chapter	An overview of reconstructing 3D human motion with video sequences from single-view camera.
8	Advances in view-invariant human motion analysis: A review (Ji and Liu, 2010)	TSMCS	A summary of human motion analysis, including human detection, view-invariant pose representation and estimation, and behavior understanding.
9	Visual analysis of humans (Moeslund et al., 2011)	Book	A comprehensive overview of human analysis, including detection and tracking, pose estimation, recognition, and applications with human body and face.
10	Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments (Holte et al., 2012)	JSTSP	A review of model-based 3D HPE and action recognition methods under multi-view.
11	A survey of human motion analysis using depth imagery (Chen et al., 2013)	PRL	A survey of traditional RGB-D-based human action recognition methods, including description of sensors, corresponding datasets, and approaches.
12	A survey on model based approaches for 2D and 3D visual human pose recovery (Perez-Sala et al., 2014)	Sensors	A survey of model-based approaches for HPE, grouped in five main modules: appearance, viewpoint, spatial relations, temporal consistence, and behavior.
13	A survey of human pose estimation: the body parts parsing based methods (Liu et al., 2015)	JVCIR	A survey of body parts parsing-based HPE methods under both single-view and multiple-view from different input sources (images, videos, depth).
14	Human pose estimation from monocular images: A comprehensive survey (Gong et al., 2016)	Sensors	A survey of monocular-based traditional HPE methods with a few deep learning-based methods.
15	3d human pose estimation: A review of the literature and analysis of covariates (Sarafianos et al., 2016)	CVIU	A review of 3D HPE methods with different type of inputs (e.g., single image or video, monocular or multi-view).
16	RGB-D-based human motion recognition with deep learning: A survey (Wang et al., 2018b)	CVIU	A survey of RGB-D-based motion recognition in four categories: RGB-based, depth-based, skeleton-based, and RGB-D-based.
17	**Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods**	**Ours**	**A comprehensive survey of deep learning-based monocular HPE research and human pose datasets, organized into four groups: 2D single HPE, 2D multi-HPE, 3D single HPE and 3D multi-HPE**

is very difficult since it is a highly nonlinear problem, while small-region representation provides dense pixel information with stronger robustness. Compared to the original image size, the detected results of small-region representation limit the accuracy of the final joint coordinates.

One-stage vs. Multi-stage: The deep learning-based one-stage methods aim to map the input image to human poses by employing end-to-end networks, while multi-stage methods usually predict human pose in multiple stages and are accompanied by intermediate supervision. For example, some multi-person pose estimation methods first detect the locations of people and then estimate the human pose for each detected person. Other 3D human pose estimation methods first predict joint locations in the 2D surface, then extend them to 3D space. The training of one-stage methods is easier than multi-stage methods, but with less intermediate constraints.

This survey reviews the recent work in two main sections: 2D human pose estimation (Section 3) and 3D human pose estimation (Section 4). For each section, we further divide them into subsections based on their respective characteristics (see a summary of all the categories and the corresponding papers in Table 2.)

2.2. Human body models

Human body modeling is a key component of HPE. Human body is a flexible and complex non-rigid object and has many specific characteristics like kinematic structure, body shape, surface texture, the position of body parts or body joints, etc. A mature model for human body is not necessary to contain all human body attributes but should satisfy the requirements for specific tasks to build and describe human body pose. Based on different levels of representations and application scenarios, as shown in Fig. 2, there are three types of commonly used human body models in HPE: skeleton-based model, contour-based model, and volume-based model. For more detailed descriptions of human body models, we refer interested readers to two well-summarized papers (Liu et al., 2015; Gong et al., 2016).

Skeleton-based Model: Skeleton-based model, also known as stick-figure or kinematic model, represents a set of joint (typically between 10 to 30) locations and the corresponding limb orientations following the human body skeletal structure. The skeleton-based model can also be described as a graph where vertices indicating joints and edges encoding constraints or prior connections of joints within the skeleton structure (Felzenszwalb and Huttenlocher, 2005). This human body topology is very simple and flexible which is widely utilized in both 2D and 3D HPE (Cao et al., 2017; Mehta et al., 2017b) and human pose datasets (Andriluka et al., 2014; Wu et al., 2017). With obvious advantages of simple and flexible representing, it also has many shortcomings such as lacking texture information which indicates there is no width and contour information of human body.

Contour-based Model: The contour-based model is widely used in earlier HPE methods which contains the rough width and contour information of body limbs and torso. Human body parts are approximately represented with rectangles or boundaries of person silhouette. Widely

Table 2
The categories of deep learning-based monocular human pose estimation.

Direction	Sub-direction	Categories	Sub-categories
2D HPE	2D Single	Regression-based	**(1) Direct prediction**: Krizhevsky et al. (2012), on video (Pfister et al., 2014) **(2) Supervision improvement**: transform heatmaps to joint coordinates (Luvizon et al., 2019; Nibali et al., 2018), recursive refinement (Carreira et al., 2016), bone-based constraint (Sun et al., 2017) **(3) Multi-task**: with body part detection (Li et al., 2014), with person detection and action classification (Gkioxari et al., 2014a), with heatmap-based joint detection (Fan et al., 2015), with action recognition on video sequences (Luvizon et al., 2018)
		Detection-based	**(1) Patch-based**: Jain et al. (2013), Chen and Yuille (2014) and Ramakrishna et al. (2014) **(2) Network design**: Tompson et al. (2015), Bulat and Tzimiropoulos (2016) and Xiao et al. (2018), multi-scale inputs (Rafi et al., 2016), heatmap-based improvement (Papandreou et al., 2017), Hourglass (Newell et al., 2016), CPM (Wei et al., 2016), PRM (Yang et al., 2017), feed forward module (Belagiannis and Zisserman, 2017), HRNet (Sun et al., 2019), GAN (Chou et al., 2018; Chen et al., 2017; Peng et al., 2018) **(3) Body structure constraint**: Tompson et al. (2014), Lifshitz et al. (2016), Yang et al. (2016), Gkioxari et al. (2016), Chu et al. (2016, 2017), Ning et al. (2018), Ke et al. (2018), Tang et al. (2018a) and Tang and Wu (2019) **(4) Temporal constraint**: Jain et al. (2014), Pfister et al. (2015) and Luo et al. (2018) **(5) Network compression**: Tang et al. (2018b), Debnath et al. (2018) and Feng et al. (2019)
	2D Multiple	Top-down	Coarse-to-fine (Iqbal and Gall, 2016; Huang et al., 2017), bounding box refinement (Fang et al., 2017), multi-level feature fusion (Xiao et al., 2018; Chen et al., 2018), results refinement (Moon et al., 2019)
		Bottom-up	**(1) Two-stage**: DeepCut (Pishchulin et al., 2016), DeeperCut (Insafutdinov et al., 2016), OpenPose (Cao et al., 2017), PPN (Nie et al., 2018), PifPafNet (Kreiss et al., 2019) **(2) Single-stage**: heatmaps and associative embedding maps (Newell et al., 2017) **(3) Multi-task**: instance segmentation (Papandreou et al., 2018), keypoint detection and semantic segmentation (Kocabas et al., 2018)
3D HPE	3D Single	Model-free	**(1) Single-stage**: direct prediction (Li and Chan, 2014; Pavlakos et al., 2017), body structure constraint (Li et al., 2015b; Tekin et al., 2016; Sun et al., 2017; Pavlakos et al., 2018a) **(2) 2D-to-3D**: Martinez et al. (2017), Zhou et al. (2017), Tekin et al. (2017), Li and Lee (2019), Qammaz and Argyros (2019), Chen and Ramanan (2017), Moreno-Noguer (2017), Wang et al. (2018a) and Yang et al. (2018)
		Model-based	**(1) SMPL-based**: Bogo et al. (2016), Tan et al. (2017), Pavlakos et al. (2018b), Omran et al. (2018), Varol et al. (2018), Kanazawa et al. (2018) and Arnab et al. (2019) **(2) Kinematic model-based**: Mehta et al. (2017a), Nie et al. (2017), Zhou et al. (2016), Mehta et al. (2017b) and Rhodin et al. (2018a) **(3) Other model-based**: probabilistic model (Tome et al., 2017)
	3D Multiple		Bottom-up (Mehta et al., 2018), top-down (Rogez et al., 2017), SMPL-based (Zanfir et al., 2018), real-time (Mehta et al., 2019)

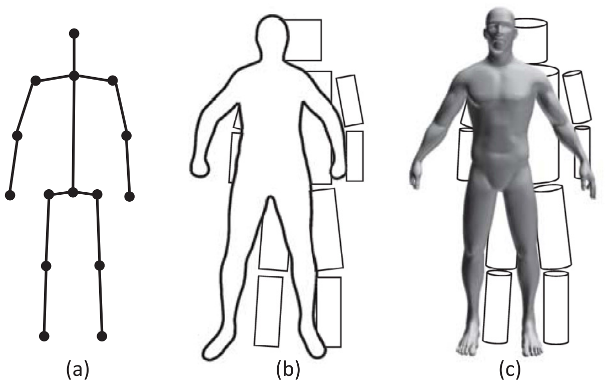

Fig. 2. Commonly used human body models. (a) skeleton-based model; (b) contour-based models; (c) volume-based models.

used contour-based models include cardboard models (Ju et al., 1996) and Active Shape Models (ASMs) (Cootes et al., 1995).

Volume-based Model: 3D human body shapes and poses are generally represented by volume-based models with geometric shapes or meshes. Earlier geometric shapes for modeling body parts include cylinders, conics, etc. (Sidenbladh et al., 2000). Modern volume-based models are represented in mesh form, normally captured with 3D scans. Widely used volume-based models includes Shape Completion and Animation of People (SCAPE) (Anguelov et al., 2005), Skinned Multi-Person Linear model (SMPL) (Loper et al., 2015), and a unified deformation model (Joo et al., 2018).

3. 2D human pose estimation

2D human pose estimation calculates the locations of human joints from monocular images or videos. Before deep learning brings a huge impact on vision-based human pose estimation, traditional 2D HPE algorithms adopt hand-craft feature extraction and sophisticated body models to obtain local representations and global pose structures (Dantone et al., 2013; Chen and Yuille, 2014; Gkioxari et al., 2014b). Here, the recent deep learning-based 2D human pose estimation methods are categorized into "single person pose estimation" and "multi-person pose estimation".

3.1. 2D single person pose estimation

2D single person pose estimation is to localize body joint positions of a single person in an input image. For images with more persons, pre-processing is needed to crop the original image so that there is only one person in the input image such as using an upper-body detector (Eichner and Ferrari, 2012a) or full-body detector (Ren et al., 2015), and cropping from original images based on the annotated person center and body scale (Andriluka et al., 2014; Newell et al., 2016). Early work of introducing deep learning into human pose estimation mainly extended traditional HPE methods by simply replaced some components of frameworks by neural networks (Jain et al., 2013; Ouyang et al., 2014).

Based on the different formulations of human pose estimation task, the proposed methods using CNNs can be classified into two categories: regression-based methods and detection-based methods. Regression-based methods attempt to learn a mapping from image to kinematic body joint coordinates by an end-to-end framework and

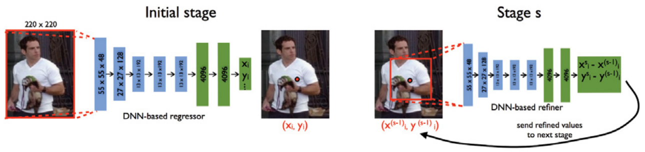

Fig. 3. The framework of DeepPose (Toshev and Szegedy, 2014).

Fig. 4. Heatmap representation of different joints.

generally directly produce joint coordinates (Toshev and Szegedy, 2014). Detection-based methods are intended to predict approximate locations of body parts (Chen and Yuille, 2014) or joints (Newell et al., 2016), usually are supervised by a sequence of rectangular windows (each including a specific body part) (Jain et al., 2013; Chen and Yuille, 2014) or heatmaps (each indicating one joint position by a 2D Gaussian distribution centered at the joint location) (Newell et al., 2016; Wei et al., 2016). Each of these two kinds of methods has its advantages and disadvantages. Direct regression learning of only one single point is a difficulty since it is a highly nonlinear problem and lacks robustness, while heatmap learning is supervised by dense pixel information which results in better robustness. Compared to the original image size, heatmap representation has much lower resolution due to the pooling operation in CNNs, which limits the accuracy of joint coordinate estimation. And obtaining joint coordinates from heatmap is normally a non-differentiable process that blocks the network to be trained end-to-end. The recent representative work for 2D single person pose estimation are summarized in Table 3, the last column is the comparisons of PCKh@0.5 scores on the MPII testing set. More details of datasets and evaluation metrics are described in Section 5.

3.1.1. Regression-based methods

AlexNet (Krizhevsky et al., 2012) was one of the early networks for deep learning-based HPE methods due to its simple architecture and impressive performance. Toshev and Szegedy (2014) firstly attempted to train an AlexNet-like deep neural network to learn joint coordinates from full images in a very straightforward manner without using any body model or part detectors as shown in Fig. 3. Moreover, a cascade architecture of multi-stage refining regressors is employed to refine the cropped images from the previous stage and show improved performance. Pfister et al. (2014) also applied an AlexNet-like network using a sequence of concatenated frames as input to predict the human pose in the videos.

Only using joints without the surrounding information lacks robustness. Converting heatmap supervision to numerical joint positions supervision can retain the advantages of both representations. Luvizon et al. (2019) proposed a Soft-argmax function to transform heatmaps to joint coordinates which can convert a detection-based network to a differentiable regression-based one. Nibali et al. (2018) designed a differentiable spatial to numerical transform (DSNT) layer to calculate joint coordinates from heatmaps, which worked well with low-resolution heatmaps.

Prediction of joint coordinates directly from input images with few constrains is very hard, therefore more powerful networks were introduced with a refinement or body model structure. Carreira et al. (2016) proposed an Iterative Error Feedback network based on GoogleNet which recursively processes the combination of the input image and output results. The final pose is improved from an initial mean pose after iterations. Sun et al. (2017) proposed a structure-aware regression approach based on a ResNet-50. Instead of using joints to represent pose, a bone-based representation is designed by involving body structure information to achieve more stable results than only using joint positions. The bone-based representation also works on 3D HPE.

Networks handling multiple closely related tasks of human body may learn diverse features to improve the prediction of joint coordinates. Li et al. (2014) employed an AlexNet-like multi-task framework to handle the joint coordinate prediction task from full images in a regression way, and the body part detection task from image patches obtained by a sliding-window. Gkioxari et al. (2014a) used a R-CNN architecture to synchronously detect person, estimate pose, and classify action. Fan et al. (2015) proposed a dual-source deep CNNs which take image patches and full images as inputs and output heatmap represented joint detection results of sliding windows together with coordinate represented joint localization results. The final estimated posture is obtained from the combination of the two results. Luvizon et al. (2018) designed a network that can jointly handle 2D/3D pose estimation and action recognition from video sequences. The pose estimated in the middle of the network can be used as a reference for action recognition.

3.1.2. Detection-based methods

Detection-based methods are developed from body part detection methods. In traditional part-based HPE methods, body parts are first detected from image patch candidates and then are assembled to fit a human body model. The detected body parts in early work are relatively big and generally represented by rectangular sliding windows or patches. We refer to Poppe (2007) and Gong et al. (2016) for a more detailed introduction. Some early methods use neural networks as body part detectors to distinguish whether a candidate patch is a specific body part (Jain et al., 2013), classify a candidate patch among predefined templates (Chen and Yuille, 2014) or predict the confidence map belonging to multiple classes (Ramakrishna et al., 2014). Body part detection methods are usually sensitive to complexity background and body occlusions. Therefore the independent image patches with only local appearance may not be sufficiently discriminative for body part detection.

In order to provide more supervision information than just joint coordinates and to facilitate the training of CNNs, more recent work employed heatmap to indicate the ground truth of the joint location (Tompson et al., 2014; Jain et al., 2014). As shown in Fig. 4, each joint occupies a heatmap channel with a 2D Gaussian distribution centered at the target joint location. Moreover, Papandreou et al. (2017) proposed an improved representation of the joint location, which is a combination of binary activation heatmap and corresponding offset. Since heatmap representation is more robust than coordinate representation, most of the recent research is based on heatmap representation.

The neural network architecture is very important to make better use of input information. Some approaches are mainly based on classic networks with appropriate improvements, such as GoogLeNet-based network with multi-scale inputs (Rafi et al., 2016), ResNet-based network with deconvolutional layers (Xiao et al., 2018). In terms of iterative refinement, some work designed networks in a multi-stage style to refine results from coarse prediction via end-to-end learning (Tompson et al., 2015; Bulat and Tzimiropoulos, 2016; Newell et al., 2016; Wei et al., 2016; Yang et al., 2017; Belagiannis and Zisserman, 2017). Such networks generally use intermediate supervision to address vanishing gradients. Newell et al. (2016) proposed a novel *stacked hourglass* architecture by using a residual module as the component unit. Wei et al. (2016) proposed a multi-stage prediction framework with input image for each stage. Yang et al. (2017) designed a Pyramid Residual Module (PRMs) to replace the residual module of the Hourglass network to enhance the invariance across scales of DCNNs by learning features on various scales. Belagiannis and Zisserman (2017) combined

Table 3
Summary of 2D single person pose estimation methods. Note that the last column shows the PCKh@0.5 scores on the Max Planck Institute for Informatics (MPII) human pose testing set.

Methods	Backbone	Input size	Highlights	PCKh (%)
Regression-based				
Toshev and Szegedy (2014)	AlexNet	220 × 220	Direct regression, multi-stage refinement	–
Carreira et al. (2016)	GoogleNet	224 × 224	Iterative error feedback refinement from initial pose.	81.3
Sun et al. (2017)	ResNet-50	224 × 224	Bone based representation as additional constraint, general for both 2D/3D HPE	86.4
Luvizon et al. (2019)	Inception-v4+Hourglass	256 × 256	Multi-stage architecture, proposed soft-argmax function to convert heatmaps into joint locations	91.2
Detection-based				
Tompson et al. (2014)	AlexNet	320 × 240	Heatmap representation, multi-scale input, MRF-like Spatial-Model	79.6
Yang et al. (2016)	VGG	112 × 112	Jointly learning DCNNs with deformable mixture of parts models	–
Newell et al. (2016)	Hourglass	256 × 256	Proposed stacked Hourglass architecture with intermediate supervision.	90.9
Wei et al. (2016)	CPM	368 × 368	Proposed Convolutional Pose Machines (CPM) with intermediate input and supervision, learn spatial correlations among body parts	88.5
Chu et al. (2017)	Hourglass	256 × 256	Multi-resolution attention maps from multi-scale features, proposed micro hourglass residual units to increase the receptive field	91.5
Yang et al. (2017)	Hourglass	256 × 256	Proposed Pyramid Residual Module (PRM) learns filters for input features with different resolutions	92.0
Chen et al. (2017)	conv-deconv	256 × 256	GAN, stacked conv-deconv architecture, multi-task for pose and occlusion, two discriminators for distinguishing whether the pose is 'real' and the confidence is strong	91.9
Peng et al. (2018)	Hourglass	256 × 256	GAN, proposed augmentation network to generate data augmentations without looking for more data	91.5
Ke et al. (2018)	Hourglass	256 × 256	Improved Hourglass network with multi-scale intermediate supervision, multi-scale feature combination, structure-aware loss and data augmentation of joints masking	92.1
Tang et al. (2018a)	Hourglass	256 × 256	Compositional model, hierarchical representation of body parts for intermediate supervision	92.3
Sun et al. (2019)	HRNet	256 × 256	High-resolution representations of features across the whole network, multi-scale fusion.	92.3
Tang and Wu (2019)	Hourglass	256 × 256	Data-driven joint grouping, proposed part-based branching network (PBN) to learn representations specific to each part group.	92.7

a 7 layers feedforward module with a recurrent module to iteratively refine the results. This model learns to predict location heatmaps for both joints and body limbs. Also, they analyzed keypoint visibility with unbalanced ground truth distribution. To keep high-resolution representations of features across the whole network, Sun et al. (2019) proposed a novel High-Resolution Net (HRNet) with multi-scale feature fusion.

Different from earlier work which attempted to fit detected body parts into body models, some recent work tried to encode human body structure information into networks. Tompson et al. (2014) jointly trained a network with a MRF-like spatial-model for learning typical spatial relations between joints. Lifshitz et al. (2016) discretized an image into log-polar bins centered around each joint and employed a VGG-based network to predict joint category confident for each pair-wise joints (binary terms). With all relative confident scores, the final heatmap for each joint can be generated by a deconvolutional network. Yang et al. (2016) designed a two-stage network. Stage one is a convolutional neural network to predict joint locations in heatmap representation. Stage two is a message-passing model connected manually according to the human body structure to find optimal joint locations with a max-sum algorithm. Gkioxari et al. (2016) proposed a convolutional Recurrent Neural Network to output joint location one by one following a chain model. The output of each step depends on both the input image and the previously predicted output. The network can handle both images and videos with different connection strategy. Chu et al. (2016) proposed to transform kernels by a bi-directional tree to pass information between corresponding joints in a tree body model. Chu et al. (2017) replaced the residual modules of the Hourglass network with more sophisticated ones. The Conditional Random Field (CRF) is utilized for attention maps as intermediate supervisions for learning body structure information. Ning et al. (2018) designed a fractal network to impose body prior knowledge to guide the network. The external knowledge visual features are encoded into the basic network by using a learned projection matrix. Ke et al. (2018) proposed a multi-scale structure-aware network based on Hourglass network with multi-scale supervision, multi-scale feature combination, structure-aware loss, and data augmentation of joints masking. On the basic framework of Hourglass network, Tang et al. (2018a) designed a hierarchical representation of body parts for intermediate supervision to replace heatmap for each joint. Thus the network learns the bottom-up/top-down body structure, rather than only scattered joints. Tang and Wu (2019) proposed a part-based branching network (PBN) to learn specific representations of each part group rather than predict all joint heatmaps from one branch. The data-driven part groups are then split by calculating mutual information of joints.

Generative Adversarial Networks (GANs) are also employed to provide adversarial supervision for learning body structure or network training. Chou et al. (2018) introduced adversarial learning with two same Hourglass networks as generator and discriminator respectively. The generator predicts heatmap location of each joint, while the discriminator distinguishes ground truth heatmaps from generated heatmaps. Chen et al. (2017) proposed a structure-aware convolutional network with one generator and two discriminators to incorporate priors of human body structure. The generator is designed from the Hourglass network to predict joint heatmaps as well as occlusion heatmaps. The pose discriminator can discriminate against reasonable body configuration from unreasonable body configuration. The confidence discriminator shows the confidence score of predictions. Peng et al. (2018) studied how to jointly optimize data augmentation and network training without looking for more data. Instead of using random data augmentation, they applied augmentations to increase the network loss while the pose network learns from the generated augmentations.

Utilization of temporal information is also very important to estimate 2D human poses in monocular video sequences. Jain et al. (2014) designed a framework contains two-branch CNNs taking multi-scale RGB frames and optical-flow maps as inputs. The extracted features are concatenated before the last convolutional layers. Pfister et al. (2015) used optical-flow maps as a guide to align predicted heatmaps from neighboring frames based on the temporal context of videos. Luo et al. (2018) exploited temporal information with a Recurrent Neural Network redesigned from CPM by changing multi-stage architecture with LSTM structure.

In order to estimate human poses on low-capacity devices, network parameters can be reduced while still maintaining competitive performance. Tang et al. (2018b) committed to improving the network structure by proposing a densely connected U-Nets and efficient usage of memory. This network is similar to the idea of the Hourglass network while utilizing U-Net as each component with a more optimized global connection across each stage resulting in fewer parameters and small model size. Debnath et al. (2018) adapted MobileNets (Howard et al., 2017) for pose estimation by designing a split stream architecture at the final two layers of the MobileNets. Feng et al. (2019) designed a lightweight variant of Hourglass network and trained it with a full teacher Hourglass network by a Fast Pose Distillation (FPD) training strategy.

In summary, the heatmap representation is more suitable for network training than coordinate representation from detection-based methods in deep learning-based 2D single person pose estimation.

3.2. 2D multi-person pose estimation

Different from single person pose estimation, multi-person pose estimation needs to handle both detection and localization tasks since there is no prompt of how many persons in the input images. According to from which level (high-level abstraction or low-level pixel evidence) to start the calculation, human pose estimation methods can be classified into top-down methods and bottom-up methods.

Top-down methods generally employ person detectors to obtain a set of the bounding box of people in the input image and then directly leverage existing single-person pose estimators to predict human poses. The predicted poses heavily depend on the precision of the person detection. The runtime for the whole system is proportional based on the number of persons. While bottom-up methods directly predict all the 2D joints of all persons and then assemble them into independent skeletons. Correct grouping of joint points in a complex environment is a challenging research task. Table 4 summarizes recent deep learning-based work about 2D multi-person pose estimation methods in both top-down and bottom-up categories. The last column of Table 4 is the Average Precision (AP) scores on the COCO test-dev dataset. More details of datasets and evaluation metrics are described in Section 5.

3.2.1. Top-down methods

The two most important components of top-down HPE methods are human body region candidate detector and a single person pose estimator. Most of the research focused on human part estimation based on existing human detectors such as Faster R-CNN (Ren et al., 2015), Mask R-CNN (He et al., 2017), FPN (Lin et al., 2017). Iqbal and Gall (2016) utilized a convolutional pose machine-based pose estimator to generate initial poses. Then integer linear programming (ILP) is applied to obtain the final poses. Fang et al. (2017) adopted spatial transformer network (STN) (Jaderberg et al., 2015), Non-Maximum-Suppression (NMS), and Hourglass network to facilitate pose estimation in the presence of inaccurate human bounding boxes. Huang et al. (2017) developed a coarse-fine network (CFN) with Inception-v2 network (Szegedy et al., 2016) as the backbone. The network is supervised in multiple levels for learning coarse and fine prediction. Xiao et al. (2018) added several deconvolutional layers over the last convolution layer of ResNet to generate heatmaps from deep and low-resolution features. Chen et al. (2018) proposed a cascade pyramid network (CPN) by employing multi-scale feature maps from different layers to obtain more inference from local and global features with an online hard keypoint mining loss for difficulty joints. Based on similar pose error distributions of different HPE approaches, Moon et al. (2019) designed PoseFix net to refine estimated poses from any methods.

Top-down HPE methods can be easily implemented by combining existing detection networks and single HPE networks. Meanwhile, the performance of this kind of methods is affected by person detection results and the operation speed is usually not real-time.

3.2.2. Bottom-up methods

The main components of bottom-up HPE methods include body joint detection and joint candidate grouping. Most algorithms handle these two components separately. DeepCut (Pishchulin et al., 2016) employed a Fast R-CNN based body part detector to first detect all the body part candidates, then labeled each part to its corresponding part category, and assembled these parts with integer linear programming to a complete skeleton. DeeperCut (Insafutdinov et al., 2016) improved the DeepCut by using a stronger part detector based on ResNet and a better incremental optimization strategy exploring geometric and appearance constraints among joint candidates. OpenPose (Cao et al., 2017) used CPM to predict candidates of all body joints with Part Affinity Fields (PAFs). The proposed PAFs can encode locations and orientations of limbs to assemble the estimated joints into different poses of persons. Nie et al. (2018) proposed a Pose Partition Network (PPN) to conduct both joint detection and dense regression for joint partition. Then PPN performs local inference for joint configurations with joint partition. Similar to OpenPose, Kreiss et al. (2019) designed a PifPaf net to predict a Part Intensity Field (PIF) and a Part Association Field (PAF) to represent body joint locations and body joint association. It works well on low-resolution images due to the fine-grained PAF and the utilization of Laplace loss.

The above methods are all following a separation of joint detection and joint grouping. Recently, some methods can do the prediction in one stage. Newell et al. (2017) introduced a single-stage deep network architecture to simultaneously perform both detection and grouping. This network can produce detection heatmaps for each joint, and associative embedding maps that contain the grouping tags of each joint.

Some methods employed multi-task structures. Papandreou et al. (2018) proposed a box-free multi-task network for pose estimation and instance segmentation. The ResNet-based network can synchronously predict joint heatmaps of all keypoints for every person and their relative displacements. Then the grouping starts from the most confident detection with a greedy decoding process based on a tree-structured kinematic graph. The network proposed by Kocabas et al. (2018) combines a multi-task model with a novel assignment method to handle human keypoint estimation, detection, and semantic segmentation tasks altogether. Its backbone network is a combination of ResNet and FPN with shared features for keypoints and person detection subnets. The human detection results are used as constraints of the spatial position of people.

Currently, the processing speed of bottom-up methods is very fast, and some (Cao et al., 2017; Nie et al., 2018) can run in real-time. However, the performance can be very influenced by the complex background and human occlusions. The top-down approaches achieved state-of-the-art performance in almost all benchmark datasets while the processing speed is limited by the number of detected people.

4. 3D human pose estimation

3D human pose estimation is to predict locations of body joints in 3D space from images or other input sources. Although commercial products such as Kinect (2019) with depth sensor, Vicon (2019) with optical sensor and TheCaptury (2019) with multiple cameras have been

Table 4
Comparison of 2D multi-person pose estimation methods. Note that the last column shows the Average Precision (AP) scores on the COCO test-dev set.

Methods	Network type	Highlights	AP score (%)
Top-down			
Iqbal and Gall (2016)	Faster R-CNN + CPM	After person detection and single HPE, refines detected local joint candidates with Integer Linear Programming (ILP).	–
Fang et al. (2017)	Faster R-CNN + Hourglass	Combines symmetric spatial transformer network (SSTN) and Hourglass model to do SPPE on detected results; proposes a parametric pose NMS for refining pose proposals; designs a pose-guided proposals generator to augment the existing training samples	63.3[a]
Papandreou et al. (2017)	Faster R-CNN + ResNet-101	Produces heatmap and offset map of each joint for SPPE and combines them with an aggregation procedure; uses keypoint-based NMS to avoid duplicate poses	64.9[a]
Huang et al. (2017)	Faster R-CNN + Inception-v2	Produces coarse and fine poses for SPPE with multi-level supervisions; multi-scale features fusion	72.2[a]
He et al. (2017)	Mask R-CNN + ResNet-FPN	An extension of Mask R-CNN framework; predicts keypoints and human mask synchronously	63.1[a]
Xiao et al. (2018)	Faster R-CNN + ResNet	Simply adds a few deconvolutional layers after ResNet to generate heatmaps from deep and low resolution features	73.7
Chen et al. (2018)	FPN + CPN	Proposes CPN with feature pyramid; two-stage network; online hard keypoints mining	73.0
Moon et al. (2019)	ResNet + upsampling	Proposes PoseFix net to refine estimated pose from any HPE methods based on pose error distributions	–
Sun et al. (2019)	Faster R-CNN + HRNet	High-resolution representations of features across the whole network, multi-scale fusion	75.5
Bottom-up			
Pishchulin et al. (2016)	Fast R-CNN	Formulate the distinguishing different persons as an ILP problem; cluster detected part candidates; combine person clusters and labeled parts to obtain final poses	–
Insafutdinov et al. (2016)	ResNet	Employs image-conditioned pairwise terms to assemble the part proposals	–
Cao et al. (2017)	VGG-19 + CPM	OpenPose; real-time; Simultaneous joints detection and association in a two-branch architecture; propose Part Affinity Fields (PAFs) to encode the location and orientation of limbs	61.8[a]
Newell et al. (2017)	Hourglass	Simultaneous joints detection and association in one branch; propose dense associative embedding tags for detected joints grouping	65.5
Nie et al. (2018)	Hourglass	Simultaneous joints detection and association in a two-branch architecture; generate partitions in the embedding space parameterized by person centroids over joint candidates; estimate pose instances by a local greedy inference approach	–
Papandreou et al. (2018)	ResNet	Multi-task (pose estimation and instance segmentation) network; simultaneous joints detection and association in a multi-branch architecture; multi-range joint offsets following tree-structured kinematic graph to guide joints grouping	68.7
Kocabas et al. (2018)	ResNet-FPN + RetinaNet	Multi-task (pose estimation, person detection and person segmentation) network; simultaneous keypoint detection and person detection in a two-branch architecture; proposes a Pose Residual Network (PRN) to assign keypoint detection to person instances	69.6
Kreiss et al. (2019)	ResNet-50	Predicts Part Intensity Fields (PIF) and Part Association Fields (PAF) to represent body joints location and body joints association; works well under low-resolution	66.7

[a] The results were obtained with COCO16 training set, while others with COCO17 training set.

employed for 3D body pose estimation, all these systems work in very constrained environments or need special markers on human body. Monocular camera, as the most widely used sensor, is very important for 3D human pose estimation. Deep neural networks have the capability to estimate the dense depth (Li et al., 2015a, 2018a, 2019) and sparse depth points (joints) as well from monocular images. Moreover, the progress of 3D human pose estimation from monocular inputs can further improve multi-view 3D human pose estimation in constrained environments. Thus, this section focuses on the deep learning-based methods that estimate 3D human pose from monocular RGB images and videos including 3D single person pose estimation and 3D multi-person pose estimation.

4.1. 3D single person pose estimation

Compared to 2D HPE, 3D HPE is more challenging since it needs to predict the depth information of body joints. In addition, the training data for 3D HPE are not easy to obtain as 2D HPE. Most existing datasets are obtained under constrained environments with limited generalizability. For single person pose estimation, the bounding box of the person in the image is normally provided, and hence it is not necessary to combine the process of person detection. In this section, we divide the methods of 3D single person pose estimation into model-free and model-based categories and summarize the recent work in Table 5. The last column of Table 5 is the comparisons of Mean Per Joint Position Error (MPJPE) in millimeter on Human3.6M dataset under protocol #1. More details of datasets and evaluation metrics are described in Section 5.

4.1.1. Model-free methods

The model-free methods do not employ human body models as the predicted target or intermediate cues. They can be roughly categorized into two types: (1) directly map an image to 3D pose, and (2) estimate depth following intermediately predicted 2D pose from 2D pose estimation methods.

Approaches that directly estimate the 3D pose from image features usually contain very few constraints. Li and Chan (2014) employed a shallow network to regress 3D joint coordinates directly with synchronous task of body part detection with sliding windows. Pavlakos et al. (2017) proposed a volumetric representation for 3D human pose and employed a coarse-to-fine prediction scheme to refine predictions with a multi-stage structure. Some researchers attempted to add body structure information or the dependencies between human joints to the deep learning networks. Li et al. (2015b) designed an embedding sub-network learning latent pose structure information to guide the 3D joint coordinates mapping. The sub-network can assign matching scores for input image-pose pairs with a maximum-margin cost function. Tekin et al. (2016) pre-trained an unsupervised auto-encoder to learn a high-dimensional latent pose representation of 3D pose for adding implicit constraints about the human body and then used a shallow network

Table 5
Comparison of 3D single person pose estimation methods. Here “E.” stands for “Extra data” and “T.” indicates “Temporal info”. The last column is the Mean Per Joint Position Error (MPJPE) in millimeter on Human3.6M dataset under protocol #1.

Methods	Backbone	E.	T.	Highlights	MPJPE (mm)
Model-free					
Li and Chan (2014)	Shallow CNNs	✗	✗	A multi-task network to predict of body part detection with sliding windows and 3D pose estimation jointly	132.2[a]
Li et al. (2015b)	Shallow CNNs	✗	✗	Compute matching score of image-pose pairs	120.2[a]
Tekin et al. (2016)	auto-encoder + shallow CNNs	✗	✗	Employ an auto-encoder to learn a high-dimensional representation of 3D pose; use a shallow CNNs network to learn the high-dimensional pose representation	116.8[a]
Tekin et al. (2017)	Hourglass	✓	✗	Predict 2D heatmaps for joints first; then use a trainable fusion architecture to combine 2D heatmaps and extracted features; 2D module is pre-trained with MPII	69.7
Chen and Ramanan (2017)	CPM	✓	✗	Estimate 2D poses from images first; then estimate depth of them by matching to a library of 3D poses; 2D module is pre-trained with MPII	82.7/57.5[b]
Moreno-Noguer (2017)	CPM	✓	✗	Use Euclidean Distance Matrices (EDMs) to encoding pairwise distances of 2D and 3D body joints; train a network to learn 2D-to-3D EDM regression; jointly trained with other 3D (Humaneva-I) dataset	87.3
Pavlakos et al. (2017)	Hourglass	✓	✗	Volumetric representation for 3D human pose; a coarse-to-fine prediction scheme; 2D module is pre-trained with MPII	71.9
Zhou et al. (2017)	Hourglass	✓	✗	A proposed loss induced from a geometric constraint for 2D data; bone-length constraints; jointly trained with 2D (MPII) dataset	64.9
Martinez et al. (2017)	Hourglass	✓	✗	Directly map predicted 2D poses to 3D poses with two linear layers; 2D module is pre-trained with MPII; process in real-time	62.9/45.5[b]
Sun et al. (2017)[c]	ResNet	✓	✗	A bone-based representation involving body structure information to enhance robustness; bone-length constraints; jointly trained with 2D (MPII) dataset	48.3
Yang et al. (2018)	Hourglass	✓	✗	Adversarial learning for domain adaptation of 2D/3D datasets; adopted generator from Zhou et al. (2017); multi-source discriminator with image, pairwise geometric structure and joint location; jointly trained with 2D (MPII) dataset	58.6
Pavlakos et al. (2018a)	Hourglass	✓	✗	Volumetric representation for 3D human pose; additional ordinal depths annotations for human joints; jointly trained with 2D (MPII) and 3D (Humaneva-I) datasets	56.2
Sun et al. (2018)	Mask R-CNN	✓	✗	Volumetric representation for 3D human pose; integral operation unifies the heat map representation and joint regression; jointly trained with 2D (MPII) dataset	40.6
Li and Lee (2019)	Hourglass	✓	✗	Multiple hypotheses of 3D poses are generated from 2D poses; the best one is chosen by 2D reprojections; 2D module is pre-trained with MPII	52.7
Model-based					
Bogo et al. (2016)[c]	DeepCut	✗	✗	SMPL model; fit SMPL model to 2D joints by minimizing the distance between 2D joints and projected 3D model joints	82.3
Zhou et al. (2016)[c]	ResNet	✗	✗	Kinematic model; embedded a kinematic object model into network for general articulated object pose estimation; orientation and rotational constrains	107.3
Mehta et al. (2017b)[c]	ResNet	✓	✓	A real-time pipeline with temporal smooth filter and model-based kinematic skeleton fitting; 2D module is pre-trained with MPII and LSP; process in real-time; provide body height	80.5
Tan et al. (2017)	Shallow CNNs	✗	✗	SMPL model; first train a decoder to predict a 2D body silhouette from parameters of SMPL; then train a encoder–decoder network with images and corresponding silhouettes; the trained encoder can predict parameters of SMPL from images	–
Mehta et al. (2017a)	Resnet	✓	✗	Kinematic model; transfer learning from features learned for 2D pose estimation; 2D pose prediction as auxiliary task; predict relative joint locations following the kinematic tree body model; jointly trained with 2D (MPII and LSP) datasets	74.1
Nie et al. (2017)	RMPE + LSTM	✓	✗	Kinematic model; joint depth estimation from global 2D pose with skeleton-LSTM and local body parts with patch-LSTM; 2D module is pre-trained with MPII	79.5
Kanazawa et al. (2018)[c]	ResNet	✓	✗	SMPL model; adversarial learning for domain adaptation of 2D images and 3D human body model; propose a framework to learn parameters of SMPL; jointly trained with 2D (LSP, MPII and COCO) datasets; process in real-time	88.0
Pavlakos et al. (2018b)[c]	Hourglass	✓	✗	SMPL model; first predict 2D heatmaps of joint and human silhouette; second generate parameters of SMPL; 2D module is trained with MPII and LSP	75.9

(*continued on next page*)

Table 5 (*continued*).

Methods	Backbone	E.	T.	Highlights	MPJPE (mm)
Omran et al. (2018)[c]	RefineNet	✗	✗	SMPL model; first predict 2D body parts segmentation from the RGB image; second take this segmentation to predict the parameters of SMPL	59.9
Varol et al. (2018)	Hourglass	✓	✗	SMPL model; first predict 2D pose and 2D body parts segmentation; second predict 3D pose; finally predict volumetric shape to fit SMPL model; 2D modules are trained with MPII and SURREAL	49.0
Arnab et al. (2019)[c]	ResNet	✓	✓	SMPL model; 2D keypoints, SMPL and camera parameters estimation; off-line bundle adjustment with temporal constraints; 2D module is trained with COCO	77.8/63.3[b]
Tome et al. (2017)	CPM	✓	✗	Pre-trained probabilistic 3D pose model; 3D lifting and projection by probabilistic model within the CPM-like network; 2D module is pre-trained with MPII; process in real-time	88.4
Rhodin et al. (2018a)	Hourglass	✗	✗	A latent variable body model learned from multi-view images; an encoder–decoder to predict a novel view image from a given one; the pre-trained encoder with additional shallow layers to predict 3D poses from images	–

[a] The results were reported from 6 actions in testing set, while others from all 17 actions.

[b] The results were reported with 2D joint ground truth.

[c] The methods report joint rotation as well.

to learn the high-dimensional pose representation. Sun et al. (2017) proposed a structure-aware regression approach. They designed a bone-based representation involving body structure information which is more stable than only using joint positions. Pavlakos et al. (2018a) trained the network with additional ordinal depths of human joints as constraints, by which the 2D human datasets can also be feed in with ordinal depths annotations.

The 3D HPE methods which intermediately estimate 2D poses gain the advantages of 2D HPE, and can easily utilize images from 2D human datasets. Some of them adopt off-the-shelf 2D HPE modules to first estimate 2D poses, then extend to 3D poses. Martinez et al. (2017) designed a 2D-to −3D pose predictor with only two linear layers. Zhou et al. (2017) presented a depth regression module to predict 3D pose from 2D heatmaps with a proposed geometric constraint loss for 2D data. Tekin et al. (2017) proposed a two-branch framework to predict 2D heatmaps and extract features from images. The extracted features are fused with 2D heatmaps by a trainable fusion scheme instead of being hand-crafted to obtain the final 3D joint coordinates. Li and Lee (2019) considered 3D HPE as an inverse problem with multiple feasible solutions. Multiple feasible hypotheses of 3D poses are generated from 2D poses and the best one is chosen by 2D reprojections. Qammaz and Argyros (2019) proposed MocapNET directly encoding 2D poses into the 3D BVH (Meredith et al., 2001) format for subsequent rendering. By consolidating OpenPose (Cao et al., 2017) the architecture estimated and rendered 3D human pose in real-time using only CPU processing.

When mapping 2D pose to 3D pose, different strategies may be applied. Chen and Ramanan (2017) used a matching strategy for an estimated 2D pose and 3D pose from a library. Moreno-Noguer (2017) encoded pairwise distances of 2D and 3D body joints into two Euclidean Distance Matrices (EDMs) and trained a regression network to learn the mapping of the two matrices. Wang et al. (2018a) predicted depth rankings of human joints as a cue to infer 3D joint positions from a 2D pose. Yang et al. (2018) adopted a generator from Zhou et al. (2017) and designed a multi-source discriminator with image, pairwise geometric structure, and joint location information.

4.1.2. Model-based methods

Model-based methods generally employ a parametric body model or template to estimate human pose and shape from images. Early geometric-based models are not included in this paper. More recent models are estimated from multiple scans of diverse people (Hasler et al., 2009; Loper et al., 2015; Pons-Moll et al., 2015; Zuffi and Black, 2015) or combination of different body models (Joo et al., 2018). These models are typically parameterized by separate body pose and shape components.

Some work employed the body model of SMPL (Loper et al., 2015) and attempted to estimate the 3D parameters from images. For example, Bogo et al. (2016) fit SMPL model to estimated 2D joints and proposed an optimization-based method to recover SMPL parameters from 2D joints. Tan et al. (2017) inferred SMPL parameters by first training a decoder to predict silhouettes from SMPL parameters with synthetic data, and then learning an image encoder with the trained decoder. The trained encoder can predict SMPL parameters from input images. Directly learning parameters of SMPL is hard, some work predicted intermediate cues as constrains. For example, intermediate 2D pose and human body segmentation (Pavlakos et al., 2018b), body parts segmentation (Omran et al., 2018), 2D pose and body parts segmentation (Varol et al., 2018). In order to overcome the problem of lacking training data for the human body model, Kanazawa et al. (2018) employed adversarial learning by using a generator to predict parameters of SMPL, and a discriminator to distinguish the real SMPL model and the predicted ones. Arnab et al. (2019) reconstructed person from video sequences which explored the multiple views information.

Kinematic model is widely used for 3D HPE. Mehta et al. (2017a) predicted relative joint locations from 2D heatmaps following the kinematic tree body model. Nie et al. (2017) employed LSTM to exploit global 2D joint locations and local body part images following kinematic tree body model which are two cues for joint depth estimation. Zhou et al. (2016) embedded a kinematic object model into a network for general articulated object pose estimation which provides orientation and rotational constrains. Mehta et al. (2017b) proposed a pipeline for 3D single HPE running in real-time. The temporal information and kinematic body model are used as a smooth filter and skeleton fitting respectively. Rhodin et al. (2018a) used an encoder–decoder network to learn a latent variable body model without 2D or 3D annotations under self-supervision, then employed the pre-trained encoder to predict 3D poses.

Additional to those typical body models, latent 3D pose model learned from data is also used for 3D HPE. Tome et al. (2017) proposed a multi-stage CPM-like network including a pre-trained probabilistic 3D pose model layer which can generate 3D pose from 2D heatmaps.

4.2. 3D multi-person pose estimation

The achievements of monocular 3D multi-person pose estimation are based on 3D single person pose estimation and other deep learning methods. This research field is pretty new and only a few methods are proposed. Table 6 summarizes these methods.

Mehta et al. (2018) proposed a bottom-up method by using 2D pose and part affinity fields to infer person instances. An occlusion-robust pose-maps (ORPM) is proposed to provide multi-style occlusion

Table 6
Summary of 3D multi-person pose estimation methods.

Methods	Network type	Highlights
Mehta et al. (2018)	ResNet	Propose an occlusion-robust pose-maps (ORPM) for full-body pose inference even under (self-)occlusions; combine 2D pose and part affinity fields to infer person instances
Rogez et al. (2017)	Faster R-CNN + VGG-16	Localize human bounding boxes with Faster R-CNN; classify the closest anchor-pose for each proposal; regress anchor-pose to get final pose
Zanfir et al. (2018)	DMHS	Feed forward process of body parts semantic segmentation and 3d pose estimates; feed backward process of refining pose and shape parameters of a body model SMPL
Mehta et al. (2019)	SelecSLS Net	Real-time; a new CNN architecture that uses selective long and short range skip connections; 2D and 3D pose features prediction along with identity assignments for all visible joints of all individuals; complete 3D pose reconstruction including occluded joints; temporal stability refinement and kinematic skeleton fitting.

information regardless of the number of people. Rogez et al. (2017) proposed a Localization–Classification–Regression Network (LCR-Net) following three-stage processing. First, Faster R-CNN is employed to detect people locations. Second, each pose proposal is assigned with the closest anchor-pose scored by a classifier. The final poses are refined with a regressor respectively. Zanfir et al. (2018) proposed a framework with feed forward and feed backward stages for 3D multi-person pose and shape estimation. The feed forward process includes semantic segmentation of body parts and 3D pose estimates based on DMHS (Popa et al., 2017). Then the feed backward process refines the pose and shape parameters of SMPL (Loper et al., 2015). Mehta et al. (2019) estimated multiple poses in real-time with three stages. First, SelecSLS Net infers 2D pose and intermediate 3D pose encoding for visible body joints. Then based on each detected person, it reconstructs the complete 3D pose, including occluded joints. Finally, refinement is provided for temporal stability and kinematic skeleton fitting.

5. Datasets and evaluation protocols

Datasets play an important role in deep learning-based human pose estimation. Datasets not only are essential for fair comparison of different algorithms but also bring more challenges and complexity through their expansion and improvement. With the maturity of the commercial motion capture systems and crowdsourcing services, recent datasets are no longer limited by the data quantity or lab environments.

This section discusses the popular publicly available human pose datasets for 2D and 3D human pose estimation, introduces the characteristics and the evaluation methods, as well as the performance of recent state-of-the-art work on several popular datasets. In addition to these basic datasets, some researchers have extended the existing datasets in their own way (Pavlakos et al., 2018a; Lassner et al., 2017). In addition, some relevant human datasets are also within the scope of this section (Güler et al., 2018). A brief description of how researchers collected all the annotated images of each dataset is also provided to bring inspiration to readers who want to generate their own datasets.

5.1. Datasets for 2D human pose estimation

Before deep learning brings significant progress for 2D HPE, there are many 2D human pose datasets for specific scenarios and tasks. Upper body pose datasets include Buffy Stickmen (Ferrari et al., 2008) (frontal-facing view, from indoor TV show), ETHZ PASCAL Stickmen (Eichner et al., 2009) (frontal-facing view, from PASCAL VOC (Everingham et al., 2010)), We Are Family (Eichner and Ferrari, 2010) (Group photo scenario), Video Pose 2 (Sapp et al., 2011) (from indoor TV show), Sync. Activities (Eichner and Ferrari, 2012b) (sports, full-body image, upper body annotation). Full-body pose datasets include PASCAL Person Layout (Everingham et al., 2010) (daily scene, from PASCAL VOC (Everingham et al., 2010)), Sport (Wang et al., 2011) (sport scenes) and UIUC people (Li and Fei-fei, 2007) (sport scenes). For detailed description of these datasets, we refer interested readers to several well-summarized papers (Andriluka et al., 2014; Gong et al., 2016).

Above earlier datasets for 2D human pose estimation have many shortcomings such as few scenes, monotonous view angle, lack of diverse activities, and limited number of images. The scale is the most important aspect of a dataset for deep learning-based methods. Small training sets are insufficient for learning robust features, unsuitable for networks with deep layers and complex design, and may easily cause overfitting. Thus in this section, we only introduce 2D human pose datasets with the number of images for training over 1000. The features of these selected 2D HPE datasets are summarized in Table 7 and some sample images with annotations are illustrated in Fig. 5.

Frames Labeled In Cinema (FLIC) Dataset (Sapp and Taskar, 2013) contains 5003 images collected from popular Hollywood movies. For every tenth frame of 30 movies, a person detector (Bourdev and Malik, 2009) was run to obtain about 20k person candidates. Then all candidates are sent to Amazon Mechanical Turk to obtain ground truth labeling for 10 upper body joints. Finally, images with person occluded or severely non-frontal views are manually deleted. The undeleted original set called FLIC-full consisting of occluded, non-frontal, or just plain mislabeled examples (20,928 examples) is also available. Moreover, in Tompson et al. (2014), the FLIC-full dataset is further cleaned to FLIC-plus to make sure that the training subset does not include any images from the same scene as the test subset.

Leeds Sports Pose (LSP) Dataset (Johnson and Everingham, 2010) contains 2000 images of full-body poses collected from Flickr by downloading with 8 sports tags (athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, and volleyball). Each image is annotated with up to 14 visible joint locations. Further, the extension version Leeds Sports Pose Extended (LSP-extended) training dataset (Johnson and Everingham, 2011) is gathered to extend the LSP dataset only for training. It contains 10,000 images collected from Flickr searches with 3 most challenging tags (parkour, gymnastics, and athletics). The annotations were conducted through Amazon Mechanical Turk and the accuracy cannot be guaranteed.

Max Planck Institute for Informatics (MPII) Human Pose Dataset (Andriluka et al., 2014) is one of current the state-of-the-art benchmarks for evaluation of articulated human pose estimation with rich annotations. First, with guidance from a two-level hierarchy of human activities from Ainsworth et al. (2011), 3913 videos spanning 491 different activities are downloaded from YouTube. Then frames that either contains different people in the video or the same person in a very different pose were manually selected which results in 24,920 frames. Rich annotations including 16 body joints, the 3D viewpoint of the head and torso and position of the eyes and nose are labeled by in-house workers and on Amazon Mechanical Turk. For corresponding joints, visibility and left/right labels are also annotated in a person-centric way. Images in MPII have various body poses and are suitable

Table 7
Popular 2D databases for human pose estimation. Selected example images with annotations are shown in Fig. 5. Here Jnt. indicates the number of joints.

Dataset name	Single/Multiple	Jnt.	Number of images/videos			Evaluation protocol	Highlights
			Train	Val	Test		
Image-based							
FLIC	Single	10	≈5k	0	≈1k	PCP&PCK	Upper body poses; Sampled from movies; FLIC-full is complete version (Sapp and Taskar, 2013); FLIC-plus is cleaned version (Tompson et al., 2014); FLIC is a simple version with no difficult poses.
FLIC-full			≈20k	0	0		
FLIC-plus			≈17k	0	0		
LSP	Single	14	≈1k	0	≈1k	PCP	Full-body poses; From Flickr with 8 sports tags (Johnson and Everingham, 2010); Extended by adding most challenging poses lie in 3 tags (Johnson and Everingham, 2011).
LSP-extended			≈10k	0	0		
MPII	Single	16	≈29k	0	≈12k	PCPm/PCKh	Various body poses; Downloaded videos from YouTube; Multiple annotations (bounding boxes, 3D viewpoint of the head and torso, position of the eyes and nose, joint locations); Andriluka et al. (2014).
	Multiple		≈3.8k	0	≈1.7k	mAP	
COCO16	Multiple	17	≈45k	≈22k	≈80k	AP	Various body poses; From Google, Bing and Flickr; Multiple annotations (bounding boxes, human body masks, joint locations); With about 120k unlabeled images for semi-supervised learning; Lin et al. (2014)
COCO17			≈64k	≈2.7k	≈40k		
AIC-HKD	Multiple	14	≈210k	≈30k	≈60k	AP	Various body poses; From Internet search engines; Multiple annotations (bounding boxes, joint locations); Wu et al. (2017)
Video-based							
Penn Action	Single	13	≈1k	0	≈1k	–	Full-body poses; From YouTube; 15 actions; Multiple annotations (joint locations, bounding boxes, action classes) (Zhang et al., 2013).
J-HMDB	Single	15	≈0.6k	0	≈0.3k	–	Full-body poses; Generated from action recognition dataset; 21 actions; Multiple annotations (joint positions and relations, optical flows, segmentation masks) (Jhuang et al., 2013).
PoseTrack	Multiple	15	292	50	208	mAP	Various body poses; Extended from MPII; Dense annotations (joint locations, head bounding boxes) (Andriluka et al., 2018).

for many tasks such as 2D single/multiple human pose estimation, action recognition, etc.

Microsoft Common Objects in Context (COCO) Dataset (Lin et al., 2014) is a large-scale dataset that was originally proposed for daily object detection and segmentation in natural environments. With improvements and extensions, the usage of COCO covers image captioning and keypoint detection. Images are collected from Google, Bing, and Flickr image search with isolated or pairwise object categories. Annotations were conducted on Amazon Mechanical Turk. The whole set contains more than 200,000 images and 250,000 labeled person instances. Suitable examples are selected for human pose estimation, thus forming two datasets: COCO keypoints 2016 and COCO keypoints 2017, corresponding to two public keypoint detection challenges respectively. The only difference between these two versions is the train/val/test splitting strategy based on community feedback (shown in Table 7), and cross-year results can be compared directly since the images in the test set are same. The COCO Keypoint Detection Challenge aims to localize keypoints of people in uncontrolled images. The annotations for each person include 17 body joints with visibility and left/right labels, and instance human body segmentation. Note that COCO dataset contains about 120k unlabeled images following the same class distribution as the labeled images which can be used for unsupervised or semi-supervised learning.

AI Challenger Human Keypoint Detection (AIC-HKD) Dataset (Wu et al., 2017) has the largest number of training examples. It contains 210,000, 30,000, 30,000, and 30,000 images for training, validation, test A, and test B respectively. The images, focusing on the daily life of people, were collected from Internet search engines. Then, after removing inappropriate examples (e.g. with the political, constabulary, violent and sexual contents; too small or too crowded human figures), each person in the images were annotated with a bounding box and 14 keypoints. Each keypoint has the visibility and left/right labels.

In addition to the datasets described above which are in static image style, datasets with densely annotated video frames are collected in closer to real-life application scenarios which offer the possibility to utilize temporal information and can be used for action recognition. Some of them focus on single individuals (Zhang et al., 2013; Jhuang et al., 2013; Charles et al., 2016) and others have pose annotations for multiple people (Insafutdinov et al., 2017; Iqbal et al., 2017; Andriluka et al., 2018).

Penn Action Dataset (Zhang et al., 2013) consists of 2326 videos downloaded from YouTube covering 15 actions: baseball pitch, baseball swing, bench press, bowling, clean and jerk, golf swing, jump rope, jumping jacks, pull up, push up, sit up, squat, strum guitar, tennis forehand, and tennis serve. Annotations for each frame were labeled by VATIC (Vondrick et al., 2013) (an annotation tool) on Amazon Mechanical Turk. Each video involves an action class label and each video frame contains a bounding box of human and 13 joint locations with the visibility and left/right labels.

Joint-annotated Human Motion Database (J-HMDB) (Jhuang et al., 2013) is based on the HMDB51 (Jhuang et al., 2011) which is originally collected for action recognition. First, 21 action categories with relatively large body movements were selected from original 51 actions in HMDB51, including: brush hair, catch, clap, climb stairs, golf, jump, kick ball, pick, pour, pull-up, push, run, shoot ball, shoot bow, shoot gun, sit, stand, swing baseball, throw, walk, and wave. Then, after a selection-and-cleaning process, 928 clips comprising 31,838 annotated frames are selected. Finally, a 2D articulated human puppet model (Zuffi et al., 2012) is employed to generate all the needed annotations using Amazon Mechanical Turk. The 2D puppet model is an articulated human body model that provides scale, pose, segmentation, coarse viewpoint, and dense optical flow for the humans in actions. The annotations include 15 joint positions and relations, 2D optical flow corresponding to the human motion, human body segmentation mask. The 70% images are used for training and the 30% images for testing. J-HMDB can also be used for action recognition and human detection tasks.

There are several video datasets annotated with human upper body pose. **BBC Pose** (Charles et al., 2014) contains 20 videos (10/5/5 for train/val/test, 1.5 million frames in total) with 9 sign language signers. 2000 frames for validation and test are manually annotated

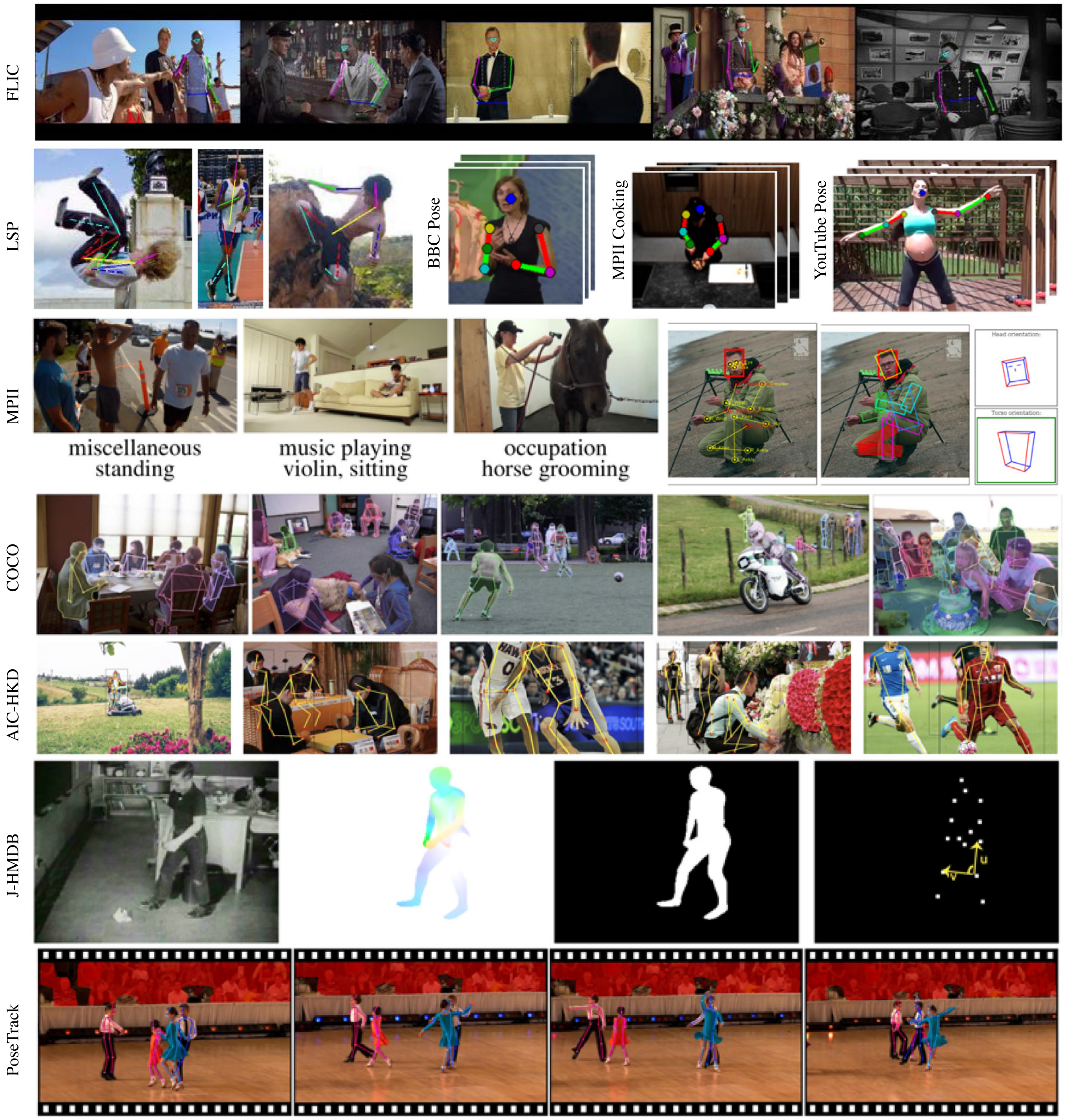

Fig. 5. Some selected example images with annotations from typical 2D human pose estimation datasets.

and the rest of the frames are annotated with a semi-automatic method. **Extended BBC Pose** dataset (Pfister et al., 2014) adds 72 additional training videos for **BBC Pose** which has about 7 million frames in total. **MPII Cooking** (Rohrbach et al., 2012) dataset contains 1071 frames for training and 1277 frames for testing with manually annotated joint locations for cooking activities. **YouTube Pose** dataset (Charles et al., 2016) contains 50 YouTube videos with single person in. The activities cover dancing, stand-up comedy, how-to, sports, disk jockeys, performing arts and dancing, and sign language signers. 100 frames of each video are manually annotated with joint locations of the upper body. The scenes of these datasets are relatively simple, with static views and the characters are normally in a small motion range.

From unlabeled MPII Human Pose (Andriluka et al., 2014) video data, there are several extended versions result in dense annotations of video frames. The general approach is to extend the original labeled frame with the connected frames both forward and backward and annotate unlabeled frames in the same way as the labeled frame. **MPII Video Pose** dataset (Insafutdinov et al., 2017) provides 28 videos containing 21 frames each by selecting the challenging labeled images and unlabeled neighboring ±10 frames from the MPII dataset. In **Multi-Person PoseTrack** dataset (Iqbal et al., 2017), each selected labeled frame is extended with unlabeled clips ranging ±20 frames, and each person has a unique ID. Also, additional videos of more than 41 frames are provided for longer and variable-length sequences. In total, it contains 60 videos with additional videos with more than 41 frames for longer and variable-length sequences. **PoseTrack** dataset (Andriluka et al., 2018) is the integrated expansion of the above two datasets and is the current largest multi-person pose estimation and tracking dataset. Each person in the video has a unique track ID with annotations of a head bounding box and 15 body joint locations. All pose annotations are labeled with VATIC (Vondrick et al., 2013). PoseTrack contains 550 video sequences with the frames mainly ranging between 41 and 151 frames in a wide variety of everyday human activities and is divided into 292, 50, and 208 videos for training, validation, and testing, following original MPII split strategy.

5.2. Evaluation metrics of 2D human pose estimation

Different datasets have different features (e.g. various range of human body sizes, upper/full human body) and different task requirements (single/multiple pose estimation), so there are several evaluation metrics for 2D human pose estimation. The summary of different evaluation metrics which are commonly used are listed in Table 8.

Percentage of Correct Parts (PCP) (Ferrari et al., 2008) is widely used in early research. It reports the localization accuracy for limbs. A limb is correctly localized if its two endpoints are within a threshold from the corresponding ground truth endpoints. The threshold can be 50% of the limb length. Besides a mean PCP, some limbs PCP (torso, upper legs, lower legs, upper arms, forearms, head) normally are also reported (Johnson and Everingham, 2010). And percentage curves for each limb can be obtained with the variation of threshold in the metric (Gkioxari et al., 2013). The similar metrics PCPm from Andriluka et al. (2014) use 50% of the mean ground-truth segment length over the entire test set as a matching threshold.

Percentage of Correct Keypoints (PCK) (Yang and Ramanan, 2013) measures the accuracy of the localization of the body joints. A candidate body joint is considered as correct if it falls within the threshold pixels of the ground-truth joint. The threshold can be a fraction of the person bounding box size (Yang and Ramanan, 2013), pixel radius that normalized by the torso height of each test sample (Sapp and Taskar, 2013) (denoted as Percent of Detected Joints (PDJ) in Toshev and Szegedy (2014)), 50% of the head segment length of each test image (denoted as **PCKh@0.5** in Andriluka et al. (2014)). Also, with the variation of a threshold, Area Under the Curve (AUC) can be generated for further analysis.

The Average Precision (AP). For systems in which there are only joint locations but no annotated bounding boxes for human bodies/heads or number of people in the image as ground truth at testing, the detection problem must be addressed as well. Similar to object detection, an Average Precision (AP) evaluation method is proposed, which is first called Average Precision of Keypoints (APK) in Yang and Ramanan (2013). In AP measure, if a predicted joint falls within a threshold of the ground-truth joint location, it is counted as a true positive. Note that correspondence between candidates and ground-truth poses are established separately for each keypoint. For multi-person pose evaluation, all predicted poses are assigned to the ground truth poses one by one based on the PCKh score order, while unassigned predictions are counted as false positives (Pishchulin et al., 2016). The mean average precision (mAP) is reported from the AP of each body joint.

Average Precision (AP), Average Recall (AR) and their variants. In Lin et al. (2014), evaluating multi-person pose estimation results as an object detection problem is further designed. AP, AR, and their variants are reported based on an analogous similarity measure: object keypoint similarity (OKS) which plays the same role as the Intersection over Union (IoU). Additional, AP/AR with different human body scales are also reported in COCO dataset. Table 8 summarizes all above evaluation metrics.

Frame Rate, Number of Weights and Giga Floating-point Operations Per Second (GFLOPs). The computational performance metrics are also very important for HPE. Frame Rate indicates the processing speed of input data, generally expressed by Frames Per Second (FPS) or seconds per image (s/image) (Cao et al., 2017). Number of Weights and GFLOPs show the efficiency of the network, mainly related to the network design and the specific used GPUs/CPUs (Sun et al., 2019). These computational performance metrics are suitable for 3D HPE as well.

5.3. Datasets for 3D human pose estimation

For a better understanding of the human body in 3D space, there are many kinds of body representations with different modern equipment. 3D human body shape scans, such as **SCAPE** (Anguelov et al., 2005), **INRIA4D** (INRIA4D, 2019) and **FAUST** (Bogo et al., 2014, 2017), 3D human body surface cloud points with time of flight (TOF) depth sensors (Shahroudy et al., 2016), 3D human body reflective markers capture with motion capture systems (MoCap) (Sigal et al., 2010; Ionescu et al., 2014), orientation and acceleration of 3D human body data with Inertial Measurement Unit (IMU) (von Marcard et al., 2016, 2018). It is difficult to summarize them all, this paper summarizes the datasets that involve RGB images and 3D joint coordinates. The details of the selected 3D datasets are summarized in Table 9 and some example images with annotations are shown in Fig. 6.

HumanEva-I&II Datasets (Sigal et al., 2010). The ground truth annotations of both datasets were captured with a commercial MoCap system from ViconPeak. The HumanEva-I dataset contains 7-view video sequences (4 grayscales and 3 colors) which are synchronized with 3D body poses. There are 4 subjects with markers on their bodies performing 6 common actions (e.g. walking, jogging, gesturing, throwing and catching a ball, boxing, combo) in an 3 m × 2 m capture area. HumanEva-II is an extension of HumanEva-I dataset for testing, which contains 2 subjects performing the action combo.

Human3.6M Dataset (Ionescu et al., 2014) was collected using accurate marker-based MoCap systems (Vicon, 2019) in an indoor laboratory setup with 11 professional actors (5 females and 6 males) dressing moderately realistic clothing. It contains 3.6 million 3D human poses and corresponding images from 4 different views. The performed 17 daily activities include discussion, smoking, taking photos, talking on the phone, etc. Main capturing devices include 4 digital video cameras, 1 time-of-flight sensor, 10 motion cameras working synchronously. The capture area is about 4 m × 3 m. The provided annotations include 3D joint positions, joint angles, person bounding boxes, and 3D laser scans of each actor. For evaluation, there are three protocols with different training and testing data splits (protocol #1, protocol #2 and protocol #3.)

TNT15 Dataset (von Marcard et al., 2016) consists of synchronized data streams from 8 RGB-cameras and 10 IMUs. It has been recorded in an office environment. The dataset records 4 actors performing five activities (e.g. walking, running on the spot, rotating arms, jumping and skiing exercises, dynamic punching.) and contains about 13k frames including binary segmented images obtained by background subtraction, 3D laser scans and registered meshes of each actor.

MPI-INF-3DHP (Mehta et al., 2017a) was collected with a markerless multi-camera MoCap system (TheCaptury, 2019) in both indoor and outdoor scenes. It contains over 1.3M frames from 14 different views. Eight subjects (4 females and 4 males) are recorded performing 8 activities (e.g. walking/standing, exercise, sitting, crouch/reach, on the floor, sports, miscellaneous.)

TotalCapture Dataset (Trumble et al., 2017) was captured in indoors with space measuring roughly 8 m × 4 m with 8 calibrated HD video cameras at a frame rate of 60 Hz. There are 4 male and 1 female subjects each performing four diverse performances, repeated 3 times: Range Of Motion (ROM), Walking, Acting, and Freestyle. There is a total of 1,892,176 frames of synchronized video, IMU and Vicon data. The variation and body motions contained in particular within the acting and freestyle sequences are very challenging with actions such as yoga, giving directions, bending over and crawling performed in both the train and test data.

MARCOnI Dataset (Elhayek et al., 2017) is a test dataset containing sequences in a variety of uncontrolled indoor and outdoor scenarios. The sequences vary according to different data modalities captured (multiple videos, video + marker positions), in the numbers and identities of actors to track, the complexity of the motions, the number of cameras used, the existence and number of moving objects in the

Table 8
Summary of commonly used evaluation metrics for 2D HPE.

Metric	Meaning	Typical datasets and description	
Single person			
PCP	Percentage of Correct Parts	LSP	Percentage of correct predicted parts which their end points fall within a threshold
PCK	Percentage of Correct Keypoints	LSP MPII	Percentage of correct predicted joints which fall within a threshold
Multiple person			
AP	Average Precision	MPII PoseTrack	mean AP (mAP) is reported by AP for each body part after assigning predicted pose to the ground truth pose by PCKh score.
		COCO	• AP_{coco}: at OKS=.50:.05:.95 (primary metric) • $AP_{coco}^{OKS=.50}$: at OKS=.50 (loose metric) • $AP_{coco}^{OKS=.75}$: at OKS=.75 (strict metric) • AP_{coco}^{medium}: for medium objects: 32^2 < area < 96^2 • AP_{coco}^{large}: for large objects: area > 96^2
AR	Average Recall	COCO	• AR_{coco}: at OKS=.50:.05:.95 • $AR_{coco}^{OKS=.50}$: at OKS=.50 • $AR_{coco}^{OKS=.75}$: at OKS=.75 • AR_{coco}^{medium}: for medium objects: 32^2 < area < 96^2 • AR_{coco}^{large}: for large objects: area > 96^2
OKS	Object Keypoint Similarity	COCO	A similar role as the Intersection over Union (IoU) for AP/AR.

Table 9
Popular databases for 3D human pose estimation. Selected example images with annotations are shown in Fig. 6 (Cams. indicates the number of cameras; Jnt. indicates the number of joints).

Dataset name	Cams.	Jnt.	Number of frames/videos			Evaluation protocol	Highlights
			Train	Val	Test		
Single person							
HumanEva-I HumanEva-II	7 4	15	≈6.8k 0	≈6.8k 0	≈24k ≈2.5k	MPJPE	4/2 (I/II) subjects, 6/1 (I/II) actions, Vicon data, indoor environment. (Sigal et al., 2010)
Human3.6M	4	17	≈1.5M	≈0.6M	≈1.5M	MPJPE	11 subjects, 17 actions, Vicon data, multi-annotation (3D joints, person bounding boxes, depth data, 3D body scans), indoor environment. (Ionescu et al., 2014)
TNT15	8	15	All (≈13k)			MPJPE	4 subjects, 5 actions, IMU data, 3D body scans, indoor environment. (von Marcard et al., 2016)
MPI-INF-3DHP	14	15	All (≈1.3M)			3DPCK	8 subjects, 8 actions, commercial markerless system, indoor and outdoor scenes. (Mehta et al., 2017a)
TotalCapture	8	26	All (≈1.9M)			MPJPE	5 subjects, 5 actions, IMU and Vicon data, indoors environment. (Trumble et al., 2017)
Multiple person							
Panoptic	521	15	All (65 videos ≈5.5 hours)			3DPCK	Up to 8 subjects in each video, social interactions, markerless studio, multi-annotation (3D joints, cloud points, optical flow), indoors environment. (Joo et al., 2017)
3DPW	1	18	All (60 videos ≈51k frames)			MPJPE MPJAE	7 subjects (up to 2), daily actions, estimated 3D poses from videos and attached IMUs, 3D body scans, SMPL model fitting, in the wild. (von Marcard et al., 2018)

background, and the lighting conditions (i.e. some body parts lit and some in shadow). Cameras differ in the types (from cell phones to vision cameras), the frame resolutions, and the frame rates.

Panoptic Dataset (Joo et al., 2017) was captured with a markerless motion capturing using multiple view systems which contains 480 VGA camera views, 31 HD views, 10 RGB-D sensors and hardware-based synchronized system. It contains 65 sequences (5.5 h) of social interaction with 1.5 millions of 3D skeletons. The annotations include 3D keypoints, cloud points, optical flow, etc.

3DPW Dataset (von Marcard et al., 2018) was captured with a single hand-held camera in natural environments. 3D annotations are estimated from IMUs attached to subjects' limbs with proposed method Video Inertial Poser. All subjects are provided with 3D scans. The dataset consists of 60 video sequences (more than 51,000 frames) with daily actions including walking in the city, going up-stairs, having coffee or taking the bus.

In addition to the datasets collected with MoCap systems, there are other approaches to create a dataset for 3D human pose estimation. **JTA** (Joint Track Auto) (Fabbri et al., 2018) is a fully synthetic dataset generated from highly photorealistic video game Grand Theft Auto V. It contains almost $10M$ annotated body poses and over 460,800 densely annotated frames. In **Human3D+** (Chen et al., 2016), the training images are obtained by integrating real background images and 3D textured models which generated from SCAPE model (Anguelov et al., 2005) with different texture deformation. The parameters for generating basic SCAPE models are captured from a MoCap system, or inferred from human-annotated 2D poses. **SURREAL** (Synthetic hUmans foR REAL) (Varol et al., 2017) contains videos of single synthetic people with real unchanged background. It contains annotations of body parts segmentation, depth, optical flow, and surface normals. The dataset employs the SMPL body model for generating body poses and shapes. **LSP-MPII-Ordinal** (Pavlakos et al., 2018a) is an extension of two 2D human pose datasets (LSP (Johnson and Everingham, 2010) and MPII (Andriluka et al., 2014)) by adding the ordinal depth relation for each pair of joints. **UP-3D** (Lassner et al., 2017) is a combination of color images from 2D human pose benchmarks like LSP (Johnson and

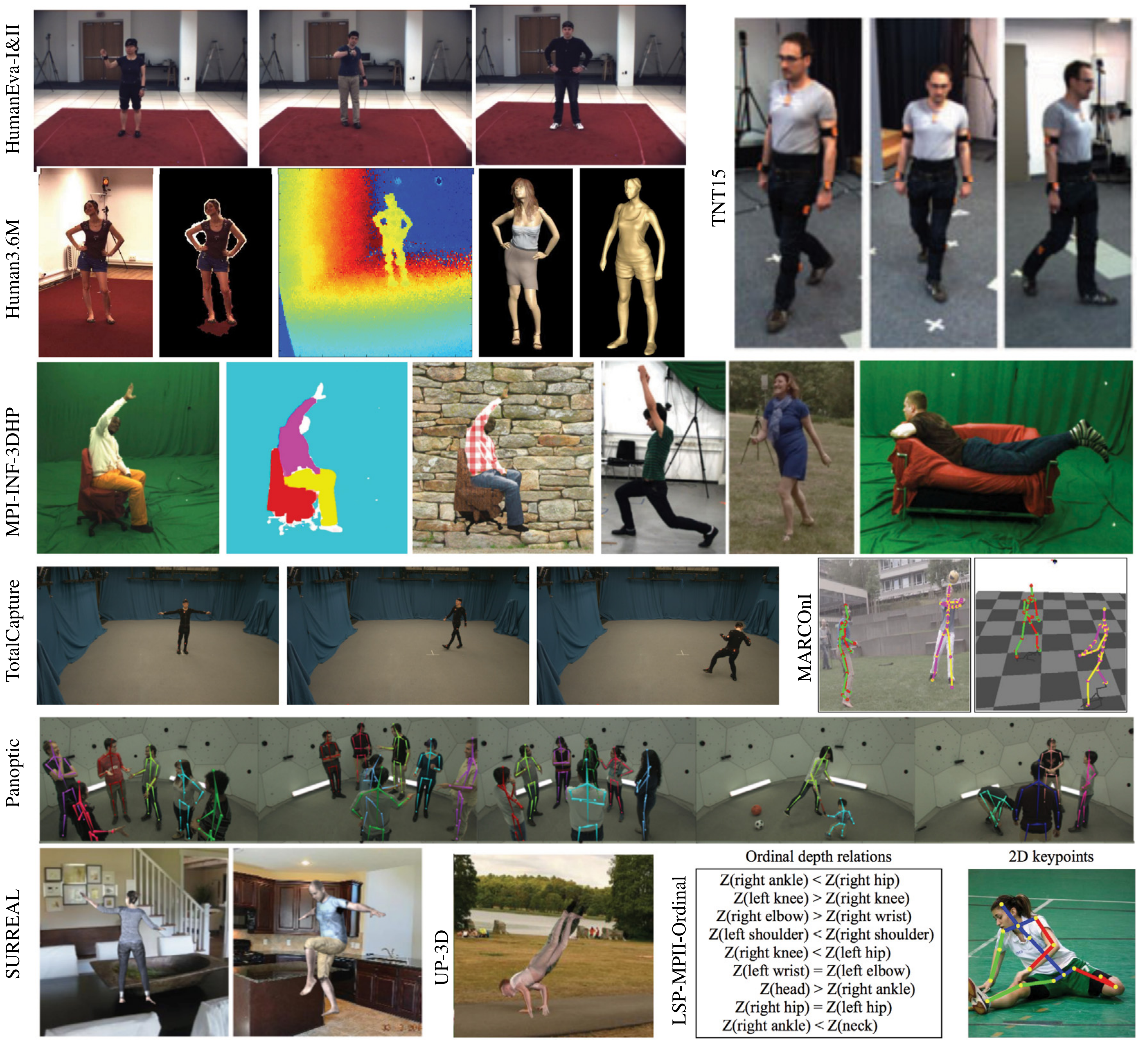
HumanEva-I&II
Human3.6M
TNT15
MPI-INF-3DHP
TotalCapture
MARCOnI
Panoptic
SURREAL
UP-3D
LSP-MPII-Ordinal
Ordinal depth relations
Z(right ankle) < Z(right hip)
Z(left knee) > Z(right knee)
Z(right elbow) > Z(right wrist)
Z(left shoulder) < Z(right shoulder)
Z(right knee) < Z(left hip)
Z(left wrist) = Z(left elbow)
Z(head) > Z(right ankle)
Z(right hip) = Z(left hip)
Z(right ankle) < Z(neck)
2D keypoints

Fig. 6. Some selected example images with annotations from typical 3D human pose estimation datasets.

Everingham, 2010) and MPII (Andriluka et al., 2014) and human body model SMPL Bogo et al. (2016). The 3D human shape candidates are fit to color images by human annotators. **DensePose** (Güler et al., 2018) is an extension on 50K COCO images with people. All RGB images are manually annotated with surface-based representations of the human body. **AMASS Dataset** (Mahmood et al., 2019) unifies 15 different optical marker-based human motion capture datasets with SMPL Loper et al. (2015) body model as a standard fitting representation for human skeleton and surface mesh. Each body joint in this rich dataset has 3 rotational Degrees of Freedom (DoF) which are parameterized with exponential coordinates.

5.4. Evaluation metrics of 3D human pose estimation

There are several evaluation metrics for 3D human pose estimation with different limitation factors. Note that we only list widely used evaluation metrics as below.

Mean Per Joint Position Error (MPJPE) is the most widely used measures to evaluate the performance of 3D pose estimation. It calculates the Euclidean distance from the estimated 3D joints to the ground truth in millimeters, averaged over all joints in one image. In the case of a set of frames, the mean error is averaged over all frames. For different datasets and different protocols, there are different data post-processing of estimated joints before computing the MPJPE. For example, in the protocol #1 of Human3.6M, the MPJPE is calculated after aligning the depths of the root joints (generally pelvis joint) (Tome et al., 2017; Yang et al., 2018), which is also called N-MPJPE (Rhodin et al., 2018a). The MPJPE in HumanEva-I and the protocol #2 & #3 of Human3.6M is calculated after the alignment of predictions and ground truth with a rigid transformation using Procrustes Analysis (Gower, 1975), which is also called reconstruction error (Kanazawa et al., 2018; Pavlakos et al., 2018b), P-MPJPE (Rhodin et al., 2018a) or PA-MPJPE (Sun et al., 2018).

Percentage of Correct Keypoints (PCK) and **Area Under the Curve (AUC)** are suggested by Mehta et al. (2017a) for 3D pose evaluation similar to PCK and AUC in MPII for 2D pose evaluation. PCK counts the percentage of points that fall in a threshold also called **3DPCK**, and AUC is computed by a range of PCK thresholds. The general threshold in 3D space is 150 mm, corresponding to roughly half of the head size.

In addition to the evaluation metrics for 3D joint coordinates, there is another evaluation measurement **Mean Per-vertex Error** to report the results of 3D body shape which report the error between predicted and ground truth meshes (Varol et al., 2018; Pavlakos et al., 2018b).

6. Conclusion and future research directions

Human pose estimation is a hot research area in computer vision that evolved recently along with the blooming of deep learning. Due to limitations in hardware device capability and the quantity and quality of training data, early networks are relatively shallow, used in a very straightforward way and can only handle small images or patches (Toshev and Szegedy, 2014; Tompson et al., 2015; Li and Chan, 2014). More recent networks are more powerful, deeper and efficient (Newell et al., 2016; Cao et al., 2017; He et al., 2017; Sun et al., 2019). In this paper, we have reviewed the recent deep learning-based research addressing the 2D/3D human pose estimation problem from monocular images or video footage and organize approaches into four categories based on specific tasks: (1) 2D single person pose estimation, (2) 2D multi-person pose estimation, (3) 3D single person pose estimation, and (4) 3D multi-person pose estimation. Further, we have summarized the popular human pose datasets and evaluation protocols.

Despite the great development of monocular human pose estimation with deep learning, there still remain some unresolved challenges and gap between research and practical applications, such as the influence of body part occlusion and crowded people. Efficient networks and adequate training data are the most important requirements for deep learning-based approaches.

Future networks should explore both global and local contexts for more discriminative features of the human body while exploiting human body structures into the network for prior constraints. Current networks have validated some effective network design tricks such as multi-stage structure, intermediate supervision, multi-scale feature fusion, multi-task learning, body structure constrains. Network efficiency is also a very important factor to apply algorithms in real-life applications.

Diversity data can improve the robustness of networks to handle complex scenes with irregular poses, occluded body limbs and crowded people. Data collection for specific complex scenarios is an option and there are other ways to extend existing datasets. Synthetic technology can theoretically generate unlimited data while there is a domain gap between synthetic data and real data. Cross-dataset supplementation, especially to supplement 3D datasets with 2D datasets can mitigate the problem of insufficient diversity of training data.

Acknowledgments

This manuscript is based upon the work supported by National Science Foundation (NSF) under award number IIS-1400802 and National Natrual Science Foundation of China (61420106007, 61671387). Yucheng Chen's contribution was made when he was a visiting student at the City University of New York, sponsored by the Chinese Scholarship Council.

References

Aggarwal, J.K., Cai, Q., 1999. Human motion analysis: A review. Comput. Vis. Image Underst. 73, 428–440.

Ainsworth, B.E., Haskell, W.L., Herrmann, S.D., Meckes, N., Basset Jr., D.R., Tudor-Locke, C., Greer, J.L., Vezina, J., Whitt-Glover, M.C., Leon, A.S., 2011. 2011 compendium of physical activities: a second update of codes and met values. Med. Sci. Sports Exerc. 43, 1575–1581.

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

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 2d human pose estimation: New benchmark and state of the art analysis. In: Proc. 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: ACM Transactions on Graphics. ACM, pp. 408–416.

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

Belagiannis, V., Zisserman, A., 2017. Recurrent human pose estimation. In: Proc. IEEE Conference on Automatic Face and Gesture Recognition. IEEE, pp. 468–475.

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: Proc. European Conference on Computer Vision. Springer, pp. 561–578.

Bogo, F., Romero, J., Loper, M., Black, M.J., 2014. FAUST: Dataset and evaluation for 3D mesh registration. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3794–3801.

Bogo, F., Romero, J., Pons-Moll, G., Black, M.J., 2017. Dynamic FAUST: Registering human bodies in motion. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 6233–6242.

Bourdev, L., Malik, J., 2009. Poselets: Body part detectors trained using 3d human pose annotations. In: Proc. IEEE International Conference on Computer Vision. IEEE, pp. 1365–1372.

Bulat, A., Tzimiropoulos, G., 2016. Human pose estimation via convolutional part heatmap regression. In: Proc. European Conference on Computer Vision. Springer, pp. 717–732.

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

Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J., 2016. Human pose estimation with iterative error feedback. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4733–4742.

Charles, J., Pfister, T., Everingham, M., Zisserman, A., 2014. Automatic and efficient human pose estimation for sign language videos. Int. J. Comput. Vis. 110, 70–90.

Charles, J., Pfister, T., Magee, D., Hogg, D., Zisserman, A., 2016. Personalizing human video pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3063–3072.

Chen, C.H., Ramanan, D., 2017. 3d human pose estimation= 2d pose estimation+ matching. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7035–7043.

Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J., 2017. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1212-1221.

Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., Chen, B., 2016. Synthesizing training images for boosting human 3d pose estimation. In: Proc. IEEE International Conference on 3D Vision. IEEE, pp. 479–488.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., 2018. Cascaded pyramid network for multi-person pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112.

Chen, L., Wei, H., Ferryman, J., 2013. A survey of human motion analysis using depth imagery. Pattern Recognit. Lett. 34, 1995–2006.

Chen, X., Yuille, A.L., 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Advances in Neural Information Processing Systems. pp. 1736–1744.

Chou, C.J., Chien, J.T., Chen, H.T., 2018. Self adversarial training for human pose estimation. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 17-30.

Chu, X., Ouyang, W., Li, H., Wang, X., 2016. Structured feature learning for pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4715–4723.

Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X., 2017. Multi-context attention for human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1831-1840.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Underst. 61, 38–59.

Dantone, M., Gall, J., Leistner, C., Va. Gool, L., 2013. Human pose estimation using body parts dependent joint regressors. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3041–3048.

Debnath, B., O'Brien, M., Yamaguchi, M., Behera, A., 2018. Adapting mobilenets for mobile based upper body pose estimation. In: Proc. IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 1–6.

Eichner, M., Ferrari, V., 2010. We are family: Joint pose estimation of multiple persons. In: Proc. European Conference on Computer Vision. Springer, pp. 228–242.

Eichner, M., Ferrari, V., 2012a. Calvin upper-body detector v1.04. URL http://groups.inf.ed.ac.uk/calvin/calvin_upperbody_detector/.

Eichner, M., Ferrari, V., 2012b. Human pose co-estimation and applications. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2282–2288.

Eichner, M., Ferrari, V., Zurich, S., 2009. Better appearance models for pictorial structures. In: Proc. British Machine Vision Conference, p. 5.

Elhayek, A., d. Aguiar, E., Jain, A., Thompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C., 2017. Marconi—convnet-based marker-less motion capture in outdoor and indoor scenes. IEEE Trans. Pattern Anal. Mach. Intell. 39, 501–514.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88, 303–338.

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: Proc. European Conference on Computer Vision, pp. 430–446.

Faessler, M., Mueggler, E., Schwabe, K., Scaramuzza, D., 2014. A monocular pose estimation system based on infrared leds. In: Proc. IEEE International Conference on Robotics and Automation. IEEE, pp. 907–913.

Fan, X., Zheng, K., Lin, Y., Wang, S., 2015. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1347-1355.

Fang, H., Xie, S., Tai, Y.W., Lu, C., 2017. Rmpe: Regional multi-person pose estimation. In: Proc. IEEE International Conference on Computer Vision, pp. 2334–2343.

Felzenszwalb, P.F., Huttenlocher, D.P., 2005. Pictorial structures for object recognition. Int. J. Compu. Vis. 61, 55–79.

Feng, Z., Xiatian, Z., Mao, Y., 2019. Fast human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

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

Gavrila, D.M., 1999. The visual analysis of human movement: A survey. Comput. Vis. Image Underst. 73, 82–98.

Gkioxari, G., Arbelaez, P., Bourdev, L., Malik, J., 2013. Articulated pose estimation using discriminative armlet classifiers. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3342–3349.

Gkioxari, G., Hariharan, B., Girshick, R., Malik, J., 2014a. R-cnns for pose estimation and action detection. arXiv preprint arXiv:1406.5212.

Gkioxari, G., Hariharan, B., Girshick, R., Malik, J., 2014b. Using k-poselets for detecting people and localizing their keypoints. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3582–3589.

Gkioxari, G., Toshev, A., Jaitly, N., 2016. Chained predictions using convolutional neural networks. In: Proc. European Conference on Computer Vision. Springer, pp. 728–743.

Gong, W., Zhang, X., Gonzàlez, A., Bouwmans, T., Tu, C., Zahzah, E.h., 2016. Human pose estimation from monocular images: A comprehensive survey. Sensors 16, 1966.

Gower, J.C., 1975. Generalized procrustes analysis. Psychometrika 40, 33–51.

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

Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P., 2009. A statistical model of human pose and body shape. In: Computer Graphics Forum. Wiley Online Library, pp. 337–346.

He, K., Gkioxari, G., Dollár, R., 2017. Mask r-cnn. In: Proc. IEEE International Conference on Computer Vision. IEEE, pp. 2980–2988.

Holte, M.B., Tran, C., Trivedi, M.M., Moeslund, T.B., 2012. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. IEEE J. Sel. Top. Signal Process. 6, 538–552.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern. Part C 34, 334–352.

Huang, S., Gong, M., Tao, D., 2017. A coarse-fine network for keypoint localization. In: Proc. IEEE International Conference on Computer Vision, pp. 3028–3037.

2019. INRIA4D. URL http://4drepository.inrialpes.fr (accessed on 2019).

Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., Schiele, B., 2017. Arttrack: Articulated multi-person tracking in the wild. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 6457–6465.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B., 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: Proc. European Conference on Computer Vision. Springer, pp. 34–50.

Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2014. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. 36, 1325–1339.

Iqbal, U., Gall, J., 2016. Multi-person pose estimation with local joint-to-person associations. In: Proc. European Conference on Computer Vision. Springer, pp. 627–642.

Iqbal, U., Milan, A., Gall, J., 2017. Posetrack: Joint multi-person pose estimation and tracking. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2011-2020.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. In: Advances in Neural Information Processing Systems. pp. 2017–2025.

Jain, A., Tompson, J., Andriluka, M., Taylor, G.W., Bregler, C., 2013. Learning human pose estimation features with convolutional networks. arXiv preprint arXiv:1312.7302.

Jain, A., Tompson, J., LeCun, Y., Bregler, C., 2014. Modeep: A deep learning framework using motion features for human pose estimation. In: Proc. Asian Conference on Computer Vision. Springer, pp. 302–315.

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J., 2013. Towards understanding action recognition. In: Proc. IEEE International Conference on Computer Vision, pp. 3192–3199.

Jhuang, H., Garrote, H., Poggio, E., Serre, T., Hmdb, T., 2011. A large video database for human motion recognition. In: Proc. IEEE International Conference on Computer Vision, p. 6.

Ji, X., Liu, H., 2010. Advances in view-invariant human motion analysis: A review. IEEE Trans. Syst. Man Cybern. Part C 40, 13–24.

Johnson, S., Everingham, M., 2010. Clustered pose and nonlinear appearance models for human pose estimation. In: Proc. British Machine Vision Conference, p. 5.

Johnson, S., Everingham, M., 2011. Learning effective human pose estimation from inaccurate annotation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1465–1472.

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

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

Ju, S.X., Black, M.J., Yacoob, Y., 1996. Cardboard people: A parameterized model of articulated image motion. In: Proc. IEEE Conference on Automatic Face and Gesture Recognition, pp. 38–44.

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

Ke, L., Chang, M.C., Qi, H., Lyu, S., 2018. Multi-scale structure-aware network for human pose estimation. In: Proc. European Conference on Computer Vision, pp. 713-728.

2019. Kinect. URL https://developer.microsoft.com/en-us/windows/kinect (accessed on 2019).

Kocabas, M., Karagoz, S., Akbas, E., 2018. Multiposenet: Fast multi-person pose estimation using pose residual network. In: Proc. European Conference on Computer Vision. Springer, pp. 437–453.

Kreiss, S., Bertoni, L., Alahi, A., 2019. Pifpaf: Composite fields for human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 11977–11986.

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: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4704–4713.

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

Li, B., Chen, H., Chen, Y., Dai, Y., He, M., 2017a. Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network. In: Proc. IEEE International Conference on Multimedia and Expo Workshops, pp. 613–616.

Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., He, M., 2017b. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In: Proc. IEEE International Conference on Multimedia and Expo Workshops, pp. 601–604.

Li, B., Dai, Y., He, M., 2018a. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. Pattern Recognit. 83, 328–339.

Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W., 2019. Learning the depths of moving people by watching frozen people. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4521-4530.

Li, L., Fei-fei, L., 2007. What, where and who? classifying events by scene and object recognition. In: Proc. IEEE International Conference on Computer Vision, p. 6.

Li, B., He, M., Dai, Y., Cheng, X., Chen, Y., 2018b. 3d skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated cnn. Multimedia Tools Appl. 77, 22901–22921.

Li, C., Lee, G.H., 2019. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 9887–9895.

Li, S., Liu, Z.Q., Chan, A.B., 2014. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 482–489.

Li, B., Shen, C., Dai, Y., Hengel, A., He, M., 2015a. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1119–1127.

Li, S., Zhang, W., Chan, A.B., 2015b. Maximum-margin structured learning with deep networks for 3d human pose estimation. In: Proc. IEEE International Conference on Computer Vision, pp. 2848–2856.

Lifshitz, I., Fetaya, E., Ullman, S., 2016. Human pose estimation using deep consensus voting. In: Proc. European Conference on Computer Vision. Springer, pp. 246–260.

Lin, T.Y., Dollár, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125.

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

Liu, Z., Zhu, J., Bu, J., Chen, C., 2015. A survey of human pose estimation: the body parts parsing based methods. J. Vis. Commun. Image Represent. 32, 10–19.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

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

Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., Pang, J., Lin, L., 2018. Lstm pose machines. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 5207–5215.

Luvizon, D.C., Picard, D., Tabia, H., 2018. 2d/3d pose estimation and action recognition using multitask deep learning. In: Proc. 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.F., Pons-Moll, G., Black, M.J., 2019. Amass: Archive of motion capture as surface shapes. arXiv preprint arXiv:1904.03278.

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: Proc. European Conference on Computer Vision, pp. 601–617.

von Marcard, T., Pons-Moll, G., Rosenhahn, B., 2016. Human pose estimation from video and imus. IEEE transactions on pattern analysis and machine intelligence 38, 1533–1547.

Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017. A simple yet effective baseline for 3d human pose estimation. In: Proc. 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: Proc. IEEE International Conference on 3D Vision. IEEE, pp. 506–516.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C., 2019. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camer. arXiv:1907.00837.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C., 2018. Single-shot multi-person 3d body pose estimation from monocular rgb input. In: International Conference on 3D Vision, 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, 44.

Meredith, M., Maddock, S., et al., 2001. Motion Capture File Formats Explained, Vol. 211. Department of Computer Science, University of Sheffield, pp. 241–244.

Moeslund, T.B., Granum, E., 2001. A survey of computer vision-based human motion capture. Comput. Vis. Image Underst. 81, 231–268.

Moeslund, T.B., Hilton, A., Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. 104, 90–126.

Moeslund, T.B., Hilton, A., Krüger, L., 2011. Visual Analysis of Humans. Springer.

Moon, G., Chang, J.Y., Lee, K.M., 2019. Posefix: Model-agnostic general human pose refinement network. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7773–7781.

Moreno-Noguer, F., 2017. 3d human pose estimation from a single image via distance matrix regression. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1561–1570.

Newell, A., Huang, Z., Deng, J., 2017. Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems. pp. 2277–2287.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: Proc. 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 preprint arXiv:1801.07372.

Nie, X., Feng, J., Xing, J., Yan, S., 2018. Pose partition networks for multi-person pose estimation. In: Proc. European Conference on Computer Vision, pp. 684–699.

Nie, B.X., Wei, P., Zhu, S.C., 2017. Monocular 3d human pose estimation by predicting depth on joints. In: Proc. IEEE International Conference on Computer Vision, pp. 3447–3455.

Ning, G., Zhang, Z., He, Z., 2018. Knowledge-guided deep fractal neural networks for human pose estimation. IEEE Trans. Multimed. 20, 1246–1259.

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: Proc. IEEE International Conference on 3D Vision. IEEE, pp. 484–494.

Ouyang, W., Chu, X., Wang, X., 2014. Multi-source deep learning for human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2329–2336.

Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K., 2018. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Proc. European Conference on Computer Vision, pp. 269-286.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K., 2017. Towards accurate multi-person pose estimation in the wild. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4903–4911.

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

Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017. Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1263–1272.

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

Peng, X., Tang, Z., Yang, F., Feris, R.S., Metaxas, D., 2018. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2226–2234.

Perez-Sala, X., Escalera, S., Angulo, C., Gonzalez, J., 2014. A survey on model based approaches for 2d and 3d visual human pose recovery. Sensors 14, 4189–4210.

Pfister, T., Charles, J., Zisserman, A., 2015. Flowing convnets for human pose estimation in videos. In: Proc. IEEE International Conference on Computer Vision, pp. 1913–1921.

Pfister, T., Simonyan, K., Charles, J., Zisserman, A., 2014. Deep convolutional neural networks for efficient pose estimation in gesture videos. In: Proc. Asian Conference on Computer Vision. Springer, pp. 538–552.

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: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4929–4937.

Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J., 2015. Dyna: A model of dynamic human shape in motion. ACM Trans. Graph. 34, 120.

Popa, A.I., Zanfir, M., Sminchisescu, C., 2017. Deep multitask architecture for integrated 2d and 3d human sensing. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4714–4723.

Poppe, R., 2007. Vision-based human motion analysis: An overview. Comput. Vis. Image Underst. 108, 4–18.

Qammaz, A., Argyros, A., 2019. Mocapnet: Ensemble of snn encoders for 3d human pose estimation in rgb images. In: Proc. British Machine VIsion Conference.

Rafi, U., Leibe, B., Gall, J., Kostrikov, I., 2016. An efficient convolutional network for human pose estimation. In: Proc. British Machine Vision Conference, p. 2.

Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J.A., Sheikh, Y., 2014. Pose machines: Articulated pose estimation via inference machines. In: Proc. European Conference on Computer Vision. Springer, pp. 33–47.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99.

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

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

Rogez, G., Weinzaepfel, P., Schmid, C., 2017. Lcr-net: Localization-classification-regression for human pose. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3433–3441.

Rohrbach, M., Amin, S., Andriluka, M., Schiele, B., 2012. A database for fine grained activity detection of cooking activities. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1194–1201.

Sapp, B., Taskar, B., 2013. Modec: Multimodal decomposable models for human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3674–3681.

Sapp, B., Weiss, D., Taskar, B., 2011. Parsing human motion with stretchable models. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1281–1288.

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.

Shahroudy, A., Liu, J., Ng, T.T., Wang, G., 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., et al., 2012. Efficient human pose estimation from single depth images. IEEE Trans. Pattern Anal. Mach. Intell. 35, 2821–2840.

Sidenbladh, H., De la Torre, F., Black, M.J., 2000. A framework for modeling the appearance of 3d articulated figures. In: Proc. IEEE Conference on Automatic Face and Gesture Recognition, IEEE, pp. 368–375.

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, 4.

Sminchisescu, C., 2008. 3d human motion analysis in monocular video: techniques and challenges. In: Human Motion. Springer, pp. 185–211.

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

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition.

Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y., 2018. Integral human pose regression. In: Proc. European Conference on Computer Vision, pp. 529–545.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Tan, J., Budvytis, I., Cipolla, R., 2017. Indirect deep structured learning for 3d human body shape and pose prediction. In: Proc. British Machine Vision Conference.

Tang, Z., Peng, X., Geng, S., Wu, L., Zhang, S., Metaxas, D., 2018b. Quantized densely connected u-nets for efficient landmark localization. In: Proc. European Conference on Computer Vision, pp. 339–354.

Tang, W., Wu, Y., 2019. Does learning specific features for related parts help human pose estimation?. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1107–1116.

Tang, W., Yu, P., Wu, Y., 2018a. Deeply learned compositional models for human pose estimation. In: Proc. European Conference on Computer Vision, pp. 190–206.

Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P., 2016. Structured prediction of 3d human pose with deep neural networks. arXiv preprint arXiv:1605.05180.

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

2019. TheCaptury. URL https://thecaptury.com/ (accessed on 2019).

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

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C., 2015. Efficient object localization using convolutional networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656.

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.

Toshev, A., Szegedy, C., 2014. Deeppose: Human pose estimation via deep neural networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660.

Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J., 2017. Total capture: 3d human pose estimation fusing video and inertial sensors. In: Proc. British Machine Vision Conference, pp. 1–13.

Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C., 2018. Bodynet: Volumetric inference of 3d human body shapes. In: Proc. European Conference on Computer Vision, pp. 20-36.

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C., 2017. Learning from synthetic humans. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4627–4635.

2019. Vicon. URL https://www.vicon.com/ (accessed on 2019).

Vondrick, C., Patterson, D., Ramanan, D., 2013. Efficiently scaling up crowdsourced video annotation. Int. J. Comput. Vis. 101, 184–204.

Wang, M., Chen, X., Liu, W., Qian, C., Lin, L., Ma, L., 2018a. Drpose3d: Depth ranking in 3d human pose estimation. arXiv preprint arXiv:1805.08973.

Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S., 2018b. Rgb-d-based human motion recognition with deep learning: A survey. Comput. Vis. Image Underst. 171, 118–139.

Wang, Y., Tran, D., Liao, Z., 2011. Learning hierarchical poselets for human parsing. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1705–1712.

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

Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al., 2017. Ai challenger: A large-scale dataset for going deeper in image understanding. arXiv preprint arXiv:1711.06475.

Xiao, B., Wu, H., Wei, Y., 2018. Simple baselines for human pose estimation and tracking. In: Proc. European Conference on Computer Vision, pp. 466–481.

Yang, W., Li, S., Ouyang, W., Li, H., Wang, X., 2017. Learning feature pyramids for human pose estimation. In: Proc. IEEE International Conference on Computer Vision, pp. 1281–1290.

Yang, W., Ouyang, W., Li, H., Wang, X., 2016. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3073–3082.

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

Yang, Y., Ramanan, D., 2013. Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. 35, 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: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2148–2157.

Zhang, W., Zhu, M., Derpanis, K.G., 2013. From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proc. IEEE International Conference on Computer Vision, pp. 2248–2255.

Zhao, M., Li, T., Ab. Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., Katabi, D., 2018. Through-wall human pose estimation using radio signals. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7356–7365.

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

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

Zuffi, S., Black, M.J., 2015. The stitched puppet: A graphical model of 3d human shape and pose. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3537–3546.

Zuffi, S., Freifeld, O., Black, M.J., 2012. From pictorial structures to deformable structures. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3546–3553.