electronics

MDPI

Article

Deep Learning Methods for 3D Human Pose Estimation under Different Supervision Paradigms: A Survey

Dejun Zhang [1], Yiqi Wu [2,3,*], Mingyue Guo [4] and Yilin Chen [5]

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, China; zhangdejun@cug.edu.cn
[2] College of Computer Science, China University of Geosciences, Wuhan 430078, China
[3] Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430078, China
[4] College of Information and Engineering, Sichuan Agricultural University, Yaan 625014, China; guomingyue@stu.sicau.edu.cn
[5] School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China; yilinchen@wit.edu.cn
* Correspondence: wuyq@cug.edu.cn

Abstract: The rise of deep learning technology has broadly promoted the practical application of artificial intelligence in production and daily life. In computer vision, many human-centered applications, such as video surveillance, human-computer interaction, digital entertainment, etc., rely heavily on accurate and efficient human pose estimation techniques. Inspired by the remarkable achievements in learning-based 2D human pose estimation, numerous research studies are devoted to the topic of 3D human pose estimation via deep learning methods. Against this backdrop, this paper provides an extensive literature survey of recent literature about deep learning methods for 3D human pose estimation to display the development process of these research studies, track the latest research trends, and analyze the characteristics of devised types of methods. The literature is reviewed, along with the general pipeline of 3D human pose estimation, which consists of human body modeling, learning-based pose estimation, and regularization for refinement. Different from existing reviews of the same topic, this paper focus on deep learning-based methods. The learning-based pose estimation is discussed from two categories: single-person and multi-person. Each one is further categorized by data type to the image-based methods and the video-based methods. Moreover, due to the significance of data for learning-based methods, this paper surveys the 3D human pose estimation methods according to the taxonomy of supervision form. At last, this paper also enlists the current and widely used datasets and compares performances of reviewed methods. Based on this literature survey, it can be concluded that each branch of 3D human pose estimation starts with fully-supervised methods, and there is still much room for multi-person pose estimation based on other supervision methods from both image and video. Besides the significant development of 3D human pose estimation via deep learning, the inherent ambiguity and occlusion problems remain challenging issues that need to be better addressed.

Keywords: 3D human pose estimation; deep learning; unsupervised; semi-supervised; fully-supervised; weakly-supervised

Citation: Zhang, D.; Wu, Y.; Guo, M.; Chen, Y. Deep Learning Methods for 3D Human Pose Estimation under Different Supervision Paradigms: A Survey. *Electronics* **2021**, *10*, 2267. https://doi.org/10.3390/electronics10182267

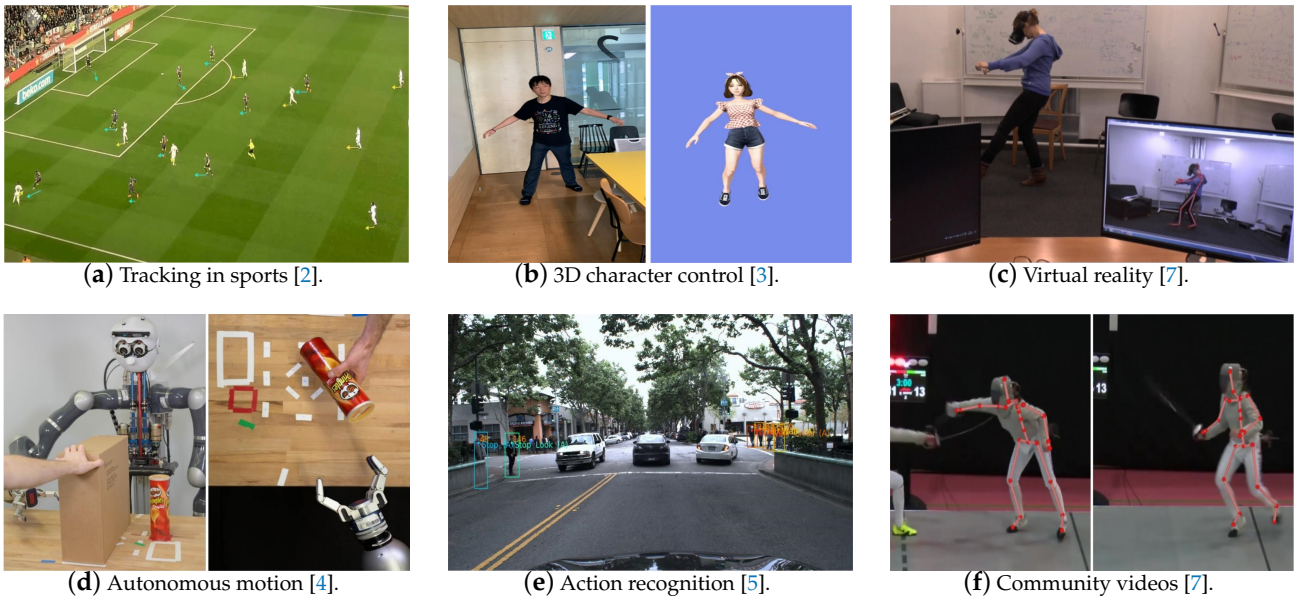
Academic Editor: Savvas A. Chatzichristofis

Received: 12 August 2021
Accepted: 6 September 2021
Published: 15 September 2021

1. Introduction

Human pose estimation is an important and widely concerned research topic in computer vision. Given an image or video input, 3D human pose estimation aims to predict the configuration of the human body. There are many real-life applications related with 3D human pose estimation [1–3], such as human-computer interaction [4], surveillance [5,6], virtual reality [7], video understanding [7], gesture recognition [8,9], etc., as shown in Figure 1. Many traditional methods have dealt with 3D human pose estimation well,

but deep learning-based 3D human pose estimation approaches have also aroused wide concern today, with the increasing development and the impressive performance of deep learning methods in varied computer vision tasks (e.g., image classification [10,11], action recognition [12,13], semantic segmentation [14,15]).



(**a**) Tracking in sports [2]. (**b**) 3D character control [3]. (**c**) Virtual reality [7].

(**d**) Autonomous motion [4]. (**e**) Action recognition [5]. (**f**) Community videos [7].

Figure 1. Some application related with 3D human estimation [2–5,7].

The main challenges in pose estimation based on deep learning method remain to be solved. (i) Lack of 3D training data: Since 3D manual annotation is expensive and time-consuming, there are not many 3D training data paired with 3D annotation. Nevertheless, the performance of deep learning methods is typically based on the scale of training data annotated with reference 3D poses. Many research studies have been proposed to resolve this issue, which will be introduced in the following content of this paper. (ii) Depth ambiguity: Depth ambiguity is an ill-posed problem in estimating 3D pose. This problem occurs because, despite given different 3D depths, joints of different poses may be projected to the same 2D location. Some works investigate this topic in different ways (e.g., References [16–18]). (iii) Occlusions: For single-person 3D pose estimation, self-occlusions (e.g., body parts occlusions) could greatly affect the performance of predicting 3D joint locations. In multi-person 3D pose estimation, besides self-occlusion, occlusion between different people is also an issue for predicting accurate 3D joint positions. There are plenty works aiming at overcoming this obstacle (e.g., References [19,20]).

Besides the above main challenges, there are other issues may be met during researching, such as variation in human appearance, articulated motion of the limbs, arbitrary camera viewpoints, and so on. Pioneer works propose various models to resolve these problems in distinct ways.

Chen et al. [21] is a recently published a survey of deep learning-based monocular human pose estimation, which contains both 2D and 3D human pose estimation. They discuss the 3D human pose estimation from two aspects, namely single-person 3D pose estimation and multi-person 3D pose estimation. For the single-person part, they divide relevant methods into model-free methods and model-based methods, while multi-person methods are not categorized. However, they describe the 3D human pose estimation very shortly. In contrast to their work, our paper discusses related research studies from distinct aspects and includes multi-view methodologies.

The most related survey with this work is offered by Sarafianos et al. [22], in which they provide an overview of predicting 3D human joint locations from image or video. On the one hand, they cover not only deep learning-based methods but also conventional

methods published in the 2008–2015 period. Instead, we only discuss research studies of 3D human pose estimation based on deep learning method published in the 2014–2021 period. On the other hand, research studies discussed in their work are categorized into generative approaches, discriminative approaches, and hybrid approaches. This kind of taxonomy does not provide readers an overview of this topic from the perspective of data, which is essential for the overview. In contrast to their work, the 3D human pose estimation methods are surveyed according to their supervision paradigms in our paper.

The main contributions of this paper are summarized as following:

- A comprehensive survey of 3D human pose estimation based on deep learning, covering the literature from 2014 to 2021, is proposed. The estimation methods are reviewed hierarchically according to diverse application targets, data sources, and supervision modes.
- Supervision mode is a crucial criterion for deep learning method taxonomy. Different supervision modes are based on entirely different philosophies and are applicable for diverse scenarios. Therefore, as far as we know, this survey is the first to classify 3D human pose estimation methods based on different supervision modes to show their characteristics, similarities, and differences.
- Eight datasets and three evaluation metrics that are widely used for 3D human pose estimation are introduced. Both quantitative results and qualitative analysis of the state-of-the-art methods mentioned in this review are presented based on the datasets mentioned above and metrics.
- As the prior processing and subsequent refinement phases, research studies of human body modeling and regularization are discussed, along with the pipeline of 3D human pose estimation.

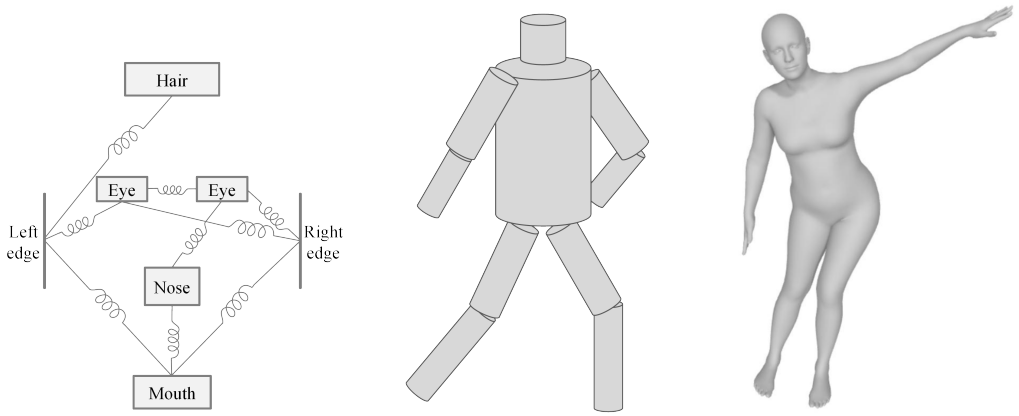
The rest of the paper is organized as follows. In Section 2, we describe the main human body models employed by some research studies to reconstruct 3D human pose. Section 3 states standard methods for both single-person and multi-person 3D pose estimation. Section 4 introduces the taxonomy of deep learning methods for 3D human pose estimation under different supervision paradigms. In Section 5, we present single-person 3D pose estimation approaches from still images and videos according to the proposed taxonomy above. Section 6 presents the recent advances and trends in multi-person 3D pose estimation from images and videos. In both sections, we discuss monocular and multi-view input approaches. In Section 7, some regularization methods found in the literature, which are employed to refine the performance, are discussed. Section 8 presents the introduction of commonly used datasets and evaluation measures for 3D human pose estimation, as well as offers experimental results summarized from the reviewed methods. Finally, Section 9 concludes the paper.

2. Human Body Models

The construction and description of human body model are basics for 3D human pose estimation. The Human body can be seen as a structural non-rigid object with particular characters. A mature human body model contains the representation of structure information, body shape, texture and so on. Existing human body models can be classified according to diverse representations together with various application scenarios. Generally, the widely adopted models are kinematics-based model, contour-based model, and volume-based model. We briefly describe the mentioned models which could be employed in some approaches that are studied in this paper below. More detailed introduction of human body models can be found in literature [21,23].

Kinematics-based models follow the skeletal structure, which can be divided into two types, the predefined model and the learned graph structure. They both can effectively represent a set of joint locations and limb orientations. A very prevailing graph model is Pictorial Structure Model (PSM) in 2D pose estimation, as described in Figure 2a, but pioneer works [24,25] extend PSM to 3D PSM. Although it is simple and flexible, there are

defects for the adoption of Kinematics-based model. An obvious one is the lost of body width and contour in the model for the lack of texture information.



(**a**) Pictorial structure model [26]. (**b**) Carboard model [27]. (**c**) SMPL model [28].

Figure 2. Three types of human body models [26–28].

Contour-based models, besides capturing the connections of body parts and the appearance, can also be learned by planar models. Usually, this class of model, with rough information of body width and contour, is widely adopted for earlier 2D pose estimation. For instance, Active Shape Models (ASMs) [29] can describe a human body and obtain the statistics while body contour deformation occurs. The cardboard model [27] represents body parts via rectangular shapes and contains the foreground color information of body, as shown in Figure 2b.

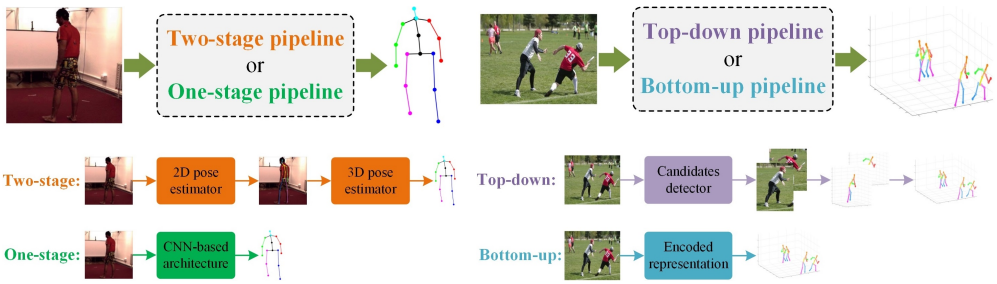
The volume-based model represents human body poses in a realistic way. Volumetric models compose of various geometric shapes, such as cylinders, conics, etc. [30], or triangulated meshes. The mesh models are suitable to deformable objects, so they fit for the non-rigid human body well. Classical volume-based models include Skinned Multi-Person Linear model (SMPL) [31], as depicted in Figure 2c, Shape Completion and Animation of People (SCAPE) [32], and a unified deformation model [33]. Especially, the SMPL is a statistical model that encodes the human subjects with two types of parameters: shape parameter and pose parameter. SMPL model has been used to estimate 3D human pose in many works [34,35].

3. Preliminary Issues

3.1. The Standard Methodologies of 3D Pose Estimation

In this work, we will discuss both single-person and multi-person 3D pose estimation. Here, we introduce standard methodologies for both classes which will be investigated further in subsequent sections.

- Single-person 3D pose estimation falls into two categories: two-stage and One-stage methods, as shown in Figure 3a. Two-stage methods involve two steps; first, we employ 2D keypoints detection models to obtain 2D joint locations, and then we lift 2D keypoints to 3D keypoints by deep learning methods. This kind of approach mainly suffers from inherent depth ambiguity in the second step, which is the key problem that many works aim to resolve. One-stage methods mean regressing 3D joint locations directly from a still image. These methods require much training data with 3D annotations, but manual annotation is costly and demanding.



(**a**) Single-person 3D pose estimation. (**b**) Multi-person 3D pose estimation.

Figure 3. The standard methodologies of single-person and multi-person 3D pose estimation. The input 2D image and the predicted 3D human pose in (**a**) come from the samples and GT in the Human3.6M Dataset [36], respectively. The input 2D image and the predicted 3D human pose in (**b**) are quoted from Reference [37].

- Multi-person 3D pose estimation is divided into two categories: top-down and bottom-up methods, as demonstrated in Figure 3b. Top-down methods first detect the human candidates and then apply single-person pose estimation for each of them. Bottom-up methods first detect all keypoints, followed by grouping them into different people. Two kinds of approaches have their advantages: top-down methods tend to be more accurate, while bottom-up methods tend to be relatively faster [38].

3.2. Input Paradigm

There are various kinds of input data, such as IMU data, images, depth data, point clouds, video sequences, 2D keypoints, etc. Nonetheless, we mainly discuss approaches that take still images or videos as input. In this survey, 3D human pose estimation methods are divided into two types based on the application scenarios, namely image-based methods and video-based methods.

- Image-based methods take static images as input, only taking spatial context into account, which differs from video-based methods.
- Video-based methods meet more challenges than image-based methods, such as temporal information processing, correspondence between spatial information and temporal information and motion changes in different frames, etc.

From the view of the application, video-based methods are more suitable to realize real-time 3D human pose estimation than image-based methods, such as diving pose estimation and scoring, as well as NBA player pose estimation.

3.3. Supervision Form

With the development of machine learning, researchers divide the supervision into different forms according to variable situations and distinct data types. For example, in the absence of ground-truth data, methods based on weak-supervision are better than fully-supervised ones. In this paper, the supervision falls into four categories: unsupervised, semi-supervised, fully-supervised, and weakly-supervised.

In general machine learning, the above four categories have standard definitions. However, definitions could change with the shift of the research scene. In this paper, on account of the research topic that this paper studies (i.e., 3D human pose estimation), specific definitions of the above four categories will subsequently be briefly discussed. We firstly identify each supervision form.

- Unsupervised methods do not require any multi-view image data, 3D skeletons, correspondences between 2D–3D points, or use previously learned 3D priors during training.
- Fully-supervised methods rely on large training sets annotated with ground-truth 3D positions coming from multi-view motion capture systems.

- Weakly-supervised methods realize the supervision through other existing or easily obtained cues rather than ground-truth 3D positions. The form of cues can be various, such as paired 2D ground-truth or camera parameters, etc.
- Semi-supervised methods use part of annotated data (e.g., 10% of 3D labels), which means labeled training data is scarce.

4. Taxonomy

Scenes with distinct limitations could yield various approaches with different supervisions. Both single-person and multi-person 3D pose estimation combined with different supervision forms could derive various branches, as described in Table 1.

Table 1. A taxonomy of deep learning methods for 3D human pose estimation.

Scenario	Input Paradigm	Degree of Supervision	3D Annotation [1]	Weak Annotation [2]	Scale [3]	Wild [4]
Single-Person (Section 5)	Image-based (Section 5.1)	Unsupervised (Section 5.1.1)	–	–	–	▬
		Semi-supervised (Section 5.1.2)	○	–	▪	▪
		Fully-supervised (Section 5.1.3)	○	–	▬	▪
		Weakly-supervised (Section 5.1.4)	–	○	▬	▬
	Video-based (Section 5.2)	Unsupervised (Section 5.2.1)	–	–	–	▬
		Semi-supervised (Section 5.2.2)	○	–	▪	▪
		Fully-supervised (Section 5.2.3)	○	–	▬	▪
		Weakly-supervised (Section 5.2.4)	–	○	▬	▬
Multi-Person (Section 6)	Image-based (Section 6.1)	Fully-supervised (Section 6.1.1)	○	–	▬	▪
		Weakly-supervised (Section 6.1.2)	–	○	▬	▬
	Video-based (Section 6.2)	Fully-supervised (Section 6.2.1)	○	–	▬	▪

[1] 3D pose ground truth data. [2] e.g., motion information, 2D annotations, 2D pose ground truth data, or multi-view images. [3] the scale of the annotated dataset. [4] generalization capabilities for in the wild images or video.

From Table 1, we summarize all the categories with four representative distinctions. The symbol of the fine dash "-" in column "3D annotation" means methods of the corresponding category do not need the 3D pose ground truth data, while the symbol of the hollow circle "○" represents the 3D pose ground truth data is required. The symbols have similar meanings in the "Weak annotation" column to whether the weak annotations, such as motion information, 2D annotations, 2D pose ground truth data, or multi-view images, are required for the corresponding category. The length of the thick line in the

column of "Scale" and "Wild" represents the scale of the required annotated dataset and the generalization capabilities for "in the wild" images or videos, respectively. Besides, the fine dash "-" in the column of "Scale" for unsupervised methods indicates that this kind of method does not need annotated dataset. For example, the fully-supervised methods for single-person 3D pose estimation from an image need annotated 3D ground truth data, but it does not require any weak annotations. The large scale of annotated data is essential for this kind of method, and the generalization capabilities for "in the wild" images are usually weak.

5. Single-Person 3D Pose Estimation

Single-person 3D pose estimation methods are categorized as image-based methods and video-based methods based on their input in this paper. We describe single-person 3D pose estimation from still images in Section 5.1 and single-person 3D pose estimation from videos in Section 5.2. Here, we first introduce image-based approaches and then video-based approaches.

5.1. Single-Person 3D Pose Estimation from Images

Images include monocular images and multi-view images. First, as for monocular 3D human pose estimation, the reconstruction of an arbitrary configuration of 3D points from a monocular RGB image has three characteristics that affect its performance: (i) it is a severely ill-posed problem because similar image projections can be derived from different 3D poses; (ii) it is an ill-conditioned problem since minor errors in the locations of the 2D body joints can have large consequences in the 3D space; and (iii) it suffers from high dimensionality. Existing approaches propose different solutions to compensate for these constraints and are discussed in the following paragraphs.

Second, finding the location correspondences between different views is a common challenge in handling multi-view images. How to exploit information between views and information of each view could be considered as a topic of balancing global and local features in multi-view scenes. Several milestone methods are illustrated in Figure 4 and will be described in the next subsections, according to different supervision formats.

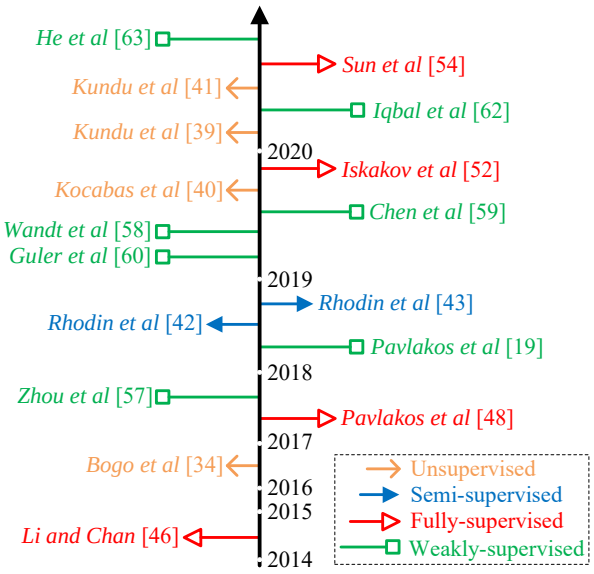


Figure 4. Chronological overview of some of the most relevant image-based single-person 3D pose estimation methods.

5.1.1. Unsupervised

Bogo et al. [34] describe the first method to estimate the 3D human pose, as well as shape, simultaneously, from a single still image. They use the DeepCut method to predict 2D joint locations, and then fit the SMPL model [31] to the joints by minimizing an objective function. Recently, Reference [39] presented a novel unsupervised monocular 3D human pose estimation method, which accesses no ground-truth data or weak cues. They only use two kinds of prior information: the kinematic skeletal structure with bone-length ratios in a fixed canonical scale and a set of 20 synthetically-rendered SMPL models. By carefully designing a kinematic structure preservation pipeline (forward-kinematic transformation, camera projection transformation, and spatial-map transformation) with training objectives, the authors realize 3D human pose estimation in an unsupervised framework.

Self-supervised methods which can also solve the issue, i.e., deficiency of 3D data, have become popular in recent years. Self-supervised methods are a form of unsupervised learning where the data provides the supervision. Kocabas et al. [40] introduce Epipolar-Pose, a self-supervised approach that creates its 3D supervision using epipolar geometry and estimated 2D poses. The architecture of EpipolarPose comprises two branches, in which the 3D pose generated by the lower branch using epipolar geometry is used as a training signal for the CNN in the upper branch. EpipolarPose simultaneously takes two consecutive images while training, but it takes a single image during inference as a monocular method. In Reference [41], a differentiable and modular self-supervised learning paradigm is proposed. The whole network contains an encoder and a decoder. The encoder takes an image as input and outputs three disentangled representations then the decoder projects encoded representations and synthesizes FG human image with 2D part segmentation. Consistency constraints across different network outputs and image pairs are designed to self-supervise the network.

5.1.2. Semi-Supervised

Semi-supervised methods are relatively direct to solve the lack of 3D annotated data. An approach introduced by Rhodin et al. [42] utilizes multi-views to replace most of the annotations and then uses a small set of images with ground-truth poses to predict correct poses. Rhodin et al. further extend their method in Reference [43]. They use unlabeled multi-view images for geometry-aware representation learning at first. Then, the mapping between the representation and actual 3D poses are learned in a supervised way with a little labeled data, which forms a semi-supervised way. Their experiments show that fully supervised state-of-the-art methods outperform their method when large amounts of annotated training data are available but suffer greatly when little labeled data are available. The devised model trained with fewer annotated training data obtained better performance in real scenarios. To be noted is that, in some works, other supervision forms can change to semi-supervised ways. For example, Reference [41] utilizes 10% of the full training set with direct 3D pose supervision referring to the semi-supervised setting.

Moreover, the semi-supervised setting proves to be comparable performance against some previous fully supervised approaches by experiment. Reference [40] also has a variant with a small set of ground-truth available as the semi-supervised setting to further illustrate the performance. To bridge the gap between the in-the-wild data without 3D ground-truth annotations and constrained data with 3D ground-truth annotations, Yang et al. [44] introduce an adversarial learning framework. The designed network distills the 3D human pose structures learned from constrained images to in-the-wild domains. Taking images from a 3D dataset and a 2D dataset as inputs, the 3D pose estimator (the generator) predicts 3D human pose that is fed to the multi-source discriminator distinguishing ground-truth poses from the predicted poses. During the learning process, only the 3D dataset has ground-truth 3D poses, while the 2D dataset only has 2D annotations.

Qiu et al. [45] propose to recover absolute 3D human poses from multi-view images by incorporating multi-view geometric priors. They first use a cross-view fusion scheme to estimate multi-view 2D poses, which are adopted to recover the 3D pose by a recursive

PSM. The introduce fusion neural network tackles the issue of acquiring corresponding locations of multiple views. Unlike PSM, the proposed RPSM could refine the estimation of 3D pose gradually.

5.1.3. Fully-Supervised

Li and Chan [46] first propose to apply deep neural networks to 3D human pose estimation from monocular images. They design a framework consisting of two types of tasks: (1) a joint regression task and (2) a joint point task. Both tasks use corresponding ground-truth data to obtain optimization objectives and are trained jointly within a multi-task learning framework.

Tekin et al. [47] introduce a deep learning regression architecture for structured prediction of 3D human pose, which directly regresses from an input RGB image to a 3D human pose. They combine traditional CNNs with auto-encoders for structured learning, which preserves the power of CNNs while accounting for dependencies. First, they encode the dependencies between human joints by learning a mapping of a 3D human pose to a high-dimensional latent space with a denoising auto-encoder. Second, they adopt a CNN to regress the image to a high-dimensional representation and, finally, re-project the latent pose estimates to the original pose space by the last decoding layers in the CNN.

Pavlakos et al. [48] are the first to cast 3D human pose estimation as a 3D keypoint localization problem in a voxel space using an end-to-end learning paradigm. Given an image, their designed network outputs a dense 3D volume with coarse-to-fine supervision, which is validated to be effective in experiments.

The method of Mehta et al. [49] implements a three-stage framework to handle the 3D human pose estimation. They first extract the actor bounding box computed from 2D joint heatmaps, obtained with a CNN called 2DPoseNet, then predict 3D pose from the bounding box cropped input with the 3DPoseNet, and, finally, align 3D to 2D pose to obtain global 3D pose coordinates.

Transfer learning is also exploited to leverage the highly relevant mid- and high-level features learned on the readily available in-the-wild 2D datasets [50,51], along with annotated 3D pose datasets. Besides, they introduce the MPI-INF-3DHP dataset to complement other datasets, through extensive appearance and pose variation, by using marker-less annotation.

For multi-view scenarios, Iskakov et.al [52] provide two solutions for multi-view 3D human pose estimation based on the idea of learnable triangulation. The first solution inputs a set of images to a CNN to obtain 2D joint heatmaps and camera-joint confidences and then passes 2D positions together with the confidences to the algebraic triangulation module to obtain the 3D pose. While the second solution differs from the first solution in the process steps of 2D joint heatmaps, here, 2D joint heatmaps are unprojected into volumes with subsequent aggregation to a fixed size volume, and then the volume is passed to a 3D CNN to output 3D heatmaps, which are used to obtain the 3D pose by computing soft-argmax. Loss functions for both solutions use ground-truth 3D data as supervision, namely fully supervised.

Remelli et al. [53] propose a light weight multi-view and image-based method for 3D pose acquisition. At first, the 3D pose representation which is independent to camera is learned. Then, the location of 2D joint is estimated via the pose representation. Finally, the 2D joints are ascended to 3D joints by the adoption of Direct Linear Transform (DLT). A fusion technique, namely canonical fusion, is proposed for efficient and joint reasoning about diverse views by the geometry information exploitation in the latent space. Besides, the DLT is realized through the deep learning method, which makes it much faster than standard implementation.

The multi-view-based 3D pose estimation methods [52,53] that rely on multi-view RGB images can effectively solve the problems encountered in the monocular human pose estimation method. In actual application scenarios, it is easier to obtain monocular RGB images than multi-view RGB images. However, monocular RGB images cannot be used

directly in multi-view methods, limiting practical applications. Sun et al. [54] propose a novel end-to-end 3D pose estimation network for monocular 3D human pose estimation to tackle this issue. They first present a multi-view pose generator to predict multi-view 2D poses from the 2D poses in a single view. Second, they adopt a simple but effective data augmentation method for generating multi-view 2D pose annotations. Finally, they employ a graph convolutional network to infer a 3D pose from multi-view 2D poses.

In the work of Luvizon et al. [55], they propose a multi-task framework for jointly estimating 2D or 3D human poses from monocular images. The architecture is composed of prediction blocks, downscaling and upscaling units, and simple connections. Middle layers of prediction blocks compute a set of body joint probability maps and depth maps, which then are used to estimate 3D joints by a differentiable and non-parametrized function. Moreover, the problem of 3D human pose estimation is decoupled to 2D pose estimation, depth estimation, and concatenation of the intermediate parts.

5.1.4. Weakly-Supervised

Tome et al. [56] introduce an end-to-end method that jointly addresses the 2D location detection and full 3D human pose estimation from a single RGB image. They propose a novel multi-stage CNN architecture that learns to combine the image appearance with the geometric 3D skeletal information encoded in a novel pre-trained model of 3D human pose. There are six stages in the CNN, and each stage consists of four distinct components, predicting CNN-based belief-maps, lifting 2D belief-maps into 3D, projected 2D pose belief maps, and a 2D fusion layer. The novel layer, based on a probabilistic 3D model of human pose, takes the predicted 2D landmarks of the sixth stage as input and lifts them into 3D space for the final estimate.

The 3D human pose estimation for in the wild scenario is generally considered to be a challenging task for the reality that training data is not readily available. Zhou et al. [57] address the problem by utilizing mixed 2D and 3D labels in a unified neural network consisting of a 2D module and a 3D module. They design an interesting 3D geometric constraint as a weakly-supervised loss for 2D data. This method realizes transferring 3D annotation from indoor images to in-the-wild images that could be a big step for investigation weakly-supervised 3D human pose estimation.

Kanazawa et al. [35] propose an end-to-end solution for 3D human pose estimation in monocular images based on the SMPL model [31]. They aim at learning a mapping from image pixels directly to model parameters, which means inferring 3D mesh parameters directly from image features. The framework could be divided into two parts. Images are first passed through a convolutional encoder and an iterative 3D regression module to infer 3D parameters that minimize the joint re-projection error. Then, 3D parameters are also sent to the discriminator to distinguish whether the inferred parameters are from a real human shape and pose. Considering the absence of accurate 3D ground truth, Pavlakos et al. [18] introduce weak supervision in the form of ordinal depth relations of the joints. The provided weak supervision are pairwise ordinal depth relations which describe relationships between any two joints.

In the work of Wandt and Rosenhahn [58], they argue that 3D poses are regressed from the input distribution (2D poses) to the output distribution (3D poses). The devised framework consists of three parts: a pose and camera estimation network, a critic network, and a re-projection network. It can be trained without 2D–3D correspondences and unknown cameras in a weakly supervised way. Some researchers employ multi-view data to deal with the data issue. A novel weakly-supervised framework is provided by Chen et al. [59]. They propose an encoder-decoder architecture to learn the geometry-aware 3D representation for the 3D human pose with multi-view data and only existing 2D labels as supervision. The architecture contains three parts: image-skeleton mapping module to obtain 2D skeleton maps, view synthesis module for learning the geometry representation, and representation consistency constraint to refine the representation. The overall framework is trained in a weakly-supervised manner.

Guler and Kokkinos [60] introduce HoloPose, a method for holistic monocular single-person 3D pose estimation, which is based on the SMPL model [31]. Firstly, they provide a part-based architecture for parameter regression, then exploit DensePose [61] and 3D joint estimation to increase the accuracy of the 3D pose, and, finally, introduce two technical modifications which simplify modeling. They address the lack of ground-truth data interestingly; they utilize DensePose as a format of supervision.

Iqbal et al. [62] propose a weakly-supervised approach which use unlabeled data for 3D poses learning without the requirement of 3D labels or paired 2D–3D annotations. Given an image, the end-to-end architecture first estimates the scale normalized 2.5D pose. Then, the 2.5D pose is used to subsequently reconstruct the scale normalized 3D pose. In the training phase, the network only consumes images with 2D pose labels and a number of unlabeled data. Taking an image as input, the designed ConvNet predicts per-voxel likelihoods for every joint in the 3D space, which are used to predict 3D Gaussian centered at the 3D location of each joint. The method is trained in an end-to-end manner.

To resolve issues of common two-step approaches for single-person 3D pose estimation based on multi-view images, He et al. [63] introduce a differentiable method, called "epipolar transformer". The epipolar transformer consists of two components: the epipolar sampler and the feature fusion module. The epipolar sampler samples features along the corresponding epipolar line in the source view after taking a point in the reference view as input, while the feature fusion module fuses the feature from the source view with the feature in the reference view, where authors give two different fusion architectures, i.e., Identity Gaussian architecture and Bottleneck Embedded Gaussian architecture. The epipolar transformer enables a 2D detector to gain 3D information in the intermediate layers of the 2D detector itself, which helps to improve 2D pose estimation.

5.2. Single-Person 3D Pose Estimation from Videos

Compared to scenes of images, more issues are posed by videos. Generally, when the target person is too small or too large, or if the motion is too fast/slow, it would affect the performance of single-person 3D pose estimation from videos. In addition, occlusion is a tough problem to handle in videos. Several milestone methods are illustrated in Figure 5 and will be described in the next subsections, according to different supervision formats.

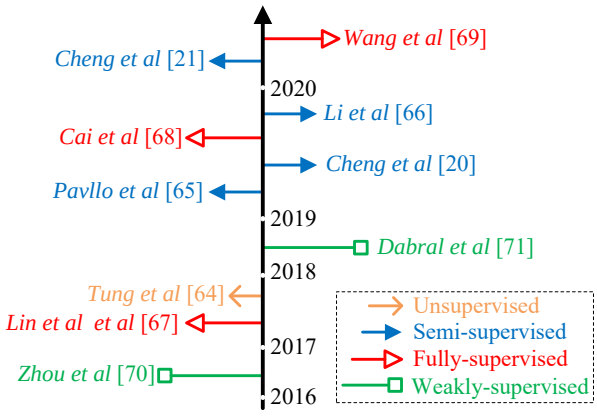


Figure 5. Chronological overview of some of the most relevant video-based single-person 3D pose estimation methods.

5.2.1. Unsupervised

Given a video sequence and a set of 2D body joint heatmaps, a network predicting the body parameters for the SMPL 3D human mesh model is implemented by Tung et al. [64]. Self-supervision coming from 3D-to-2D rendering, consistency checks against 2D estimates of keypoints, segmentation, and optical flow solves the deficiency of 3D annotations. Specifically, self-supervision consists of three components: keypoint re-projection error, motion re-projection error, and segmentation re-projection error.

5.2.2. Semi-Supervised

In the work of Pavllo et al. [65], they propose an efficient architecture based on dilated temporal convolutions on 2D keypoint trajectories for 3D human pose estimation in video. To investigate circumstances that there is only a small amount of labeled data, a novel scheme with semi-supervised training is proposed by using unlabeled video data. For semi-supervision, two objectives are jointly optimized: one is the supervised loss where the ground-truth 3D poses available, and the other is an autoencoder loss. The 3D poses are first transferred to 2D, and then the loss can be obtained by the comparison with the input. In their experiments related to semi-supervision, they consider various subsets of the training set to validate the model performance. Based on the work of Pavllo et al. [65], Cheng et al. [19] introduce an end-to-end framework which can be trained in a semi-supervised way. They explicitly handle the occlusion issue by feeding incomplete 2D keypoints to the 2D and 3D temporal convolutional networks (2D and 3D TCNs), which enforce temporal smoothness to produce a complete 3D pose. As no dataset has the occlusion ground-truth labels, they propose a novel "Cylinder Man Model" to generate pairs of virtual 3D joints and 2D keypoints with explicit occlusion labels. Because of the 3D joint ground-truth is not always available or sufficient, they design a loss considering both cases to enable semi-supervised training.

To exploit monocular videos to complement the training datasets for single-image 3D human pose estimation tasks, Li et al. [66] propose an automatic method that collects accurate annotations of human motions from monocular videos. They first pre-train a baseline model with a small set of annotations and then optimize the output of the base model to perform as pseudo annotations for further training. Considering errors caused by the optimized predictions, they introduce the weighted fidelity pseudo-supervision term that weighs the pseudo-supervision term in the loss function by the confidence score of each prediction. Training with both annotated 3D and only 2D data prevails in single-person 3D pose estimation from monocular videos.

Cheng et al. [20] introduce a spatio-temporal network for robust 3D human estimation. In the beginning, they use the High-Resolution Network (HR-Net) to exploit multi-scale spatial features to produce one heatmap for each joint. Then, they encode heatmaps into a latent space to obtain latent features, which are used with different strides in temporal convolutional networks [65] to predict 3D poses. Besides, they utilize a discriminator to check the pose validity spatio-temporally. Because of the occlusion issue, some keypoints are masked to simulate the occlusions in real world.

5.2.3. Fully-Supervised

Unlike image-based models, video-based methods are supervised by a long sequence of 3D pose. Lin et al. [67] introduce a novel Recurrent 3D Sequence Machine (RPSM) model to recurrently integrate rich spatial and temporal long-range dependencies using a multi-stage sequential refinement. They first predict 3D human poses for all monocular frames, and then sequentially refine them with multi-stage recurrent learning with the proposed unified architecture with 2D poses, feature adaption, and 3D pose recurrent modules. The whole RPSM model is fully-supervised by the loss which is defined as the Euclidean distances between the prediction for all joints and ground-truth.

In the work of Cai et al. [68], they provide a novel graph-based method to tackle the problem of the 3D human body and 3D hand pose estimation from a short sequence. Specifically, they construct a spatial-temporal graph on skeleton sequences and design a hierarchical "local-to-global" architecture with graph convolutional operations. The "local-to-global" architecture, aiming at learning multi-scale features but from the graph-based representations, is conceptually similar to the Stack Hourglass network for 2D pose estimation. The whole framework is trained in a fully-supervised way.

Recently, Wang et al. [69] propose a new pipeline for estimating 3D poses from consecutive 2D poses with an introduced motion loss. The pipeline first obtains a 2D skeleton sequence from a pose estimator, then structures 2D skeletons by a spatial-temporal

graph, and predicts 3D locations via U-shaped Graph Convolution Networks (UGCN). Additionally, to better consider the similarity of temporal structure between the estimated pose sequence and the ground-truth data, a motion loss evaluating motion reconstruction quality is proposed.

5.2.4. Weakly-Supervised

The method proposed by Zhou et al. [70] addresses the problem of 3D human pose estimation from a monocular video in a distinctive way. They take a video as input, and then obtain 2D heatmaps throughout a CNN, which are then used to estimate the 3D human pose, along with the 3D pose dictionary. Different from pioneer works that perform 3D human pose estimation in two disjoint steps, they cast the 2D joint locations as latent variables that could help to handle the 2D estimation uncertainty with prior 3D geometry. Besides, the EM algorithm is applied to estimate 3D human pose from 2D heatmaps and the pose dictionary.

Based on the work of Zhou et al. [57], Dabral et al. [71] introduce a temporal motion model for 3D human pose estimation from videos. The model is composed of three parts: Structure-Aware PoseNet (SAP-Net), Temporal PoseNet (TP-Net), and Skeleton fitting. They propose two novel anatomical loss functions, namely illegal-angle and symmetry loss, which allow training in a weakly-supervised way with 2D annotated data samples.

6. Multi-Person 3D Pose Estimation

Compared to single-person 3D pose estimation, multi-person 3D pose estimation is a more challenging problem due to the much larger state space and partial occlusions, as well as across view ambiguities, when not knowing the identity of the humans in advance. Multi-person 3D pose estimation methods could also be partitioned into two categories: image-based methods and video-based methods; the methods are shown in Figure 6, in the chronological view. Undoubtedly, multi-person scenes present more complex issues compared with single-person scenes. Image-based multi-person 3D pose estimation needs to handle detections of different peoples and occlusions between people, not to mention body joints estimation itself, while multi-person videos involve different occlusions and challenges posed by videos. We describe multi-person 3D pose estimation from images and from video in Sections 6.1 and 6.2, respectively.

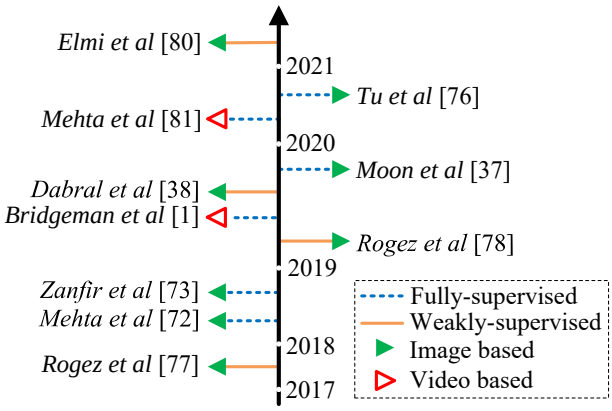


Figure 6. Chronological overview of some of the most relevant multi-person 3D pose estimation methods.

6.1. Multi-Person 3D Pose Estimation from Still Images

6.1.1. Fully-Supervised

Mehta et al. [72] propose the first multi-person dataset of real person images with 3D ground truth. The dataset consists of the training set (MuCo-3DHP) and the test set (MuPoTS-3D). Besides, a CNN-based single-shot multi-person pose estimation method is implemented which forms a novel multi-person 3D pose-map to jointly predict 2D and 3D

locations of all persons in the scene. The network is supervised by the loss function which measures the distance between the estimated poses and the ground-truth data.

An interesting approach from Zanfir et al. [73] addresses the problem of multi-person 3D pose estimation in images by presenting a feed-forward, multi-task architecture, namely MubyNet. The model is feed-forward and supports simultaneous multi-person localization, as well as 3D pose and shape estimation. MubyNet mainly contains three parts: deep volume encoding, limb scoring, and 3D pose and shape estimation. Multiple datasets with different types of annotations are used for training MubyNet, such as COCO [74], Human80K [36], and CMU Panoptic datasets [75].

The first fully learning-based, camera distance-aware top-down approach for multi-person is introduced by Moon et al. [37]. To recover the absolute camera-centered coordinates of multiple persons' keypoints, they design a system that consists of DetectNet, RootNet, and PoseNet. The DetectNet detects a human bounding box of each person in the input image that then is input to the RootNet and PoseNet. After receiving the root of the human and root-relative 3D pose from RootNet and PoseNet, respectively, the final absolute 3D pose can be obtained by simple back-projection. All three submodules use corresponding ground-truth data to supervise their training.

Tu et al. [76] propose VoxelPose to estimate 3D poses of multiple people from multiple camera views. VoxelPose directly operates in the 3D space and, therefore, avoids making incorrect decisions in each camera view. Specifically, they first predict 2D pose heatmaps for all views (the stage one), then warp the heatmaps to a common 3D space and construct a feature volume fed into a Cuboid Proposal Network (CPN) to localize all people instances (the stage two), and, finally, construct a finer-grained feature volume and estimate a 3D pose for each proposal (the final stage). They propose two novel networks, i.e., CPN and Pose Regression Network (PRN). Specifically, CPN is used to coarsely localize all people in the scene by predicting some 3D cuboid proposals from the 3D feature volume, while PRN takes the finer-grained feature volume as input to estimate a detailed 3D pose for each proposal.

6.1.2. Weakly-Supervised

Rogez et al. [77] introduce an end-to-end architecture, named Localization Classification Regression Network (LCR-Net), which detects 2D and 3D poses simultaneously in still images. As its name implies, the network has three modules: namely the pose proposal generator, a classifier, and a regressor. Taking an image as input, the network first extracts candidate regions using an RPN and obtains pose proposals in the module of localization, then scores these pose proposals by a classifier, and, finally, regresses pose proposals through a regressor. The final output contains both 2D and 3D poses per image that are collected by aggregating similar pose proposals, in terms of location and 3D pose. To train the network, they provide pseudo ground-truth 3D poses which are inferred from 2D annotations using a Nearest Neighbor (NN) search performed on the annotated joints. Rogez et al. further extend their work in Reference [78], where they improve the performance of their method in four ways: (1) addition of synthetic data obtained by rendering the SMPL 3D model to augment training data, (2) adding an iterative process to the network to further refine regression and classification results, (3) the use of improved alignment, and (4) a ResNet [79] backbone.

In the work of Dabral et al. [38], they introduce a two-stage approach that first estimates the 2D keypoints in every Region of Interest (RoI) and then lift the estimated 2D joints to 3D. For the first stage, they adopt the augmented Mask-RCNN with a shallow HourGlass Network to output 2D heatmaps, which are then inputted to a 3D pose module to regress the root-relative 3D joint coordinates in the second stage. With the existing multi-person 2D pose datasets and single-person 3D pose datasets, they do not require any multi-person datasets for training. In this case, their designed model could perform in-the-wild multi-person 3D pose estimation without the requirement of costly 3D annotations in the in-the-wild setting.

Elmi et al. [80] present the first complete bottom-up approach performing multi-person 3D pose estimation from a few calibrated camera views. They first process each 2D view separately with two modules, namely the 2D pose backbone and the reduction module, then they aggregate feature maps into a 3D input representation of the scene through the re-projection layer, and, finally, adopt volumetric processing to obtain 3D human pose estimation. Volumetric processing consists of three modules: the volumetric network predicting a set of 3D Gaussians and a set of 3D Part Affinity Fields (PAFs), the sub-voxel joint detection applied to each joint map to obtain a set of the peak for each joint type, and the skeleton decoder predicting the final poses. The post-processing strategy leads to a sub-voxel localization, addressing the issue of a quantized 3D space.

6.2. Multi-Person 3D Pose Estimation from Video

6.2.1. Fully-Supervised

Bridgeman et al. [1] propose a novel approach to tackle the problem of 3D multi-person pose estimation and tracking from multi-view video. First, they correct errors in the output of the pose detector, then they apply a label to every 2D pose ensuring consistency between views, and, finally, use the labeled 2D poses to produce a sequence of tracked 3D skeletons. A greedy search is employed to find correspondences between 2D poses in different camera views which are proved to be efficient. The algorithm is capable of tracking players in crowded soccer scene with missing and noisy detections.

Mehta et al. [81] provide a real-time approach for multi-person 3D pose estimation, which is robust to difficult occlusion both by other people and objects. The architecture consists of three stages, with the first two stages performing per-frame local and global reasoning, respectively, and the third stage implementing temporal reasoning across frames. Specifically, stage one infers 2D pose and 3D pose encoding through a new SelecSLS Net, stage two runs in parallel for each detected person and reconstructs the complete 3D poses, and the final stage uses the kinematic skeleton fitting to provide temporal stability, location, and a joint angle parameterization. As for training, the first two stages both use 2D and 3D ground-truth data, while the final stage is not CNN-based.

7. Regularization

In recent research studies on 3D human pose estimation, different types of regularizations or constraints are needed to produce accurate 3D joints. Here, we divide these regularizations into two types: kinematic constraints and loss of regularization terms.

Kinematic constraints: Wandt and Rosenhahn [58] add the kinematic chain space (KCS) proposed by Wandt et al. [82]. They develop a KCS layer with a successive fully connected network which is added in parallel to the fully connected path. The KCS matrix is a representation of a human pose containing joint angles and bone lengths and can be computed by only two matrix multiplications. By adding the discriminator network with the KCS matrix, it enforces symmetry between the left and right sides of the body.

Loss regularization terms: A 3D geometric constraint induced loss is introduced in the work of Zhou et al [57]; it helps to effectively regularize depth prediction in the absence of ground-truth depth label. The geometric loss is based on the fact that ratios between bone lengths remain relatively fixed in a human skeleton. To penalize 3D pose predictions that drift too far away from the initial ones from training, Rhodin et al. [42] provide a regularization loss. Cheng et al. [19] design a pose regularization scheme by adding occlusion constraints in the loss function. The principle is that, if a missing keypoint is occluded, we have a reasonable explanation for its failure of detection or unreliability; otherwise, it is less likely to be missed by the 2D keypoint estimator and should be penalized. Geometric loss usually acts as a regularizer and penalizes the pose configurations that violate the consistency of bone-length ratio priors. However, some geometric loss designs ignore certain other strong anatomical constraints of the human body. Hence, Dabral et al. [71] provide constraints containing joint-angle limits, left-symmetry of the human body, and bone-length ratio priors, which lead to better 3D pose configuration.

In the work of Chen et al. [83], they design a joint shift loss to ensure consistency between predicted bone lengths and directions.

8. Existing Datasets and Evaluation Metrics

8.1. Evaluation Datasets

For single-person 3D pose estimation, several datasets provide different views for subjects with different poses, which enables monocular and multi-view research studies. In some datasets, in-the-wild scenes are also presented for research studies. Datasets for multi-person 3D pose estimation provide indoor scenes or in-the-wild scenes, but the variety of data is less abundant than single-person datasets. Below, both single-person datasets and multi-person datasets will be discussed.

Here, we first introduce several benchmark datasets for single-person 3D pose estimation.

- The Human3.6M Dataset [36]. This dataset contains 3.6 million video frames with the corresponding annotated 3D and 2D human joint positions, from 11 actors. Each actor performs 15 different activities captured from 4 unique camera views. It allows us to capture data from 15 sensors (4 digital video cameras, 1 time-of-flight sensor, and 10 motion cameras), using hardware and software synchronization. The capture area was about 6 m $\times$ 5 m, and within it, we had roughly 4 m $\times$ 3 m of effective capture space, where subjects were fully visible in all video cameras.
- HumanEva-I & II Datasets [84]. The ground truth annotations of both datasets were captured using ViconPeak's commercial MoCap system. The HumanEva-I dataset contains 7-view video sequences (4 grayscales and 3 colors) synchronized with 3D body poses. There are four subjects with markers on their bodies performing six common actions (e.g., walking, jogging, gesturing, throwing and catching a ball, boxing, combo) in a 3 m $\times$ 2 m capture area. HumanEva-II is an extension of HumanEva-I dataset for testing, which contains two subjects performing the action combo.
- The MPI-INF-3DHP Dataset [49]. It is a recently proposed 3D dataset, including constrained indoor and complex outdoor scenes. It records eight actors (four females and four males) performing eight activities (e.g., walking/standing, exercise, sitting, crouch/reach, on the floor, sports, miscellaneous) from 4 camera views. It is collected with a markerless multi-camera MoCap system in both indoor and outdoor scenes.
- The 3DPW Dataset [85]. It is a very recent dataset, captured mostly in outdoor conditions, using IMU sensors to compute pose and shape ground truth. It contains nearly 60,000 images containing one or more people performing various actions in the wild.

Several benchmark datasets for multi-person 3D pose estimation are introduced in the following paragraphs.

- The CMU Panoptic Dataset [75]. This dataset consists of 31 full HD and 480 VGA video streams from synchronized cameras, with a speed of 29.97 FPS and a total duration of 5.5 h. It provides high-quality 3D pose annotations, which are computed using all the camera views. It is the most complete, open, and free-to-use dataset that can be used for 3D human pose estimation tasks. However, considering that they released annotations quite recently, most works in literature use it only for qualitative evaluations [86], or for single-person pose detection [52], discarding multi-person scenes.
- The Campus Dataset [87]. This dataset uses three cameras to capture the interaction of three people in an outdoor environment. This small training data tend to be over-fitting. As for evaluation metrics, this dataset adopts the Average 3D Point Error and the Percentage of Correct Parts for evaluation.
- The Shelf Dataset [87]. This dataset includes 3D annotated data and adopts the same metrics as the Campus dataset. However, it is more complex than Campus dataset. In the dataset, four people close together are disassembling a shelf, and five calibrated cameras are deployed in the scene, but each view is heavily occluded.

- The MuPoTS-3D Dataset [72]. This dataset is a recently released multi-person 3D pose test dataset containing 20 test sequences captured by an unmarked motion capture system in five indoor and fifteen outdoor environments. Each sequence contains various activities for 2–3 people. The evaluation metric is the 3D PCK (the percentage of correct keypoints within a radius of 15 cm) of all annotated personnel. When there is a missed detection, all joints of the missed person will be considered incorrect. Another evaluation mode is to evaluate only the detected joints.

8.2. Evaluation Metrics

Mean Per Joint Position Error (MPJPE) is the most broadly used measure to evaluate 3D human pose estimation performance. It computes 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 collection 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. The MPJPE in protocol 1 of Human3.6M is calculated after aligning the depths of the root joints, which is called NMPJPE. The MPJPE in HumanEva-I and protocol 2 and 3 of Human3.6M are calculated after aligning predictions and ground truth with a rigid transformation using Procrustes Analysis, also called reconstruction error P-MPJPE or PA-MPJPE.

$$E_{MPJPE} = \frac{1}{N}\sum_{i=1}^{N}\left\|P_i - \hat{P}_i\right\|_2, \tag{1}$$

where N represents the number of poses, P_i represents the predicted 3D pose, and $\hat{P}_i$ represents the ground-truth 3D pose.

Percentage of Correct Keypoints (PCK) and Area Under the Curve (AUC) are introduced to be the metric for 3D pose evaluation in Reference [49], as they are adopted in 2D evaluation on dataset of MPII. The metric of PCK is also known as 3DPCK. The correct point refers to the point that drops into a predefined threshold. Then, the proportion of correct points is counted to obtain the PCK, while AUC is obtained by a set of PCK thresholds. Generally, the threshold of PCK is set to 150 mm, close to half of an adult's head. Except for evaluating the coordinates, the difference between estimated body shape and the ground truth can be also used a metric, namely Mean Per-vertex.

Percentage of Correct Parts (PCP) is used as the metric to evaluate the estimated 3D pose. Specifically, for each ground-truth 3D pose, it finds the closest pose estimation and computes the percentage of correct parts. However, this metric does not penalize false positive pose estimations. Some works also extend the Average Precision (AP) to the multi-person 3D pose estimation task.

8.3. Summary on Datasets

In this section, we present the state-of-the-art results for methods mentioned in this review on several datasets, such as Human3.6M, MPI-INF-3DHP, MuPoTS-3D, Shelf, and Campus datasets. In order to aid scholars in finding most of the state-of-the-art methods, we collected the links of all the papers discussed in this section and the link of their source code (if available). All the links are posted in our community (https://github.com/djzgroup/HumanPoseSurvey (accessed on 1 August 2021)).

Table 2 presents results of representative methods proposed between 2014 and 2021. It is noted that weakly-supervised methods and fully-supervised methods account for a large proportion, which implies that learning without any supervision is still difficult to obtain good results. Hence, unsupervised deep learning-based research studies need more investigation. From the aspect of views, multi-view methods attract more attention than they used to, although monocular studies take up a greater proportion. With the development of human body models, more and more research studies consider not only the 3D joint locations but also the shape of the human body, which may benefit prediction 3D human pose. It can be seen that performance on the Human3.6M dataset has been significantly improved due to the fast development of deep learning. As for the evaluation

metric, namely average MPJPE, leaving little room for improvement that indicates the future research direction on single-person 3D pose estimation may focus more on the use of data, real-time mobile deployment, and innovation of methods than accuracy only.

The MPI-INF-3DHP dataset is a relatively new dataset that has fewer SOTA methods results. Table 3 enumerates SOTA methods which perform on this dataset. From Table 3, weak supervision is more prevailing compared to other supervision forms, which may be owing to the complex outdoor training data without ground-truth labels. Monocular methods are still hot in the research area of single-person 3D pose estimation. Obviously, according to the 3DPCK metric, some approaches already have achieved excellent results, but it does not mean that the existing methods handle the challenge posed by this dataset well.

There are not many datasets providing training data with annotated paired data for multi-person 3D pose estimation. The MuPoTS-3D dataset is a recently released multi-person 3D pose test set, which has the quantitative evaluation metric for multi-person 3D pose estimation methods. Table 4 collects results for SOTA methods, and it can be observed that monocular methods prefer to experiment on this dataset with different supervisions. However, more views will be helpful for multi-person 3D pose estimation. Meanwhile, it could be seen that investigating unsupervised ways for predicting multi-person 3D pose needs more time to develop as complex challenges are existing in the scene of multi-person.

Table 2. Summary of the state-of-the-arts methods on Human3.6M dataset.

METHOD	MPJPE (mm)	SUPERVISION	TYPE
Li et al. [46] (2014)	132.2	fully-supervised	monocular
Zhou et al. [70] (2016)	113.0	weakly-supervised	monocular
Tekin et al. [47] (2016)	116.8	fully-supervised	monocular
Tome et al. [56] (2017)	88.4	weakly-supervised	monocular
Zhou et al. [57] (2017)	64.9	weakly-supervised	monocular
Kanazawa et al. [35] (2017)	56.8	weakly-supervised	monocular
Pavllo et al. [65] (2018)	46.8	semi-supervised	monocular
Pavlakos et al. [18] (2018)	44.7	weakly-supervised	monocular
Cheng et al. [19] (2019)	42.9	semi-supervised	monocular
RepNet [58] (2019)	50.9	weakly-supervised	monocular
HoloPose [60] (2019)	50.4	weakly-supervised	monocular
Luvizon et al. [55] (2020)	48.6	fully-supervised	monocular
Cheng et al. [20] (2020)	40.1	semi-supervised	monocular
Sun et al. [54] (2020)	35.8	fully-supervised	monocular
Martinez et al. [88] (2017)	87.3	fully-supervised	multi-view
Kocabas et al. [40] (2019)	76.6	semi-supervised	multi-view
Iskakov et.al [52] (2019)	17.7	fully-supervised	multi-view
Remelli et al. [53] (2020)	30.2	fully-supervised	multi-view
He et al. [63] (2020)	19.0	weakly-supervised	multi-view

Table 3. Summary of the state-of-the-arts methods on MPI-INF-3DHP datasets. '-' means no published data.

METHOD	3DPCK	AUC	MPJPE (mm)	SUPERVISION	TYPE
Mehta et al. [49] (2017)	76.5	40.8	-	fully-supervised	monocular
Zhou et al. [57] (2017)	69.2	32.5	-	weakly-supervised	monocular
Yang et al. [44] (2018)	69.0	32.0	-	semi-supervised	monocular
MargiPose [89] (2019)	95.1	62.2	60.1	weakly-supervised	monocular
SPIN [90] (2019)	92.5	55.6	67.5	weakly-supervised	monocular
RepNet [58] (2019)	82.5	58.5	97.8	weakly-supervised	monocular
Chen et al. [91] (2019)	71.1	36.3	-	unsupervised	monocular
Chen et al. [83] (2021)	87.9	54.0	78.8	fully-supervised	monocular
Wang et al. [92] (2019)	71.2	33.8	-	weakly-supervised	multi-view

Table 4. Summary of the state-of-the-art multi-person 3D pose estimation methods on MuPoTS-3D dataset.

METHOD	Total 3DPCK (a)	Total 3DPCK (b)	SUPERVISION	TYPE
LCR-Net [77] (2017)	53.8	62.4	weakly-supervised	monocular
ORPM [72] (2018)	65.0	69.8	fully-supervised	monocular
Moon et al. [37] (2019)	81.8	82.5	fully-supervised	monocular
XNect [81] (2020)	70.4	75.8	semi-supervised	monocular
Lcr-net++ [78] (2019)	70.6	74.0	weakly-supervised	monocular
HG-RCNN [38] (2019)	71.3	74.2	weakly-supervised	monocular

The Campus dataset and the Shelf dataset differ in the scene aspect; the former was collected indoors, while the latter was collected outdoors. Tables 5 and 6 present results of these two datasets, respectively. Noted that several methods are not involved for they are not in the scope we discuss in this work. Listing their results here is to give readers a more clear view of these two datasets. It is apparent that multi-view approaches favor to perform on these two datasets, and several works (e.g., References [76,86]) could extraordinarily handle challenges of two datasets. The Campus and the Shelf datasets are relatively classical and original datasets for multi-person 3D pose estimation, where almost every kind of supervision forms has been studied, except semi-supervision. From the PCP metric, the performance on these two datasets is gradually reaching the upper limit. However, from the timeline, these two datasets are outstanding benchmarks to test multi-person 3D pose estimation methods.

Table 5. Summary of the state-of-the-art multi-person 3D pose estimation methods on Campus dataset. '-' means no published data.

METHOD	Average PCP	AP	SUPERVISION	TYPE
3DPS [87] (2014)	75.8	-	fully-supervised	multi-view
Belagiannis et al. [25] (2014)	78.0	-	weakly-supervised	multi-view
Belagiannis et al. [24] (2015)	84.5	-	fully-supervised	multi-view
Ershadi et al. [93] (2018)	90.6	-	weakly-supervised	multi-view
Dong et al. [86] (2019)	96.3	61.6	weakly-supervised	multi-view
Bridgeman et al. [1] (2019)	92.6	-	unsupervised	multi-view
Tu et al. [76] (2020)	96.7	91.4	fully-supervised	multi-view
Chen et al. [94] (2020)	96.6	-	unsupervised	multi-view

Table 6. Summary of the state-of-the-art multi-person 3D pose estimation methods on Shelf dataset. '-' means no published data.

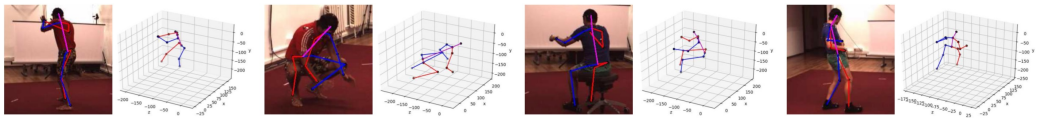
METHOD	Average PCP	AP	SUPERVISION	TYPE
3DPS [87] (2014)	71.4	-	fully-supervised	multi-view
Belagiannis et al. [25] (2014)	76.0	-	weakly-supervised	multi-view
Belagiannis et al. [24] (2015)	77.5	-	fully-supervised	multi-view
Ershadi et al. [93] (2018)	88	-	weakly-supervised	multi-view
Dong et al. [86] (2019)	96.9	75.4	weakly-supervised	multi-view
Bridgeman et al. [1] (2019)	96.7	-	unsupervised	multi-view
Tu et al. [76] (2020)	97.0	80.5	fully-supervised	multi-view
Light3DPose [80] (2021)	89.8	-	weakly-supervised	multi-view
Chen et al. [94] (2020)	96.8	-	unsupervised	multi-view

After comprehensively analyzing the quantitative results of multitudes of state-of-the-art methods, the qualitative results of some representative works are presented and subsequently analyzed.

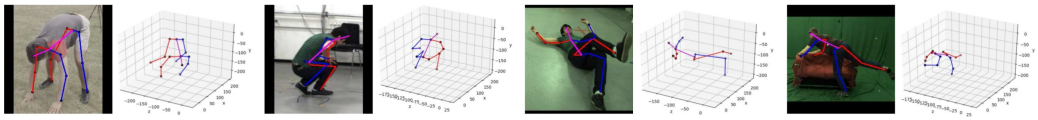
Figure 7 shows the performance of the method proposed by Zhou et al. [57] for single-person 3D pose estimation on several datasets. The impressive performance is not only achieved on the benchmark dataset Human3.6M, as shown in Figure 7a, but also on MPI-INF-3DHP, the dataset without 3D ground truth, with the correct extraction of

complex poses, as shown in Figure 7b. Moreover, the method can also process in the wild images, such as outdoor sports images, as shown in Figure 7c. Besides, the method obtains remarkable performance for the self-occlusions, as shown in the figure.

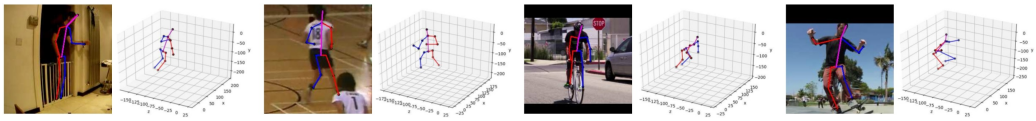
As many research studies focus on multi-person 3D pose estimation in recent years, the research of Dong et al. [86] is a representative one. As shown in Figure 8, their work estimates the 3D poses of multiple people accurately on the video-based datasets. As a multi-person pose estimation method, it tackles the problem of self-occlusions and deals with occlusions between crowded people well.



(**a**) Single-person 3D pose estimation on Human3.6M dataset.

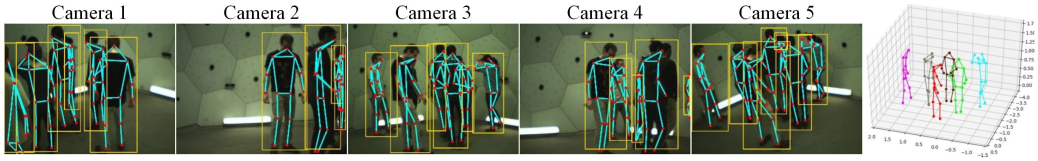


(**b**) Single-person 3D pose estimation on MPI-INF-3DHP dataset.

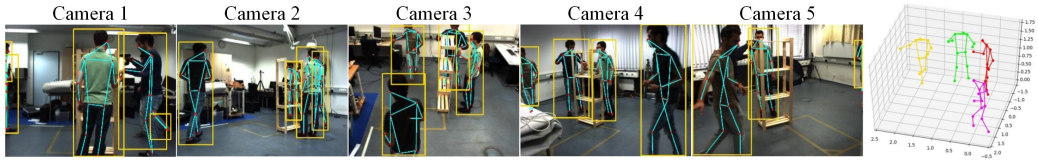


(**c**) Single-person 3D pose estimation on MPII dataset.

Figure 7. Qualitative results from three different single-person datasets. We show the 2D pose on the original image and the 3D pose from Reference [57].



(**a**) Multi-person 3D pose estimation on CMU panoptic dataset.



(**b**) Multi-person 3D pose estimation on Shelf dataset.

Figure 8. Qualitative results from two different multi-person datasets. The five leftmost columns, respectively, show the 2D bounding boxes and pose obtained from five different cameras. The rightmost column shows the estimated 3D poses from Reference [86].

9. Conclusions

Three-dimensional human pose estimation is a hot research area in computer vision that developed with the blooming of deep learning. This paper has presented a contemporary survey of the state-of-the-art methods for 3D human pose estimation, including single-person and multi-person 3D pose estimation, which both are categorized into two parts based on the input signal: images and videos. In each category, we group the methods based on their supervision forms, which are full supervision, semi-supervision, weak supervision, self-supervision, and no supervision. A comprehensive taxonomy and performance comparison of these methods have been presented. We first reviewed different human body

models. Then, the standard pipelines and taxonomy of single-person and multi-person 3D pose estimation are briefly introduced. Next, single-person and multi-person 3D pose estimation approaches are separately described. Finally, we introduce datasets widely used in 3D human pose estimation. Notwithstanding the great development of 3D human pose estimation with deep learning, there remain some pending challenges and gap between research and practical applications, such as occlusions between body parts and between different people, inherent ambiguity, and crowded people.

Author Contributions: D.Z. conceived of and designed the algorithm and the experiments. D.Z. and M.G. analyzed the data. D.Z. and M.G. wrote the manuscript. Y.W. supervised the research. Y.W. and Y.C. provided suggestions for the proposed method and its evaluation and assisted in the preparation of the manuscript. All authors approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National National Science Foundation of China (Grant No. 61802355 and 61702350) and the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing(KLIGIP-2019B04).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bridgeman, L.; Volino, M.; Guillemaut, J.Y.; Hilton, A. Multi-Person 3D Pose Estimation and Tracking in Sports. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 15–20 June 2019; pp. 2487–2496.
2. Arbués-Sangüesa, A; Martín, A; Fernández, J; Rodríguez. Always Look on the Bright Side of the Field: Merging Pose and Contextual Data to Estimate Orientation of Soccer Players. *arXiv* **2020**, arXiv:2003.00943.
3. Hwang, D.H.; Kim, S.; Monet, N.; Koike, H.; Bae, S. Lightweight 3D Human Pose Estimation Network Training Using Teacher-Student Learning. *arXiv* **2020**, arXiv:2001.05097.
4. Kappler, D.; Meier, F.; Issac, J.; Mainprice, J.; Cifuentes, C.G.; Wüthrich, M.; Berenz, V.; Schaal, S.; Ratliff, N.; Bohg, J. Real-time Perception meets Reactive Motion Generation. *arXiv* **2017**, arXiv:1703.03512.
5. Hayakawa, J.; Dariush, B. Recognition and 3D Localization of Pedestrian Actions from Monocular Video. *arXiv* **2020**, arXiv:2008.01162.
6. Andrejevic, M. Automating surveillance. *Surveill. Soc.* **2019**, *17*, 7–13. [CrossRef]
7. Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.P.; Xu, W.; Casas, D.; Theobalt, C. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *arXiv* **2017**, arXiv:1705.01583.
8. Li, G.; Tang, H.; Sun, Y.; Kong, J.; Jiang, G.; Jiang, D.; Tao, B.; Xu, S.; Liu, H. Hand gesture recognition based on convolution neural network. *Clust. Comput.* **2019**, *22*, 2719–2729. [CrossRef]
9. Skaria, S.; Al-Hourani, A.; Lech, M.; Evans, R.J. Hand-gesture recognition using two-antenna doppler radar with deep convolutional neural networks. *IEEE Sens. J.* **2019**, *19*, 3041–3048. [CrossRef]
10. Sun, Y.; Xue, B.; Zhang, M.; Yen, G.G.; Lv, J. Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification. *IEEE Trans. Cybern.* **2020**, *50*, 3840–3854. [CrossRef]
11. Sun, X.; Xv, H.; Dong, J.; Zhou, H.; Chen, C.; Li, Q. Few-shot Learning for Domain-specific Fine-grained Image Classification. *IEEE Trans. Ind. Electron.* **2020**, *68*, 3588–3598. [CrossRef]
12. Han, F.; Zhang, D.; Wu, Y.; Qiu, Z.; Wu, L.; Huang, W. Human Action Recognition Based on Dual Correlation Network. In *Asian Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 211–222.
13. Zhang, D.; He, L.; Tu, Z.; Zhang, S.; Han, F.; Yang, B. Learning motion representation for real-time spatio-temporal action localization. *Pattern Recognit.* **2020**, *103*, 107312. [CrossRef]
14. Fu, J.; Liu, J.; Wang, Y.; Zhou, J.; Wang, C.; Lu, H. Stacked deconvolutional network for semantic segmentation. *IEEE Trans. Image Process.* **2019**.. [CrossRef]
15. Zhang, D.; He, F.; Tu, Z.; Zou, L.; Chen, Y. Pointwise geometric and semantic learning network on 3D point clouds. *Integr. Comput. Aided Eng.* **2020**, *27*, 57–75. [CrossRef]
16. Duvenaud, D.K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R.P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; pp. 2215–2223.
17. Yang, W.; Ouyang, W.; Li, H.; Wang, X. End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3073–3082.
18. Pavlakos, G.; Zhou, X.; Daniilidis, K. Ordinal Depth Supervision for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7307–7316.

19. Cheng, Y.; Yang, B.; Wang, B.; Wending, Y.; Tan, R. Occlusion-Aware Networks for 3D Human Pose Estimation in Video. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 723–732.
20. Cheng, Y.; Yang, B.; Wang, B.; Tan, R.T. 3D Human Pose Estimation Using Spatio-Temporal Networks with Explicit Occlusion Training. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 10631–10638.
21. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [CrossRef]
22. Sarafianos, N.; Boteanu, B.; Ionescu, B.; Kakadiaris, I.A. 3D human pose estimation: A review of the literature and analysis of covariates. *Comput. Vis. Image Underst.* **2016**, *152*, 1–20. [CrossRef]
23. Gong, W.; Zhang, X.; Gonzàlez, J.; Sobral, A.; Bouwmans, T.; Tu, C.; Zahzah, E.h. Human pose estimation from monocular images: A comprehensive survey. *Sensors* **2016**, *16*, 1966. [CrossRef]
24. Belagiannis, V.; Amin, S.; Andriluka, M.; Schiele, B.; Navab, N.; Ilic, S. 3D pictorial structures revisited: Multiple human pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1929–1942. [CrossRef]
25. Belagiannis, V.; Wang, X.; Schiele, B.; Fua, P.; Ilic, S.; Navab, N. Multiple human pose estimation with temporally consistent 3D pictorial structures. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 742–754.
26. Fischler, M.A.; Elschlager, R.A. The representation and matching of pictorial structures. *IEEE Trans. Comput.* **1973**, *100*, 67–92. [CrossRef]
27. Ju, S.X.; Black, M.J.; Yacoob, Y. Cardboard people: A parameterized model of articulated image motion. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, Killington, VT, USA, 14–16 October 1996; pp. 38–44.
28. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. Bashirov, R.; Ianina, A.; Iskakov, K.; Kononenko, Y.; Strizhkova, V.; Lempitsky, V.; Vakhitov, A. Real-time RGBD-based Extended Body Pose Estimation. *arXiv* **2021**, arXiv:2103.03663.
29. Cootes, T.; Taylor, C.; Cooper, D.; Graham, J. Active Shape Models-Their Training and Application. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59. [CrossRef]
30. Sidenbladh, H.; De la Torre, F.; Black, M.J. A framework for modeling the appearance of 3D articulated figures. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, 28–30 March 2000; pp. 368–375.
31. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *Acm Trans. Graph. (TOG)* **2015**, *34*, 1–16. [CrossRef]
32. Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; Davis, J. SCAPE: Shape completion and animation of people. *Acm Trans. Graph. (TOG)* **2005**, *24*, 408–416. [CrossRef]
33. Joo, H.; Simon, T.; Sheikh, Y. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8320–8329.
34. Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; Black, M.J. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 561–578.
35. Kanazawa, A.; Black, M.J.; Jacobs, D.W.; Malik, J. End-to-End Recovery of Human Shape and Pose. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7122–7131.
36. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef]
37. Moon, G.; Chang, J.Y.; Lee, K.M. Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation From a Single RGB Image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 10132–10141.
38. Dabral, R.; Gundavarapu, N.B.; Mitra, R.; Sharma, A.; Ramakrishnan, G.; Jain, A. Multi-person 3D human pose estimation from monocular images. In Proceedings of the International Conference on 3D Vision (3DV), Quebec City, QC, Canada, 16–19 September 2019; pp. 405–414.
39. Kundu, J.N.; Seth, S.; Rahul, M.; Rakesh, M.; Radhakrishnan, V.B.; Chakraborty, A. Kinematic-Structure-Preserved Representation for Unsupervised 3D Human Pose Estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11312–11319.
40. Kocabas, M.; Karagoz, S.; Akbas, E. Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1077–1086.
41. Kundu, J.N.; Seth, S.; Jampani, V.; Rakesh, M.; Babu, R.V.; Chakraborty, A. Self-Supervised 3D Human Pose Estimation via Part Guided Novel Image Synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6152–6162.
42. Rhodin, H.; Meyer, F.; Spörri, J.; Müller, E.; Constantin, V.; Fua, P.; Katircioglu, I.; Salzmann, M. Learning Monocular 3D Human Pose Estimation from Multi-view Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8437–8446.
43. Rhodin, H.; Salzmann, M.; Fua, P. Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 765–782.

44. Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; Wang, X. 3D Human Pose Estimation in the Wild by Adversarial Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5255–5264.
45. Qiu, H.; Wang, C.; Wang, J.; Wang, N.; Zeng, W. Cross View Fusion for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 27 October–2 November 2019; pp. 4341–4350.
46. Li, S.; Chan, A.B. 3D human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 332–347.
47. Bugra, T.; Isinsu, K.; Mathieu, S.; Vincent, L.; Pascal, F. Structured Prediction of 3D Human Pose with Deep Neural Networks. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016; pp. 130.1–130.11.
48. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1263–1272.
49. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3D human pose estimation in the wild using improved cnn supervision. In Proceedings of the international conference on 3D vision (3DV), Qingdao, China, 10–12 October 2017; pp. 506–516.
50. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
51. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the International Conference on Machine learning, Lille, France, 7–9 July 2015; pp. 1180–1189.
52. Iskakov, K.; Burkov, E.; Lempitsky, V.; Malkov, Y. Learnable Triangulation of Human Pose. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 7717–7726.
53. Remelli, E.; Han, S.; Honari, S.; Fua, P.; Wang, R. Lightweight Multi-View 3D Pose Estimation through Camera-Disentangled Representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6040–6049.
54. Sun, J.; Wang, M.; Zhao, X.; Zhang, D. Multi-View Pose Generator Based on Deep Learning for Monocular 3D Human Pose Estimation. *Symmetry* **2020**, *12*, 1116. [CrossRef]
55. Luvizon, D.; Picard, D.; Tabia, H. Multi-task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2752–2764. [CrossRef] [PubMed]
56. Tome, D.; Russell, C.; Agapito, L. Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5689–5698.
57. Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; Wei, Y. Towards 3D Human Pose Estimation in the Wild: A Weakly-supervised Approach. *arXiv* **2017**, arXiv:1704.02447.
58. Wandt, B.; Rosenhahn, B. RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7774–7783.
59. Chen, X.; Lin, K.Y.; Liu, W.; Qian, C.; Lin, L. Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10887–10896.
60. Güler, R.A.; Kokkinos, I. HoloPose: Holistic 3D Human Reconstruction In-The-Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10876–10886.
61. Güler, R.A.; Neverova, N.; Kokkinos, I. DensePose: Dense Human Pose Estimation in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7297–7306.
62. Iqbal, U.; Molchanov, P.; Kautz, J. Weakly-Supervised 3D Human Pose Learning via Multi-view Images in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5243–5252.
63. He, Y.; Yan, R.; Fragkiadaki, K.; Yu, S.I. Epipolar Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7779–7788.
64. Tung, H.Y.F.; Tung, H.W.; Yumer, E.; Fragkiadaki, K. Self-supervised learning of motion capture. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5242–5252.
65. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7745–7754.
66. Li, Z.; Wang, X.; Wang, F.; Jiang, P. On Boosting Single-Frame 3D Human Pose Estimation via Monocular Videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 2192–2201.
67. Lin, M.; Lin, L.; Liang, X.; Wang, K.; Cheng, H. Recurrent 3D Pose Sequence Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5543–5552.

68. Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.J.; Yuan, J.; Thalmann, N.M. Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 2272–2281.
69. Wang, J.; Yan, S.; Xiong, Y.; Lin, D. Motion Guided 3D Pose Estimation from Videos. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 764–780.
70. Zhou, X.; Zhu, M.; Leonardos, S.; Derpanis, K.G.; Daniilidis, K. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4966–4975.
71. Dabral, R.; Mundhada, A.; Kusupati, U.; Afaque, S.; Sharma, A.; Jain, A. Learning 3D Human Pose from Structure and Motion. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 679–696.
72. Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Sridhar, S.; Pons-Moll, G.; Theobalt, C. Single-shot multi-person 3D pose estimation from monocular rgb. In Proceedings of the International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 120–130.
73. Zanfir, A.; Marinoiu, E.; Zanfir, M.; Popa, A.I.; Sminchisescu, C. Deep network for the integrated 3D sensing of multiple people in natural images. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 8420–8429.
74. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
75. Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; Sheikh, Y. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3334–3342.
76. Tu, H.; Wang, C.; Zeng, W. VoxelPose: Towards Multi-camera 3D Human Pose Estimation in Wild Environment. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 197–212.
77. Rogez, G.; Weinzaepfel, P.; Schmid, C. LCR-Net: Localization-Classification-Regression for Human Pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1216–1224.
78. Rogez, G.; Weinzaepfel, P.; Schmid, C. Lcr-net++: Multi-person 2D and 3D pose detection in natural images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1146–1161. [CrossRef]
79. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
80. Elmi, A.; Mazzini, D.; Tortella, P. Light3DPose: Real-time Multi-Person 3D Pose Estimation from Multiple Views. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 2755–2762. [CrossRef]
81. Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Elgharib, M.; Fua, P.; Seidel, H.P.; Rhodin, H.; Pons-Moll, G.; Theobalt, C. XNect: Real-Time Multi-Person 3D Motion Capture with a Single RGB Camera. *ACM Trans. Graph.* **2020**, *39*. [CrossRef]
82. Wandt, B.; Ackermann, H.; Rosenhahn, B. A Kinematic Chain Space for Monocular Motion Capture. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 31–47.
83. Chen, T.; Fang, C.; Shen, X.; Zhu, Y.; Chen, Z.; Luo, J. Anatomy-aware 3D Human Pose Estimation with Bone-based Pose Decomposition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**. [CrossRef]
84. Sigal, L.; Balan, A.O.; Black, M.J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. J. Comput. Vis.* **2010**, *87*, 4. [CrossRef]
85. Von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, pp. 614–631.
86. Dong, J.; Jiang, W.; Huang, Q.; Bao, H.; Zhou, X. Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views. *arXiv* **2019**, arXiv:1901.04111.
87. Belagiannis, V.; Amin, S.; Andriluka, M.; Schiele, B.; Navab, N.; Ilic, S. 3D Pictorial Structures for Multiple Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1669–1676.
88. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A Simple Yet Effective Baseline for 3D Human Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2659–2668.
89. Nibali, A.; He, Z.; Morgan, S.; Prendergast, L. 3D human pose estimation with 2D marginal heatmaps. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 1477–1485.
90. Kolotouros, N.; Pavlakos, G.; Black, M.; Daniilidis, K. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 2252–2261.
91. Chen, C.H.; Tyagi, A.; Agrawal, A.; Drover, D.; Rohith, M.; Stojanov, S.; Rehg, J.M. Unsupervised 3D Pose Estimation With Geometric Self-Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5707–5717.
92. Wang, L.; Chen, Y.; Guo, Z.; Qian, K.; Lin, M.; Li, H.; Ren, J.S. Generalizing Monocular 3D Human Pose Estimation in the Wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27 October–2 November 2019; pp. 4024–4033.

93. Ershadi-Nasab, S.; Noury, E.; Kasaei, S.; Sanaei, E. Multiple human 3D pose estimation from multiview images. *Multimed. Tools Appl.* **2018**, *77*, 15573–15601. [CrossRef]
94. Chen, L.; Ai, H.; Chen, R.; Zhuang, Z.; Liu, S. Cross-View Tracking for Multi-Human 3D Pose Estimation at over 100 FPS. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3279–3288.