Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.020857

REVIEW

Overview of 3D Human Pose Estimation

Jianchu Lin[1,2], Shuang Li[3], Hong Qin[3,4], Hongchang Wang[3], Ning Cui[6], Qian Jiang[7], Haifang Jian[3,*] and Gongming Wang[5,*]

[1]Huaiyin Institute of Technology, Huai'an, 223000, China

[2]Jiangsu Outlook Shenzhou Big Data Technology Co., Ltd., Nanjing, 210002, China

[3]Institute of Semiconductors, Chinese Academy of Sciences, Beijing, 100083, China

[4]Center of Materials Science and Optoelectronics Engineering & School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing, 100049, China

[5]Inspur Software Group Company, Ltd., Jinan, 250104, China

[6]China Great Wall Industry Corporation, Beijing, 100054, China

[7]China Great Wall Industry Corporation Navigation Co., Ltd., Beijing, 100144, China

*Corresponding Authors: Haifang Jian. Email: jhf@semi.ac.cn; Gongming Wang. Email: gongmingwang@126.com

Received: 16 December 2021 Accepted: 26 April 2022

ABSTRACT

3D human pose estimation is a major focus area in the field of computer vision, which plays an important role in practical applications. This article summarizes the framework and research progress related to the estimation of monocular RGB images and videos. An overall perspective of methods integrated with deep learning is introduced. Novel image-based and video-based inputs are proposed as the analysis framework. From this viewpoint, common problems are discussed. The diversity of human postures usually leads to problems such as occlusion and ambiguity, and the lack of training datasets often results in poor generalization ability of the model. Regression methods are crucial for solving such problems. Considering image-based input, the multi-view method is commonly used to solve occlusion problems. Here, the multi-view method is analyzed comprehensively. By referring to video-based input, the human prior knowledge of restricted motion is used to predict human postures. In addition, structural constraints are widely used as prior knowledge. Furthermore, weakly supervised learning methods are studied and discussed for these two types of inputs to improve the model generalization ability. The problem of insufficient training datasets must also be considered, especially because 3D datasets are usually biased and limited. Finally, emerging and popular datasets and evaluation indicators are discussed. The characteristics of the datasets and the relationships of the indicators are explained and highlighted. Thus, this article can be useful and instructive for researchers who are lacking in experience and find this field confusing. In addition, by providing an overview of 3D human pose estimation, this article sorts and refines recent studies on 3D human pose estimation. It describes kernel problems and common useful methods, and discusses the scope for further research.

KEYWORDS

3D human pose estimation; monocular camera; deep learning; multi-view; indicator

1 Introduction

Human pose estimation is a research hotspot in the field of computer vision, similar to face recognition [1–4]. It is useful for estimating human pose information from images or videos, and it can be integrated with other tasks such as target recognition [5–7], segmentation [8], regression [9], classification [10] and detection [11]. Human pose estimation can be categorized into two types, namely 2D human pose estimation and 3D human pose estimation, using image channels. Common RGB images are usually referred to as 2D images, whereas a channel with greater depth is added to obtain a 3D RGB-D image, as shown in Fig. 1. In recent years, human pose estimation has attracted increasing attention for applications such as driver condition monitoring [12] and pedestrian recognition [13–15].

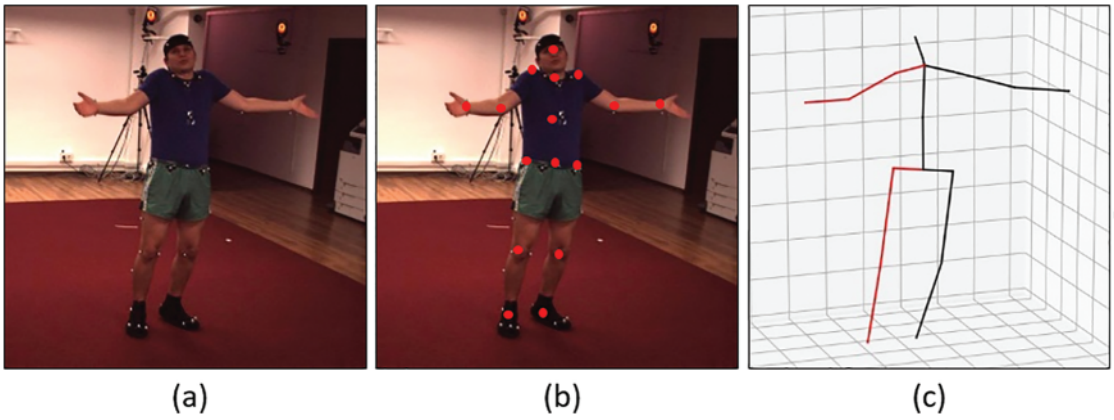

Figure 1: (a) Human body image; (b) 2D human pose estimation; (c) 3D human pose estimation

In particular, 3D pose estimation has witnessed rapid development. It involves estimating the 3D joint positions of the human body from a single view. Compared with 2D detection, 3D detection includes depth information, which is used to calculate the 3D coordinates of the human joint positions. Therefore, the precision of 3D human pose estimation is higher than that of 2D human pose estimation [16]. Moreover, its application scope and research value are also greater. In real scenarios, depth information can usually be obtained in two ways. One is to collect the depth data using special hardware equipment, such as 3D lidar and RGB-D [17] cameras; this approach usually has considerable hardware requirements and involves high costs. The other is to estimate the depth using single or multiple visible-light image sequences from the same scene [18–20], this approach requires correct device calibration and uses specific information for the scenario or device [21].

Currently, many breakthroughs have been achieved in studies on 2D human pose estimation [22–24]. By contrast, 3D human pose estimation has encountered several bottlenecks owing to the inherent depth ambiguity in mapping a single view from 2D to 3D, self-occlusion in a 3D space, and a lack of 3D data for specific scenarios. The main area of concern is single-view 3D pose estimation [25]. As it only uses a monocular camera, it involves issues such as uncertain depth and person-to-camera proportion. Nevertheless, the demand for human pose estimation from a single view has increased. Therefore, how to analyze and calculate the 3D human pose effectively and correctly using various algorithms has long been the focus of research in this field.

2 Research Overview

As deep neural networks have good feature extraction capabilities, many methods [26–31] directly employ deep convolutional neural networks (CNNs) to estimate 3D images from 2D images or other sources (such as point clouds [32,33]). Existing 3D estimation methods are based on two main methods. One method is to directly predict the 3D pose from an image. To avoid the acquisition of 2D–3D matching data, many studies [34–40] have decomposed the task of 3D pose estimation into two independent stages: First, existing 2D pose estimation methods are used to predict joint positions in the image space; then, a mapping is learned to extend them to a 3D space. The other method is to reduce the complexity of the entire task. This approach extracts features from images and directly outputs the information of 3D human joint positions. It is easy for a network to learn 2D-to-3D mapping. Meanwhile, 2D pose estimation is more mature than 3D estimation; hence, this method can easily use reprojection to perform semi-supervised learning, which makes it more mainstream than the first method. Many previous studies have investigated 3D human pose estimation. Inspired by [41], this article classifies the input methods mainly as image-based input and video-based input. These two input methods have an inherent logical relationship. Image-based input methods mainly adopt mainstream estimation techniques. The problems that need to be solved are the self-occlusion of joints in single-person photos and interpersonal occlusion in multi-person photos. The key is to use multi-view images [9], which yields excellent results. Moreover, owing to the lack of datasets, how to improve the generalization ability of the model has emerged as a research hotspot. Researchers are committed to creating new datasets or pre-processing data. Another effective method is to use weakly supervised learning for training. Video-based 3D human pose estimation focuses on how to use the temporal and spatial information of video sequences. Compared with image-based estimation, the advantages of video-based estimation are as follows: (1) The depth information contained in a single image is limited; however, a network can learn additional depth information from a video sequence. (2) Allowing the model to "see" additional images of people at different times from the same perspective reduces the depth ambiguity and narrows the spatial range of 3D poses; in fact, a 2D image can correspond to an infinite number of 3D poses. In this regard, our analysis focuses on the video-based method of using prior knowledge to improve the accuracy of the network as well as the use of weakly supervised learning for training. Considering that multi-view classification is similar to image-based estimation, a detailed description is not provided here.

3 Image-Based Input

Existing studies on 3D pose estimation with image-based input are mainly conducted from two aspects. One is to solve the problem of human occlusion, and the other is to improve the generalization ability of the model [42–61]. Most existing methods are based on regression. The self-occlusion of human joints and interpersonal occlusion contribute toward ambiguity in 3D human pose estimation. As image-based estimation cannot consider sequence information to solve the occlusion problem, it is necessary to consider multiple perspectives or further investigate the image information. It is extremely difficult to manually annotate 3D human poses in 2D images. Compared with 2D pose datasets, existing 3D pose datasets have low diversity in terms of poses and environments. This severely limits the use of supervised 3D pose estimation models. Many methods have attempted to train the models in a weakly supervised manner. Dataset bias is an important challenge. To address this problem, weakly supervised learning methods and new datasets are usually employed in order to improve the generalization ability of the model. This section also discusses some solutions to the problem of identification ambiguity caused by human diversity.

3.1 Obtaining the 3D Coordinates Based on the Violent Regression Method

The detection-based model directly returns the key points of the human body via 2D joint detection, and the performance of this method has been verified. In a 3D space, owing to its high degree of nonlinearity, the output space is large, which leads to challenges in the detection of human joints. At present, the regression-based model is highly popular. The regression task is to estimate the position of a joint relative to the root joint. It makes the relative position relationship between the parent and child joints easier to obtain. This method usually employs a multi-task training framework by combining detection and regression. For example, in reference [62], a multi-task framework was employed and a pre-training strategy was adopted to address the dependencies between various parts of the body, and the correlation between them was learned subsequently. Because the traditional regression method does not obtain the structure information effectively, the author proposed a structure-aware regression method that returns not the root-related joint but the parent-related joint. As the data variance of the parent-related joint is smaller than that of the root-related joint, the network regression becomes easier. Besides the mean square error of each bone, the author constrained some long-distance joint pairs to effectively mitigate the cumulative error [63].

3.2 Solving the Problem of Occlusion

The occlusion problem is essentially a problem of missing human key points. Because the pose estimation information of a monocular image is not perfect, the estimation algorithm can usually obtain only the relative coordinates of the human body key points (not the absolute coordinates). To address this problem, related studies have adopted a multi-view method in order to obtain the absolute position of the 3D pose of the human body. The multi-view method uses monocular cameras to capture the same object from different viewpoints at the same time. This process uses a 2D detector to locate the 2D joints in each view and then performs robust triangulation for 2D detection in each view to obtain the 3D joint positions. Some researchers have adopted multi-view modeling to overcome the problem of relying on single-view RGB images to train the network [64]. Although the multi-view method is effective, its acquisition cost is relatively high. Furthermore, researchers have used specific image data, i.e., mirror data, which comprise monocular images with multi-view information. In addition, the ground-truth values have been converted into the Mirrored-Human dataset by using an optimization method [65]. For a small number of cameras, in reference [66], a differentiable epipolar transformer was proposed to improve the multi-view pose estimation so that the 2D detector can use the 3D perception features to improve the 2D pose estimation. When a 2D position p in the current view is given, the corresponding point p_0 is first found in the adjacent view. Then, the feature at p_0 is combined with the feature at p to generate a 3D perceptual feature at p. An epipolar transformer uses epipolar constraints and feature matching to approximate the feature at p_0. However, this method has some limitations. When the perspective of adjacent camera views is too large, a certain 3D key point might be blocked, which makes feature matching more difficult. The method proposed in reference [67] can overcome this limitation to some extent. Microsoft Research Asia has proposed a 3D human pose estimation method based on cross-view information fusion [68]. This method first establishes a CNN with multi-view images as input, which can fuse the information from other viewpoints with that from the current viewpoint to obtain a more accurate 2D posture. A recursive pictorial structure model (RPSM) has been proposed to iteratively optimize the 3D pose obtained by the current PSM. The quantization error is reduced step by step from coarse to fine in order to obtain a more accurate 3D pose.

In contrast to the traditional multi-view method, it has been assumed in reference [67] that the video frames of different cameras are independent, although this method is based on video input. The video frames are obtained in chronological order and input by iteration. Given a certain frame in the

camera, the detected 2D key points are matched with the 3D pose predicted using the historical frames. If the matching is successful, the 3D joint positions of the target are updated according to the latest 2D key points. If the result is unmatched, the detected 2D key points will be retained and updated to a new 3D pose after the threshold is reached. Using its self-made dataset, this approach can achieve 154 fps with 12 cameras and 34 fps with 28 cameras.

Other excellent methods [69] have also been proposed to use part of the UV map to describe humans occluded by objects. High-latitude features are extracted by the UV map recovery branch to supervise the training of the color image encoder. The estimation of the complete human body shape is converted into a UV map restoration problem. In reference [70], the bottom-top method was adopted to detect the 3D human pose. It directly uses an absolute root map to sort each person's distance and performs calculations for the nearby human first in order to avoid overlap. This method can overcome the problem of inaccurate calculation of PAFs caused by human occlusion and overlap (commonly encountered in the case of dense crowds). The swoop network in reference [71] uses multi-person pose estimation, which can effectively deal with human occlusion and interaction caused by detection errors. The introduced discriminator enhances the effectiveness of human poses by closely interacting with the camera center coordinates.

In the case of severe occlusion, researchers have proposed an innovative multi-view method to construct a voxel-based expression of the scene (including people), as shown in Fig. 2. First, the detection network is used to calculate the approximate position of the human in 3D. Second, the pose estimation network is used to detect the fine 3D human pose around each detected position. This method can make direct inferences in a 3D space without making any hard decisions on 2D images. The process is divided into two steps: (1) The key points in each perspective are clustered into multiple instances; (2) the key points of the same person in different perspectives are associated. The limitation of this method is that the anchor size must be set in advance for different objects. This prolongs the calculation process and thus makes it difficult to use the method in practical applications [72].

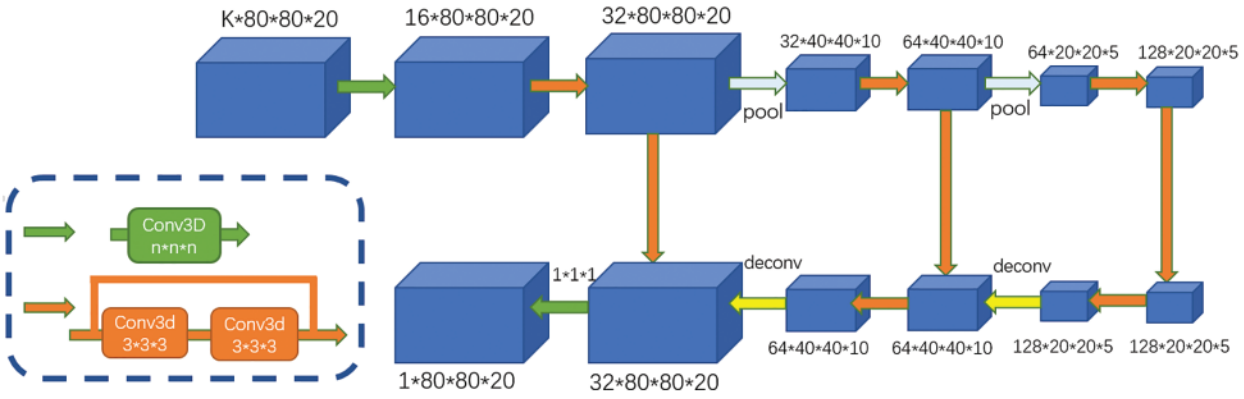

Figure 2: Network structure of CPN

3.3 Improving the Model Generalization Ability

Supervised learning models are strongly dependent on labeled samples and suffer from dataset bias. To improve their generalization ability, some recent studies [73,74] have used weakly supervised learning techniques in order to reduce the need for 3D-pose-GT annotations. Most of these studies involve auxiliary tasks, such as multi-view 2D pose estimation to train the 3D pose estimator [75,76]. Instead of using 3D-pose-GT for monitoring, the loss function is employed to monitor the 3D-pose

network based on the 2D pose of the multi-view projection. Therefore, some of these methods still require many annotations in terms of paired 2D-pose-GT [77–79], multi-view images [80], and known camera parameters [3].

In general, the representative datasets of existing machine vision techniques are uniformly distributed; however, there is usually a large amount of long-tailed data in actual training. This disagreement in the distribution results in data imbalance. Previous studies have adopted a rebalancing strategy to adjust the network training. Thus, they can make expectations closer to the distribution of tests by resampling samples in small batches or by using reweighted sample losses. However, this method affects the learning of deep features to a certain extent, such as by increasing the risk of over-fitting and under-fitting [81]. For the 3D human pose estimation task, the "local" human pose does not suffer from severe long-tail problems statistically, because each local pose may have been learned from the training data even if the overall pose may not have occurred in the data. The structure decomposes all the key points of the human body into several groups initially, because each key point in the group has a strong correlation, whereas the correlation between groups is relatively weak. First, the key points in each group pass through an independent sub-network to strengthen the calculation of local relationships (features). Second, the "low-dimensional global information" is calculated from the remaining key points of other groups and added back to this group in order to express the weak correlation between the key points inside and outside the group. The intra-group key point learning process reduces the dependence on weakly correlated key points outside the group without loss of global consistency by controlling the "global information" dimension. As the dependence on the weakly correlated key points is reduced, the model can reflect the distribution of "local" poses more effectively. Thus, it can achieve better generalization to new combined poses [82]. The flowchart of the algorithm is shown in Fig. 3.

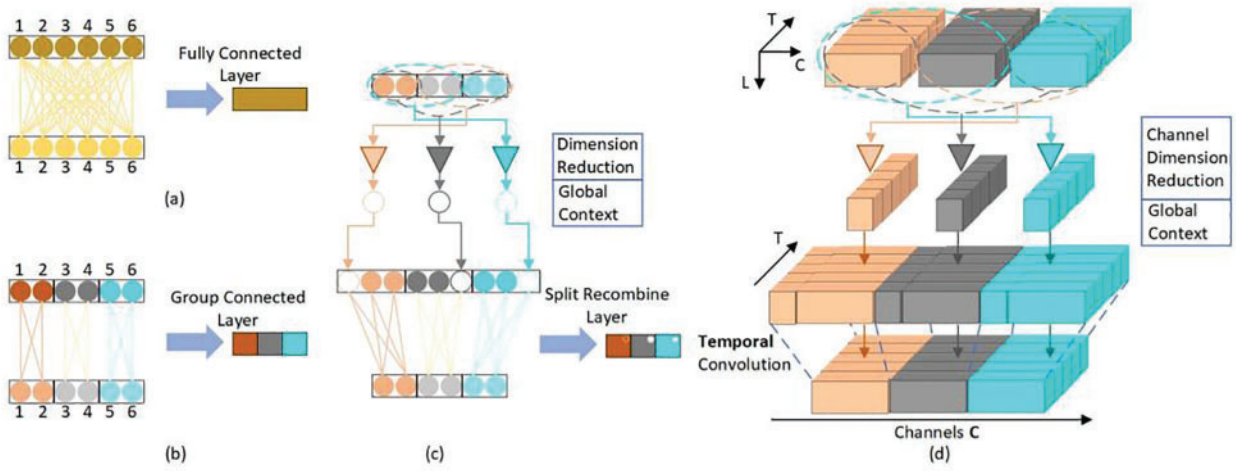

Figure 3: SRNet structure diagram: (a) Fully connected layer; (b) Group connected layer; (c) Split-recombine layer; (d) Convolution layer for temporal models

In existing methods [82], the parameters of the fusion model depend on a specific camera, and it is difficult to generalize them to a new environment. To solve this problem, in reference [83], a method was proposed to decompose the original fusion model into two parts: (1) A common model shared by all the cameras and (2) a lightweight transformation matrix for specific cameras. Moreover, the

meta-learning algorithm is used to pre-train large-scale multi-camera data in order to maximize the generalization ability of the model.

3.3.1 Weakly Supervised Learning

In reference [84], a network structure called RepNet was proposed on the basis of reprojection adversarial training in a weakly supervised manner, as shown in Fig. 4. It does not require the corresponding 3D pose as a supervision signal; however, it has good generalization ability for unknown data. This model consists of three parts. The first part involves pose and camera estimation. The network employs a dual-branch structure; one branch performs the network estimation of poses, and the other branch performs the network estimation of the camera parameters. The second part is the reprojection layer. Its main function is to map the generated 3D pose to a 2D pose. Thus, the input and output can be linked to facilitate model training. The third part is the critic network. It employs a dual-branch structure, where the first branch introduces the KCS condition (specifically, the input is a 3n 3D pose structure), while the second branch is a fully connected layer with the input 3n-dimensional matrix.

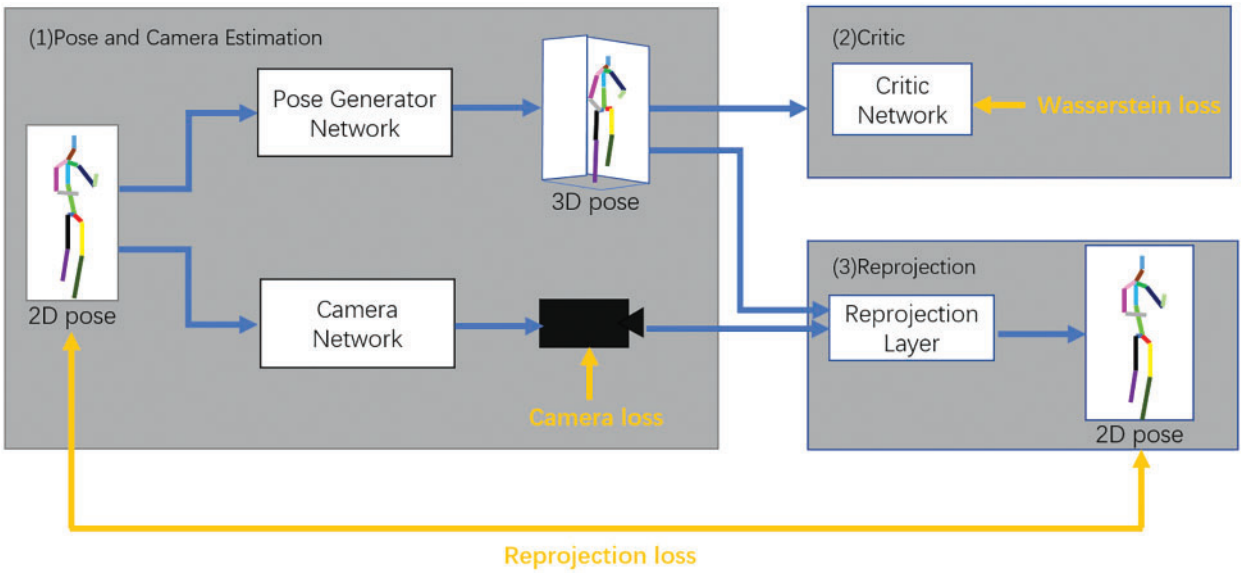

Figure 4: Network structure of RepNet. (1) A pose and camera estimation network; (2) A critic network; (3) A reprojection network

3.3.2 Solving the Problem of Insufficient Data

Although 3D human pose estimation has achieved considerable success, the 3D annotation for RGB images is a labor-intensive, time-consuming, and expensive process. Existing datasets are biased, and most of them are indoor datasets. Moreover, the actions included only cover a few selected daily behaviors. In reference [85], a new dataset called AMASS was proposed. This dataset is biased and can only deal with problems involving limited datasets. In reference [86], a dataset evolution framework was proposed for data enhancement operations (crossover, transformation, etc.), and new 3D skeletons were generated in a 3D space. For enhanced data, a cascade network (TAGNet) was proposed to predict the final 3D skeleton. Pavlakos et al. used the weakly ordered deep relationship between key points for supervision [87]. Hemlets used a heat map triplet loss as the ground truth to

encode the explicit depth ordering of adjacent key points. In reference [84], a generative adversarial training method was proposed; it uses a discriminator for weak supervision to overcome the problem of insufficient 3D annotation datasets.

3.4 Solving the Problem of Human Posture Diversity

One of the problems in the process of 3D human pose estimation is that different 3D poses may exhibit similar 2D projections. Human motion in the real world follows the laws of kinematics, including static/dynamic structures. Many studies [88–90] have employed simple structural constraints, such as symmetrical bone length [90] and limited joint angles [91], to promote 3D joint prediction. Most methods directly express this task as a coordinate regression problem without fully considering the inherent kinematic structure of the human body. However, this approach often leads to ineffective results. Therefore, it is of great significance to integrate the prior knowledge of kinematics into the deep model. Regarding the prior knowledge of human structure, Luvizon et al. [92] proposed a gradual method that clearly explains the different degrees of freedom between the various parts of the body. Sharma et al. [37] synthesized different reasonable 3D posture samples under the estimated 2D posture by generating the model of the automatic encoder based on the depth condition. Fang et al. [35] designed a deep grammatical network to explicitly encode the human dependencies in order to enhance the spatial consistency of the estimated 3D human poses by combining the geometric dependencies between different body parts. In reference [84], the motion chain characterization was encoded through the network layer, where the bone length and motion angle information was introduced. A generic combination of GCN and FCN [69] can also improve the representation ability. However, most of the aforementioned methods have ignored the kinematic structure and view correspondence. The intrinsic effectiveness of 3D poses has not been studied from a systematic and comprehensive perspective. In reference [93], optimization of the kinematic structure of 2D input with noise was proposed as the key to obtaining accurate 3D estimation. This process is shown in Fig. 5.

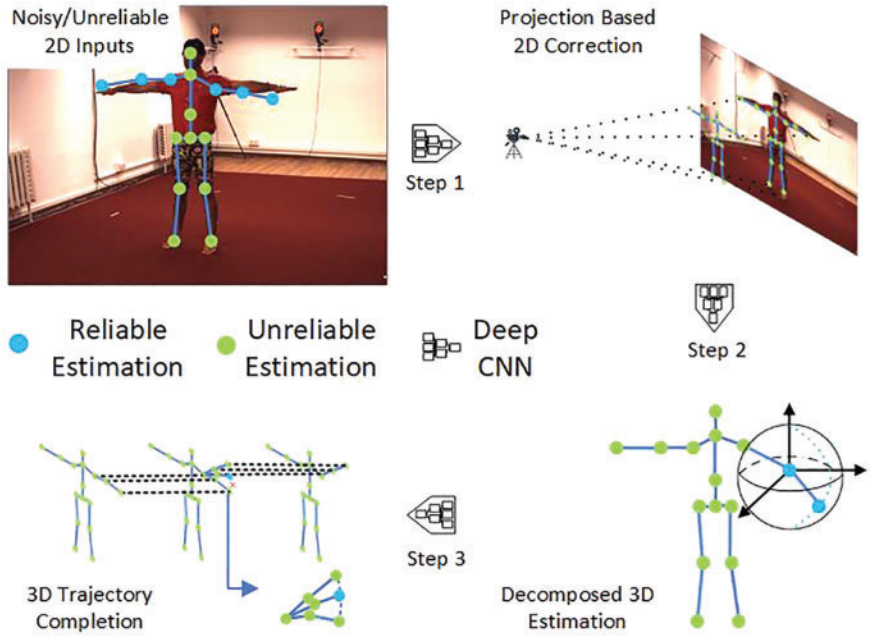

Figure 5: Overview of the framework

Initially, perspective projection is used to correct the 2D input containing noise (represented as red dots) and the 2D joints. Then, the joint motion and the human topology are explicitly decomposed. Finally, the unreliable 3D poses (represented by red crosses) are eliminated to complete the entire task. The three aforementioned steps are seamlessly integrated into the deep neural model to form a deep kinematics analysis pipeline that considers the static/dynamic structures of the 2D input and 3D output simultaneously. This experiment pioneered the use of perspective projection to refine 2D joints.

Table 1: Quantitative comparison of MPJPE

Protocol 1: MPJPE	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez et al. ICCV'17 (T = 1) [88]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	63.4	59.1	65.1	49.5	52.4	62.9
Luvizon et al. CVPR'18 [92]	49.2	51.6	47.6	50.5	51.8	60.3	48.5	51.7	61.5	70.9	53.7	48.9	57.9	44.4	48.9	53.2
Hossain et al. ECCV'18(T = 5) [94]	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Lee et al. ECCV'18 (T = 5) [95]	40.2	49.2	47.8	52.6	50.1	75.0	50.2	43.0	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Pavllo et al. CVPR'19 (T = 1) [90]	47.1	50.6	49.0	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	51.8
Pavllo et al. CVPR'19 (T = 9) [90]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.8
Protocol 2: PA-MPJPE	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Sun et al. ICCV'17 (T = 1) [63]	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
Fang et al. AAAI'18 (T = 1) [35]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Pavlakos et al. CVPR'18 (T = 1) [87]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1.	36.5	41.8
Hossain et al. ECCV'18(T = 5) [96]	35.7	39.3	44.6	43.0	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.4	44.1
Pavllo et al. CVPR'19 (T = 1) [90]	36.0	38.7	38.0	41.7	40.1	45.9	37.1	35.4	46.8	53.4	41.4	36.9	43.1	30.3	34.8	40.0

As shown in Table 1, the mean per-joint position error (MPJPE) between the estimated pose and the ground truth is quantitatively compared on the Human3.6 M dataset with millimeter as the unit, where T represents the number of input frames used in each method. When T = 1, it is better than existing methods in terms of the evaluation of P1 and P2 [97]. When T = 9 and T = 7, the estimation accuracy is improved and it is higher than that of Pavllo et al. [90], who considered the time smoothness but did not consider the reliability.

The latest study [98] has achieved improvements by reviewing existing methods for 3D human pose estimation and using limb-joint context information from a macro perspective. By combining deep neural networks with prior knowledge of human limbs, a general 3D human pose estimation formula is derived on the basis of "context" modeling. As shown in Fig. 6, when estimating a certain joint position, features are collected from its "context" joint positions (defined by the input body structure). These joint positions are integrated with the features and updated by these features.

A graph convolutional network (GCN) is a deep-learning-based method that performs convolution operations on graphs. Compared with the traditional CNN, GCNs have a unique convolution operator for irregular data structures. GCNs can be categorized into two types: Spectral-based GCN [99,100] and non-spectral-based GCN [101,102]. The latter attempts to expand the spatial definition of

convolution by rearranging the vertices of a graph into a certain grid form in order to directly perform conventional convolution operations, whereas the former uses the Fourier transform to perform the convolution process. Spectral GCN is usually suitable for processing graphs with fixed topology, whereas non-spectral GCN is suitable for processing graphs with topological changes. GCNs extends CNNs to any graph structure. They have attracted considerable attention and are widely used in many fields, such as image classification, document classification, and semi-supervised learning. However, these methods are based on a fixed curve graph as input.

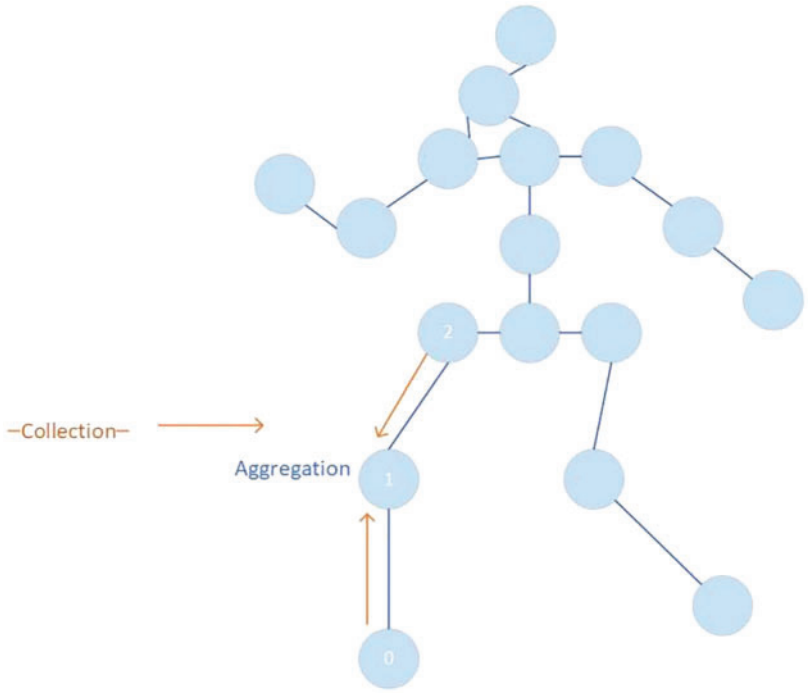

Figure 6: Update process of key nodes

Previous studies have used only the first-level edge of each node. This limits the receptive field to one dimension, and it is not conducive to learning global features. In reference [36], semantic GCNs were used to learn semantic information that is not explicitly represented in the graph, such as local and global node relationships. First, the pre-trained 2D pose network is used to extract the 2D skeleton from the image. Second, the 2D pose is input to the semantic GCN to return the 3D pose. Finally, summing of the joint position loss and bone length loss is performed to train the network. The experimental results are summarized in Table 2. It can be seen that SemCN achieves the performance of SOTA, while the parameter magnitude is reduced by 90%.

Table 2: SemCN performance comparison

Protocol 1 #1	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
Lonesdtal et al. [103]	132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Tekin et al. CVPR' 16 [104]	102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Zhou et al. CVPR' 16 [105]	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Du et al. ECCV'16 [106]	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Chen et al. CVPR'17 [107]	89.9	97.6	89.9	107.9	107.3	139.2	93.6	136.0	133.1	240.1	106.6	106.2	87.0	114.0	90.5	114.1
Pavlakos et al. [26]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Mehta et al. [27]	52.6	64.1	55.2	62.2	71.6	79.5	52.8	68.6	91.8	118.4	65.7	63.5	49.4	76.4	53.5	68.6
Zhou et al. [108]	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.1	66.0	51.4	63.2	55.3	64.9
Martinez et al. ICCV'17 [88]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Sun et al. [63]	52.8	54.8	54.2	**54.3**	61.8	53.1	53.6	71.7	86.7	61.5	67.2	**53.4**	47.1	61.6	53.4	59.1
Fang et al. AAAI' 18 [35]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Yang et al. CVPR'18 [109]	51.5	58.9	**50.4**	57.0	62.1	65.4	49.8	52.7	69.2	85.2	**57.4**	58.4	43.6	60.1	47.7	58.6
Hossain et al. [96]	48.4	**50.7**	57.2	55.2	63.1	72.6	53.0	**51.7**	**66.1**	80.9	59.0	57.3	62.4	**46.6**	49.6	58.3
SemCN (HG)	48.2	60.8	51.8	64.0	64.6	53.6	51.1	67.4	88.7	57.7	73.2	65.6	48.9	64.8	51.9	60.8
SemCN (RN w/FP)	**47.3**	60.7	51.4	60.5	**61.1**	**49.9**	**47.3**	68.1	86.2	**55.0**	67.8	61.0	**42.1**	60.6	**45.3**	**57.6**
SemCN (GT)	37.8	49.4	37.6	40.9	45.1	41.4	40.1	48.3	50.1	42.2	53.5	44.3	40.5	47.3	39.0	43.8

In contrast to the study of the unified GCN [110,111] for dense hand mesh reconstruction or spatial map LSTM, in reference [97], the GCN was used for spatiotemporal graphs with semantic grouping and sequence 3D pose estimation. In reference [97], the bone joint sequence was defined as a spatiotemporal graph. The graph topology takes the joints as the graph nodes, represents the spatial edges of the spatial dependence between different joints, and represents time edges that connect the same joints in adjacent frames. The adjacent nodes are classified according to their semantics to solve the problem of sparse connections and changing graph edges of functions in 3D pose estimation. Instead of different adjacent nodes with different kernels, the general graph convolution operator shares kernel weights to deal with adjacent nodes of the same degree.

A study [112] was conducted by Peking University and Microsoft Research Asia along with deep medical cooperation to solve the problem of 3D pose estimation of people in a scene from a single image. This work improved on the method proposed in reference [38], and a generalized formula, $y = X(S \odot W)$, was proposed on the basis of GCNs. Furthermore, a local connection network (LCN) was proposed to assign special filters to different joints in order to overcome the insufficient representation ability of GCNs. By using the common spatial integral method, end-to-end training was carried out by combining the existing 2D pose estimator and the LCN. Thus, the network representation and generalization ability was improved considerably. This approach has been successfully applied to different scenes.

3.5 Unsupervised Methods

Owing to the small amount of human posture data compared with other studies as well as the complexity of data annotation, the data richness that is available for training is not sufficient. By contrast, unsupervised learning can directly derive the properties of the data from the data itself and then summarize them, thereby enabling researchers to use these properties to make data-driven decisions. Therefore, some unsupervised-learning-based human pose estimation methods have been studied extensively [113–115]. Unsupervised learning usually treats pose estimation as a template matching problem that can be learned. A CNN is used to extract deep features, and the human body composition template with 2D Gaussian distribution features is then used to obtain the location information of key nodes. This method has achieved significant results.

3.6 Other Methods

In reference [116], a differentiable and modular self-supervised learning framework was proposed for monocular 3D human pose estimation. The encoder network takes an image as input and outputs three separate representations: (1) The unchanged 3D human pose observed in the standard coordinate system; (2) the camera parameters; (3) the potential code representing the human appearance in the foreground (FG). Then, the decoder network receives the aforementioned coded representation, projects it into 2D space, and synthesizes the FG human images. At the same time, the decoder also generates the 2D partial segmentation. In reference [117], during the training process, the predicted 3D pose representation and the sampled real 3D pose were alternately used to guide the model toward a reasonable 3D pose distribution. In reference [118], a novel bottom-up multi-person 3D human pose estimation method was proposed on the basis of monocular RGB images. It uses high-resolution volumetric heat maps to model joint positions and designs a simple and effective compression method. The essence of this method is a fully convolutional network-volume [119] at map automatic encoder, which is responsible for compressing the real heat map into a dense intermediate representation. In reference [120], a self-supervised estimation method called EpipolarPose was proposed for obtaining the 3D pose from a 2D image using multi-view polar geometry. This approach can directly predict the

3D human pose from a single image. During training, it does not require 3D supervision or external camera parameters. Epipolar geometry and 2D pose information are used to obtain the 3D pose similarity. The network is inferred from a single perspective, whereas multiple perspectives are used during training. Thus, even without any 3D ground-truth data and external camera parameters, this method can be used for the self-supervision of multi-view images, as shown in Fig. 7.

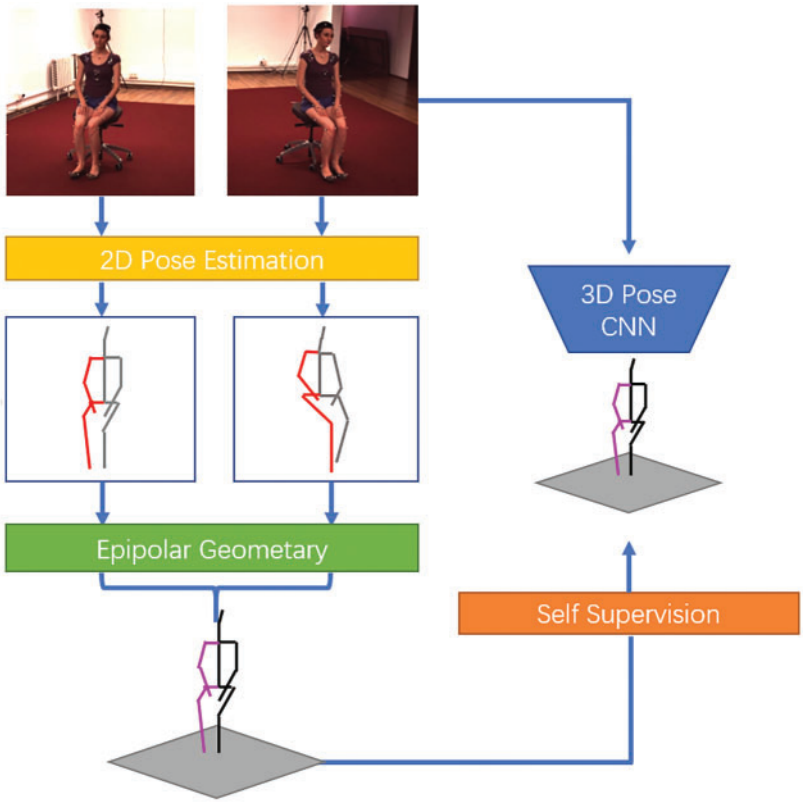

Figure 7: Epipolarpose algorithm flow chart

4 Video-Based Input

The 3D pose estimation of monocular video has attracted considerable attention in recent decades. It focuses on how to explore the temporal information from the video to generate more stable predictions and reduce the sensitivity to noise. In particular, it involves the estimation of human key point trajectories in a 3D space. Owing to motion blur and self-occlusion in video sequences, 2D detection is usually noisy and unreliable. The essence of video-based 3D pose estimation is how to use spatio-temporal information. One method is to use RNN or LSTM [121–123] to model the information of different frames. The other method is convolutional modeling. Compared with image-based methods, video-based methods can provide a larger amount of time and space information. Existing methods include RNN or LSTM and time-domain convolution to obtain sequence information [124]. By mining the information from the time dimension of the video, the accuracy of 3D human pose estimation can be further improved. Predicting the corresponding 3D joint position from the monocular video yields good results and requires less training resources than other methods that use RGB images [125]. Prior knowledge in space can not only reduce the possibility of generating

physically impossible 3D structures but also alleviate the problem of self-occlusion. Using temporal reasoning can help solve challenging problems, such as depth blur and visible jitter. Studies of video data input generally do not focus on solving the occlusion problem.

Recently, temporal information in monocular video has attracted increasing attention [90,126–128]. The difficulty in studies on video input is due to drastic changes in the human body shape in consecutive frames. To deal with the estimation of incoherence and jitter, some studies have used the temporal information between sequences. For example, Hossain et al. [96] designed a sequence-to-sequence network to predict 3D joint positions. They imposed a temporal smoothness constraint to ensure temporal consistency of the sequence during training. In reference [94], a sequence-to-sequence network composed of layer-normalized LSTM units was designed. It connects the input to the output on the decoder side and imposes a temporal smoothness constraint during the training process. XNect [129] was the first model to use RGB cameras to capture real-time 3D motion in a multi-person scene. This method can be divided into three stages. In the first stage, each person's 2D and 3D pose features are estimated through a CNN. This part designs a fast network structure called SelecSLS. In the second stage, the occluded parts of the 2D pose and 3D pose features are inferred as a complete 3D pose through the fully connected neural network. In the third stage, the spatio-temporal skeleton model is used to simulate the predicted 2D and 3D poses of each person. Its purpose is to further coordinate the 2D and 3D poses and improve the time consistency, and it finally returns the whole skeletal posture with joint angle information. In reference [130], GCNs were extended to a spatial-temporal graph model called ST-GCN. This model is constructed on a sequence of skeletal images, and each node corresponds to a joint of the human body. The model uses graph convolution to learn the skeleton data as well as the features of the established graph data and then identifies the behavior. In the relevant space, the key points between the skeletons are used as the input of the spatial relationship, and the video data are used in the temporal relationship between the input images.

4.1 Using Human Prior Knowledge

Existing methods take advantage of simple structural constraints. They employ symmetrical bone length [90] and limited joint angles [91] to facilitate 3D joint estimation; however, this is not sufficient to achieve significant improvement in this task. Most existing methods directly regard this task as a coordinate regression problem and fail to consider the inherent kinematic structure of the human body, which often leads to ineffective results. Studies have also investigated RNN methods, which consider prior knowledge based on the structural connectivity of body parts. Hossain et al. [96] used a temporal smoothness constraint across 2D joint position sequences to estimate 3D pose sequences. Pavllo et al. [90] transformed a series of 2D poses through temporal convolution. Thus, the computational complexity is independent of the spatial resolution of the key points, and the kinematics analysis pipeline can be explicitly combined for 3D pose estimation. In reference [116], a good pose prior constraint and a differentiable parent-relative local limb kinematics model were used to clarify the modeling as 3D rigid and non-rigid body posture transformation. This can reduce the ambiguity in the learning representation. In reference [97], field knowledge of the hand (body) structure was used in graph convolution operations to meet the specific requirements of 3D pose estimation.

4.2 Weakly Supervised Training

The 3D pose can be effectively predicted in videos using a fully convolutional model based on the 2D joint positions in the dilated temporal domain. In reference [90], a semi-supervised training method that can use unlabeled video data was proposed, as shown in Fig. 8. This method estimates the 2D joint positions based on unlabeled videos, then predicts the 3D pose, and finally performs reverse

mapping back to the 2D joint positions. A fully convolutional structure with residual connections has been proposed to perform time-domain convolution and linear projection layer alternation on the 2D nodes in the video. This extended convolution is used to model long-term dependencies and achieve accurate 3D pose estimation. This method is also compatible with any 2D joint position detector, which can process large contextual information through time-domain convolution. Compared with RNN, the convolution model of this method can process 2D pose information and time dimension information simultaneously. It can reduce the risk of not only gradient disappearance but also gradient explosion. This method is highly accurate, simple, and effective, and it also has other advantages such as low computational complexity and a small number of parameters.

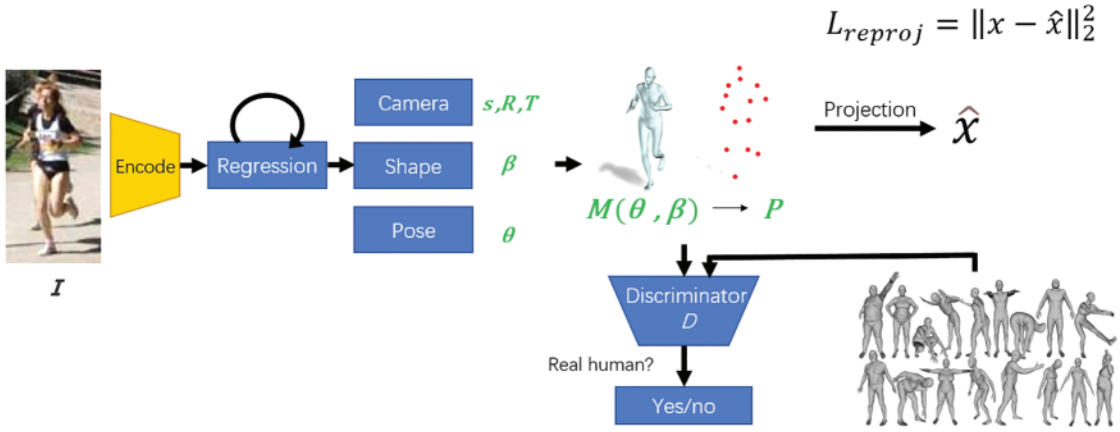

Figure 8: Human mesh restoration (HMR) framework

In reference [131], how to estimate human 3D actions more accurately using monocular video was discussed according to the time information and training of an adversarial learning network. AMASS [85] has been used to distinguish real human actions from human actions generated by the regression network. Thus, the regression can output more real actions by minimizing the error of adversarial training. The discriminator is trained in a semi-supervised manner. In reference [132], a monocular human motion capture method was proposed on the basis of a weakly supervised neural network, as shown in Fig. 9. The entire network architecture is divided into two parts, namely PoseNet and DefNet, which perform human pose estimation from monocular images (expressed on the basis of the angles of joint positions) and non-rigid surface deformation (expressed on the basis of the embedded deformation graph), respectively. The training process is performed in a weakly supervised manner based on multi-view images in order to avoid the use of 3D annotation data. The author trained a neural network to achieve this goal. Thus, the author proposed a differentiable human deformation and rendering model. This model can render the human body model and achieve a backpropagation loss compared with the 2D image. Compared to other methods in terms of the accuracy of the skeleton pose, the proposed method achieved better results, as shown in Table 3.

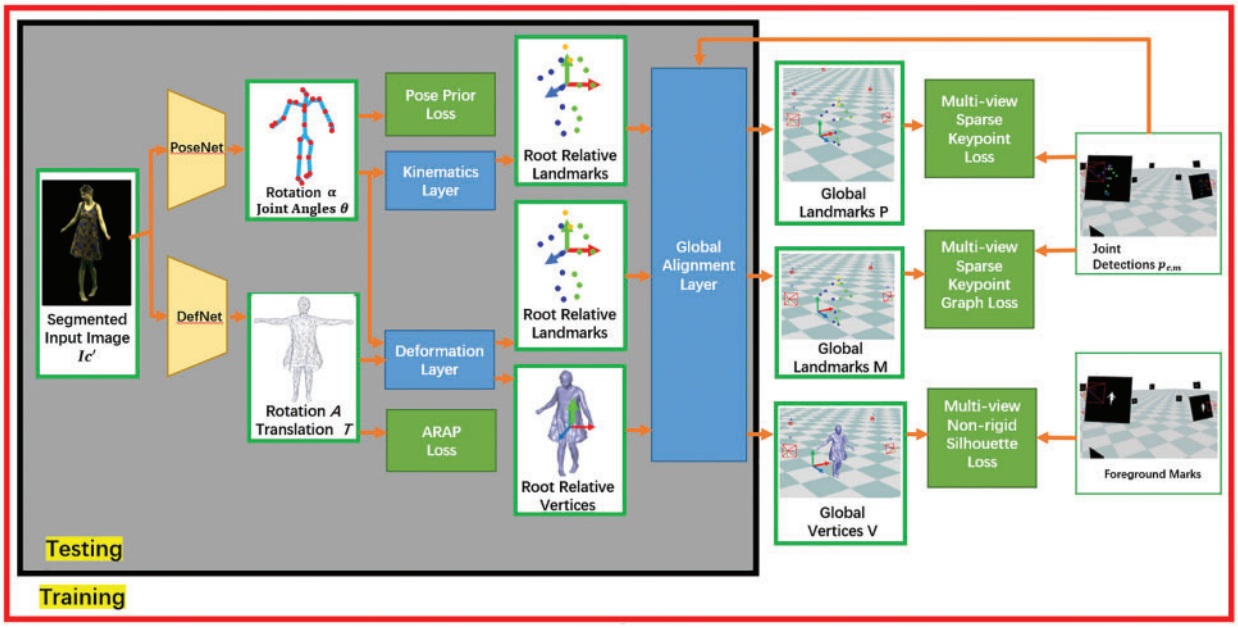

Figure 9: Deepcap is realized by a neural network composed of two parts

Table 3: MVBL performance comparison

MPJPE/GLE (in mm) and 3DPCK/AUC (in %) on S1				
Method	GLE↓	3DPCK↑	AUC↑	MPJPE↓
VNect	-	66.06	28.02	77.19
HMR	-	82.39	43.61	72.61
HMMR	-	87.48	45.33	72.40
LiveCap	317.01	71.13	37.90	92.84
MVBL	76.03	99.17	57.79	45.44

5 Dataset

Currently, the Human3.6 M dataset is the most widely used datin 3D pose estimation. It includes 3.6 million images captured by four cameras from different perspectives (50 fps video). It covers 15 actions: Directions, discussion, eating, greeting, phoning, posing, purchases, sitting, sitting down, smoking, taking photos, waiting, walking, walking dogs, and walking together. Further, it contains 11 individuals, 7 of which have 3D tags. Therefore, S1, S5, S6, S7, and S8 are generally used as training sets, while S9 and S11 are used as test sets. The dataset is available at http://vision.imar.ro/human3.6m/description.php. The data sample is shown in Fig. 10.

Figure 10: Human3.6 M data sample

MPI-INF-3DHP was developed by the Max Planck Institute for Informatics. It is a 3D human pose estimation dataset composed of constrained indoor and complex outdoor scenes. It covers 8 actors performing 8 activities from 14 camera views, including walking and sitting postures as well as complex sports postures and dynamic actions. It contains 1.3 million frames, covering more posture categories than Human3.6 M. This dataset uses multiple unmarked cameras to capture actors in a green screen studio. By calculating the masks of different areas and independently synthesizing different textures of the background, chair, upper body, and lower body areas, the captured image can be enhanced. The camera captures the actions of wearing everyday clothes, and each actor has two sets of clothes to wear in the activity: One set of clothing is casual daily clothing, while the other set is plain color clothing. In contrast to the existing dataset, this dataset allows automatic segmentation and expansion and provides true 3D annotation as well as a universal skeleton compatible with Human3.6 M. Compared with the unmarked records proposed by Joo et al. [133], this dataset provides a wider range for foreground and background enhancement. It uses images sampled from the Internet for background enhancement as well as a simplified internal decomposition for foreground enhancement. For plain clothes, the change in intensity is only caused by the shadow; hence, the average pixel intensity is used as a substitute for the shadow component. In reference [27], a new test set containing ground-truth annotations from a multi-view unmarked motion capture system was created. It complements existing test sets with more diverse movements (standing/walking, sitting/-lying, exercising, dynamic postures on the floor, dancing, etc.). It also includes camera perspective changes, clothing changes, and outdoor recordings in an unconstrained environment captured by Robertinis et al. [134]. The dataset is available at http://gvv.mpi-inf.mpg.de/3dhp-dataset. The data sample is shown in Fig. 11.

Owing to the high cost of collecting a large-scale multi-person 3D pose estimation dataset, MPI-INF-3DHP is synthesized using the MuCo-3D-HP dataset with data enhancement. The data enhancement methods include background enhancement and perceptual shadow enhancement of human contours. This method uses single-person image data of real people in MuCo-3D-HP to synthesize numerous multi-person interactive images under the control of the user. It also includes 3D pose annotations. A new shooting (non-synthesis) multi-person test set is proposed, including 20 general real-world scenes with ground-truth 3D poses, which can be obtained by up to three subjects using a multi-view unmarked motion capture system. In addition, occlusion instructions are provided for each joint. This set of scenes includes 5 indoor and 15 outdoor scenes. The background includes trees, office buildings, roads, people, vehicles, and other fixed and moving entities. The new test set is called MuPoTS-3D [135]. The dataset is available at https://paperswithcode.com/dataset/mupots-3d. The data sample is shown in Fig. 12.

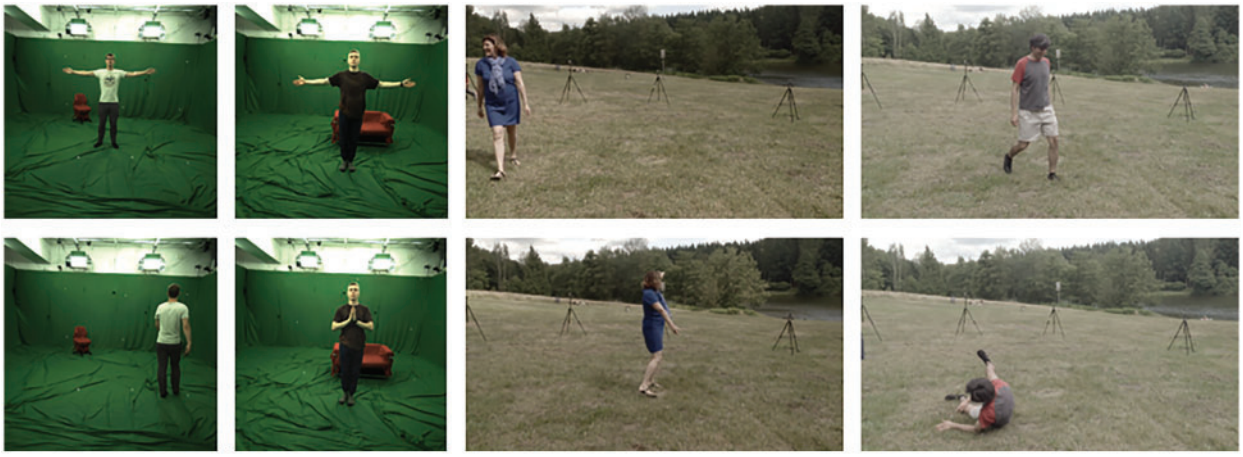

Figure 11: MPI-INF-3DHP data sample

Figure 12: MuPoTS-3D data sample

CMU Panopoptic: This point cloud dataset is captured by 10 synchronous Kinect (Kinoptic Studio) devices installed using Panoptic Studio. The Kinect data are captured by more than 500 RGB cameras. They share temporal-spatial and 3D world coordinates. The point cloud output can be used with the output of an RGB camera (such as RGB video and 3D skeleton). It contains 10 synchronized RGB-D videos, 3D point clouds from 10 RGB-D videos, 31 synchronized HD videos of the same scene from other viewpoints, calibration parameters for 10 RGB-D cameras and 31 HD cameras, and synchronization tables for all the RGB-D and HD videos [136]. The dataset is available at http://domedb.perception.cs.cmu.edu/dataset.html. The data sample is shown in Fig. 13.

AMASS: It is a large and diverse human motion database, which is standardized and parameterized by representing 15 different motion capture (mocap) datasets based on optical tags in a common framework. It uses a new method called MoSh++ to convert the mocap data into a real 3D human body mesh initially, and finally maps a large amount of labeled data to the common SMPL posture, shape, and soft tissue parameters. It is represented by an assembled human model, providing a standard skeleton representation and a fully assembled curved surface mesh. The consistent representation

of AMASS makes it very useful for animation, visualization, and generating training data for deep learning. Compared with other human motion datasets, it is much richer, with more than 40 h of motion data and recordings, covering 344 subjects and 11265 actions [85]. The dataset is available at https://amass.is.tue.mpg.de.

Figure 13: CMU Panopoptic data sample

Leeds Sports Pose (LSP) dataset: It is a single human body key point detection dataset. The number of key points is 14, and the number of samples is 2000. It is the second-most commonly used dataset in current research. It covers many sports postures, including athletics, badminton, baseball, gymnastics, parkour, football, volleyball, and tennis. It contains around 2000 posture annotations. The images are all obtained from athletes on Flickr. Each image is a three-channel color image. The row range of pixels is 23, and the column range is 16. Each image is marked with 14 joint positions, and the left and right joints are always marked according to the human center [85]. The dataset is available at http://sam.johnson.io/research/lsp.html. The data sample is shown in Fig. 14.

MPII Human Pose dataset: It is based on single/multiple human body key point detection datasets. The entire human body has a total of 16 joints. The MPII Human Pose dataset is a benchmark for human pose estimation. It includes 25,000 annotated images of more than 40,000 people, which are extracted from YouTube videos. The test set also includes body part occlusion, 3D torso, and head direction annotations. The dataset is available at http://human-pose.mpi-inf.mpg.de. The data sample is shown in Fig. 15.

The PoseTrack dataset includes videos around key frames from the MPII Human Pose dataset. These videos contain multiple individuals and non-static scenes. It uses MPII to select 41–298 adjacent frames of video clips and crowded scenes. This dataset contains an unconstrained evaluation protocol (without any prior assumptions about the size, location, and number of people, which are arbitrary). The scenes contain multiple people, and the people are articulated with each other and participate in a variety of dynamic activities. The video contains many body movements and postures, appearance changes, as well as high mutual occlusion and truncation. Part or all of the target disappears and reappears as much as possible. The dataset annotates the head boundary of each person in the video. Each person who appears in the video is assigned a unique tracking ID until that person leaves the field of view of the camera. For each person being tracked, 15 parts are annotated in the video,

including the head, nose, neck, shoulders, elbows, wrists, hips, knees, and ankles. Finally, the VATIC tool is used to accelerate the annotation process. The annotation between adjacent frames is completed by interpolation. The dataset contains 550 videos, 66,374 frames, divided into 292 training videos, 50 verification videos, and 208 test videos. For each video sequence, 30 frames in the middle are annotated to obtain a total of 23,000 annotated frames and 153,615 annotated poses. In addition, in the verification and test set, dense annotations are made every 4 frames to test the ability and stability of long-term tracking of body joints. The dataset is available at https://posetrack.net.

Figure 14: LSP data sample

Figure 15: MPII data sample

The Martial Arts, Dancing and Sports (MADS) dataset is provided by the City University of Hong Kong. It contains five categories, namely tai chi, karate, jazz, hip hop, and sports, with a total of 53,000 frames. The motion capture of this dataset is recorded by two martial arts masters, two dancers, and one athlete using multiple cameras or stereo cameras. The motions in the MADS dataset are more complex and challenging than ordinary motions. First, they have a larger range of motion, and some postures do not appear in normal actions. Second, there are more self-occlusions and interactions between limbs. Third, the motions are relatively fast. A Gaussian Mixture Model (GMM) [29] and shadow detection [30] are used to remove the background and obtain the human contours. This dataset can be used for research in human pose estimation and other fields [137]. It is available at http://visal.cs.cityu.edu.hk/research/mads/#download.

The CrowdPose dataset was constructed by a team at Shanghai Jiaotong University. It is used for multi-person joint position recognition in crowded scenes, with 14 joint positions per person. According to the crowd index of MSCOCO (person subset), MPII, and AI Challenger, the images are divided into 20 groups, ranging from 0 to 1, with a step size of 0.05 between the groups. Then, 30,000 images are evenly extracted from these groups. Further, 20,000 high-quality images are selected from the 30,000 images, each person in the image is cropped, and the key points of interference in each bounding box are marked. The dataset consists of 20,000 images and contains approximately 80,000 people. The training, verification, and test subsets are divided in a ratio of 5:1:4. The crowd index satisfies a uniform distribution in [0,1]. The dataset is designed to not only improve the performance in crowded situations but also extend the model to different scenarios [138]. It is available at https://github.com/MVIG-SJTU/AlphaPose/blob/pytorch/doc/CrowdPose.md.

PedX is a large multi-modal pedestrian dataset based on a complex urban intersection. It consists of more than 5,000 pairs of high-resolution (12 MP) stereo images and lidar data, and it provides 2D image tags and pedestrian 3D tags in the global coordinate system. The data were captured at the intersection of three four-way parking lanes with considerable interaction between pedestrians and vehicles. The author also proposed a 3D model-fitting algorithm to automatically label the constraints of different modes as well as novel shapes and time priors. All the annotated 3D pedestrians are in the real-world metric space, and the generated 3D model is validated using a motion capture system configured in a controlled outdoor environment to simulate pedestrians at urban intersections. The manual 2D image tags also can be replaced by advanced automatic labeling methods, which facilitate the automatic generation of large-scale datasets [139]. The dataset is available at http://pedx.io.

6 Evaluation Indicators

6.1 PSS

The pose structure score (PSS) is proposed to measure the structural similarity. The traditional distance evaluation indicators (MPJPE and PCK) deal with each joint position independently, and they cannot evaluate the structural accuracy of the posture as a whole.

Therefore, the PSS indicator is designed to measure structural similarity:

$$PSS(p,q) = \delta(C(p), C(q)) \tag{1}$$

$$\text{where } C(p) = \arg\min \|p - \mu_k\|_2^2, \delta(i,j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{2}$$

Calculating PSS requires the pose distribution of the ground truth as a reference. The ground-truth set is assumed to have n poses, and each vector is standardized as $\hat{q}_i = \frac{q_i}{\|q_i\|} q_i$. After k-means clustering, the PSS is calculated between the estimated pose p and the true value q [41].

6.2 Mean Per-Joint Position Error (MPJPE)

It is the average Euclidean distance between the joint position coordinates output by the network and the ground truth (usually converted to camera coordinates).

6.3 Procrustes Analysis MPJPE (P-MPJPE)

First, the network output is rigidly aligned (translation, rotation, and scaling) with the ground truth; then, the MPJPE is calculated.

6.4 Percentage of Correct Key Points (PCK)

When the distance between the predicted joint and the ground truth is within a certain threshold, the detected joint is considered to be correct.

6.5 Percentage of Correct Parts (PCP)

If the distance between the two predicted joint positions and the ground truth is less than half of the limb length, the limb is considered to be detected.

6.6 3DPCK

The principle of 3DPCK is that if a joint is located within a 15 cm sphere centered at the ground-truth joint position, the joint prediction is regarded as correct and the common minimum set of 14 marked joints is evaluated. 3DPCK is more robust than MPJPE, and it [103] also helps offset the jitter effect in all non-synthetic annotations (including our annotations).

6.7 Evaluation Indicator Analysis

Both MPJPE and P-MPJPE are commonly used evaluation indicators that represent the error value of the results. PSS focuses on the structural accuracy of the overall result rather than the average error of each point position. PCK and 3DPCK represent the percentage of correct key points. These evaluation methods can alleviate the problem of short limbs. Further, PCP penalizes shorter limbs more than PCK.

7 Conclusion

Research on 3D human pose estimation is attracting increasing attention. This article systematically introduced recent advancements in 3D human pose estimation on the basis of monocular cameras. Different data input formats lead to different research focus areas, namely image-based methods and video-based methods. The solutions to the problems faced by these two methods are similar. For landmark networks, this article compared the performances of some algorithms in order to prove their effectiveness.

Image-based input mainly focuses on estimation using the regression algorithm. The well-known existing algorithm is divided into two steps. The first step is to detect the 2D key points, and the second step is to map the 2D key points to the 3D key points. To address occlusion problems, researchers often use the multi-view method in order to improve the estimation results. To address the problem of insufficient training data, some methods [86–87,140] use data enhancement. Meanwhile, some other excellent methods use weakly supervised learning to improve the generalization ability of the model [73–74,84]. In addition, image-based input often encounters the problem of human pose diversity, which can be solved using kinematic constraints [88–93]. In recent years, some studies have investigated graph convolution [38]; these studies are summarized here. Compared with image data, video data has time-series information, which can be used to estimate human poses more effectively as well as to alleviate the problem of self-occlusion to a certain extent. Video-based input methods focus on taking advantage of human prior knowledge to constrain the estimation process [90,91] and adopt weakly supervised training to improve the performance of the model [131].

Directions for future research on 3D human pose estimation based on a monocular camera are as follows: (1) Owing to the limitations of 3DHPE, the HPE method cannot be effectively extended to different fields. Therefore, how to reduce the model parameter compression to ensure real-time performance must be investigated. (2) The interaction between humans and 3D scenes must be

explored. (3) Visual tracking and analysis can be achieved using physical constraints. (4) The problem of inaccurate estimation using low-resolution input must be solved. (5) Another inevitable problem is that noise has a significant impact on the performance of HPE. Therefore, how to improve the robustness of the HPE network is a topic for future research.

Acknowledgement: The authors would like to thank TopEdit (www.topeditsci.com) for its linguistic assistance during the preparation of this manuscript.

Funding Statement: This project is supported by the Program of Entrepreneurship and Innovation Ph.D. in Jiangsu Province (JSSCBS20211175); the School Ph.D. Talent Funding (Z301B2055); the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (21KJB520002).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Zhang, L., Li, W., Yu, L., Sun, L., Dong, X. et al. (2021). GmFace: An explicit function for face image representation. *Displays, 68,* 102022. DOI 10.1016/j.displa.2021.102022.
2. Li, S., Ning, X., Yu, L., Zhang, L. P., Dong, X. L. et al. (2020). Multi-angle head pose classification when wearing the mask for face recognition under the COVID-19 coronavirus epidemic. *2020 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*, pp. 1–5. Shenzhen, China, IEEE.
3. Ning, X., Duan, P., Li, W., Shi, Y., Li, S. (2020). A CPU real-time face alignment for mobile platform. *IEEE Access, 8,* 8834–8843. DOI 10.1109/Access.6287639.
4. Li, S., Dong, X., Shi, Y., Lu, B., Sun, L. et al. (2021). Multi-angle head pose classification with masks based on color texture analysis and stack generalization. *Concurrency and Computation: Practice and Experience, 2021,* e6331. DOI 10.1002/cpe.6331.
5. Cai, W., Liu, D., Ning, X., Wang, C., Xie, G. (2021). Voxel-based three-view hybrid parallel network for 3D object classification. *Displays, 69,* 102076. DOI 10.1016/j.displa.2021.102076.
6. Qi, S., Ning, X., Yang, G., Zhang, L., Long, P. et al. (2021). Review of multi-view 3D object recognition methods based on deep learning. *Displays, 69,* 102053. DOI 10.1016/j.displa.2021.102053.
7. Li, S., Zhang, B. (2021). Joint discriminative sparse coding for robust hand-based multimodal recognition. *IEEE Transactions on Information Forensics and Security, 16,* 3186–3198. DOI 10.1109/TIFS.2021.3074315.
8. Jiang, Y., Zhao, K., Xia, K., Xue, J., Zhou, L. et al. (2019). A novel distributed multitask fuzzy clustering algorithm for automatic MR brain image segmentation. *Journal of Medical Systems, 43(5)*. DOI 10.1007/s10916-019-1245-1.
9. Ning, X., Duan, P., Li, W., Zhang, S. (2020). Real-time 3D face alignment using an encoder-decoder network with an efficient deconvolution layer. *IEEE Signal Processing Letters, 27,* 1944–1948. DOI 10.1109/LSP.97.
10. Li, S., Sun, L., Ning, X., Shi, Y., Dong, X. (2019). Head pose classification based on line portrait. *2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*, pp. 186–189. Shenzhen, China, IEEE.
11. Wang, G., Li, W., Zhang, L., Sun, L., Chen, P. et al. (2021). Encoder-x: Solving unknown coefficients automatically in polynomial fitting by using an autoencoder. *IEEE Transactions on Neural Networks and Learning Systems, 2021,* 1–13. DOI 10.1109/TNNLS.5962385.

12. Jiang, Y., Zhang, Y., Lin, C., Wu, D., Lin, C. T. (2021). EEG-Based driver drowsiness estimation using an online multi-view and transfer TSK fuzzy system. *IEEE Transactions on Intelligent Transportation Systems, 22(3),* 1752–1764. DOI 10.1109/TITS.6979.
13. Ning, X., Gong, K., Li, W., Zhang, L., Bai, X. et al. (2021). Feature refinement and filter network for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology, 31(9),* 3391–3402. DOI 10.1109/TCSVT.2020.3043026.
14. Ning, X., Gong, K., Li, W., Zhang, L. (2021). JWSAA: Joint weak saliency and attention aware for person re-identification. *Neurocomputing, 453,* 801–811. DOI 10.1016/j.neucom.2020.05.106.
15. Zhang, J., Huang, B., Ye, Z., Kuang, L., Ning, X. (2021). Siamese anchor-free object tracking with multiscale spatial attentions. *Scientific Reports, 11(4),* 1–14. DOI 10.1038/s41598-021-02095-4.
16. Wang, M., Sun, T., Song, K., Li, S., Jiang, J. et al. (2022). An efficient sparse pruning method for human pose estimation. *Connection Science, 34,* 960–974. DOI 10.1080/09540091.2021.2012423.
17. Mur-Artal, R., Tardos, J. D. (2017). ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics, 33(5),* 1255–1262. DOI 10.1109/TRO.2017.2705103.
18. Ning, X., Wang, X., Xu, S., Cai, W., Zhang, L. et al. (2021). A review of research on co-training. *Concurrency and Computation: Practice and Experience, 2021,* e6276.
19. Liao, L., Su, L., Xia, S. (2017). Individual 3D model estimation for realtime human motion capture. *2017 International Conference on Virtual Reality and Visualization (ICVRV)*, pp. 235–240. Zhengzhou, China, IEEE.
20. Lu, Y., Li, W., Ning, X., Dong, X., Zhang, L. et al. (2021). Blind image quality assessment based on the multiscale and dual-domains features fusion. *Concurrency and Computation: Practice and Experience, 2021,* e6177.
21. Chen, J., Zhang, Y., Wu, L., You, T., Ning, X. (2021). An adaptive clustering-based algorithm for automatic path planning of heterogeneous UAVs. *IEEE Transactions on Intelligent Transportation Systems, 2021,* 1–12. DOI 10.1109/TITS.2021.3131473.
22. Biswas, S., Sinha, S., Gupta, K., Bhowmick, B. (2019). Lifting 2D human pose to 3D: A weakly supervised approach. *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. Budapest, Hungary, IEEE.
23. Newell, A., Yang, K., Jia, D. (2016). Stacked hourglass networks for human pose estimation, *European Conference on Computer Vision*, vol. 2016, pp. 483–499. Springer, Cham.
24. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B. (2018). Neural body fitting: Unifying deep learning and model based human pose and shape estimation. *2018 International Conference on 3D Vision (3DV)*, pp. 484–494. Verona, Italy, IEEE.
25. Zhang, T., Li, S., Long, P. (2021). 3D human pose estimation in motion based on multi-stage regression. *Displays, 69,* 102067. DOI 10.1016/j.displa.2021.102067.
26. Pavlakos, G., Zhou, X., Derpanis, K., Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3D human pose. *CVPR, 2017,* 7025–7034. DOI 10.1109/CVPR.2017.139.
27. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O. et al. (2017). Monocular 3D human pose estimation in the wild using improved cnn supervision. *2017 International Conference on 3D Vision (3DV)*, pp. 506–516. Qingdao, China, IEEE.
28. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K. (2018). Learning to estimate 3D human pose and shape from a single color image. *CVPR, 2018,* 459–468. DOI 10.1109/CVPR.2018.00055.
29. Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C. (2019). In the wild human pose estimation using explicit 2D features and intermediate 3D representations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10905–10914. Long Beach California, USA.
30. Zhou, K., Han, X., Jiang, N., Jia, K., Lu, J. (2019). Hemlets pose: Learning part-centric heatmap triplets for accurate 3D human pose estimation. *ICCV, 2019,* 2344–2353.

31. Cheng, Y., Yang, B., Wang, B., Yan, W., Tan, R. (2019). Occlusion-aware networks for 3D human pose estimation in video. *ICCV, 2019,* 723–732.
32. Liu, J., Ni, B., Li, C., Yang, J., Tian, Q. (2019). Dynamic points agglomeration for hierarchical point sets learning. *IEEE International Conference on Computer Vision*, pp. 7546–7555. Seoul, South Kerean.
33. Yan, Y., Zhuang, N., Zhang, J., Xu, M., Zhang, Q. et al. (2019). Fine-grained video captioning via graph-based multi-granularity interaction learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(2),* 666–683.
34. Tekin, B., Neila, P., Salzmann, M., Fua, P. (2017). Learning to fuse 2D and 3D image cues for monocular body pose estimation. *ICCV, 2017,* 3941–3950. DOI 10.1109/ICCV.2017.425.
35. Fang, H. S., Xu, Y., Wang, W., Liu, X., Zhu, S. C. (2018). Learning pose grammar to encode human body configuration for 3D pose estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. Lousiana, USA.
36. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D. N. (2019). Semantic graph convolutional networks for 3D human pose regression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3425–3435. Long Beach California, USA.
37. Sharma, S., Varigonda, P. T., Bindal, P., Sharma, A., Jain, A. (2019). Monocular 3D human pose estimation by generation and ordinal ranking. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2325–2334. South Kerean.
38. Ci, H., Wang, C., Ma, X., Wang, Y. (2019). Optimizing network structure for 3D human pose estimation. *ICCV, 2019,* 2262–2271.
39. Wang, C., Wang, X., Zhang, J., Zhang, L., Bai, X. et al. (2022). Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition, 124,* 108498. DOI 10.1016/j.patcog.2021.108498.
40. Yu, Z., Zhang, L., Li, S., Zhang, Y., Ning, X. (2021). 2D-3DMatchingNet: Multimodal point completion with 2D geometry matching. *2021 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*, pp. 94–99. Macao, China.
41. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I. A. (2016). 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding, 152,* 1–20. DOI 10.1016/j.cviu.2016.09.002.
42. Yao, P., Fang, Z., Wu, F., Feng, Y., Li, J. (2019). Densebody: Directly regressing dense 3D human pose and shape from a single color image. arXiv:1903.10153.
43. Liu, K., Ding, R., Zou, Z., Wang, L., Tang, W. (2020). A comprehensive study of weight sharing in graph networks for 3D human pose estimation. *European Conference on Computer Vision*, pp. 318–334. Springer, Cham.
44. Chen, Z., Huang, Y., Yu, H., Xue, B., Han, K. et al. (2020). Towards part-aware monocular 3D human pose estimation: An architecture search approach. *European Conference on Computer Vision*, pp. 715–732. Springer, Cham.
45. Wang, J., Yan, S., Xiong, Y., Lin, D. (2020). Motion guided 3D pose estimation from videos. *European Conference on Computer Vision*, pp. 764–780. Springer, Cham.
46. Nie, Q., Liu, Z., Liu, Y. (2020). Unsupervised human 3D pose representation with viewpoint and pose disentanglement. *Computer Vision–ECCV 2020*, pp. 102–118.
47. Weinzaepfel, P., Brégier, R., Combaluzier, H., Leroy, V., Rogez, G. (2020). Dope: Distillation of part experts for whole-body 3D pose estimation in the wild. *European Conference on Computer Vision*, pp. 380–397. Springer, Cham.
48. Chen, H., Guo, P., Li, P., Lee, G. H., Chirikjian, G. (2020). Multi-person 3D pose estimation in crowded scenes based on multi-view geometry. *European Conference on Computer Vision*, pp. 541–557. Springer, Cham.

49. Kundu, J. N., Revanur, A., Waghmare, G. V., Venkatesh, R. M., Babu, R. V. (2020). Unsupervised cross-modal alignment for multi-person 3D pose estimation. *Computer Vision–ECCV 2020*, pp. 35–52.
50. Huang, C., Jiang, S., Li, Y., Zhang, Z., Traish, J. et al. (2020). End-to-end dynamic matching network for multi-view multi-person 3D pose estimation. *European Conference on Computer Vision*, pp. 477–493. Springer, Cham.
51. Moon, G., Lee, K. M. (2020). I2l-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. *Computer Vision–ECCV 2020*, pp. 752–768.
52. Choi, H., Moon, G., Lee, K. M. (2020). Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. *Computer Vision–ECCV 2020*, pp. 769–787.
53. Clever, H. M., Erickson, Z., Kapusta, A., Turk, G., Liu, K. et al. (2020). Bodies at rest: 3D human pose and shape estimation from a pressure image using synthetic data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6215–6224. Seattle, WA, USA.
54. Gupta, V. (2020). Back to the future: Joint aware temporal deep learning 3D human pose estimation. arXiv:2002.11251.
55. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W. et al. (2021). Pose recognition with cascade transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1944–1953.
56. Liu, Z., Chen, H., Feng, R., Wu, S., Ji, S. et al. (2021). Deep dual consecutive network for human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 525–534.
57. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L. et al. (2021). Hybrik: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3383–3393.
58. Cheng, Y., Wang, B., Yang, B., Tan, R. T. (2021). Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7649–7659.
59. Su, K., Yu, D., Xu, Z., Geng, X., Wang, C. (2019). Multi-person pose estimation with enhanced channel-wise and spatial information. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5674–5682.
60. Feng, Z., Zhu, X., Mao, Y. (2018). Fast human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3517–3526.
61. Nie, X., Feng, J., Zhang, J., Yan, S. (2019). Single-stage multi-person pose machines. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6951–6960. Seoul, South Kerean.
62. Li, S., Chan, A. B. (2014). 3D human pose estimation from monocular images with deep convolutional neural network. *Asian Conference on Computer Vision*, vol. 2014, pp. 332–347. Springer, Cham.
63. Sun, X., Shang, J., Liang, S., Wei, Y. (2017). Compositional human pose regression. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2602–2611. Venice, Italy.
64. Li, Z., Oskarsson, M., Heyden, A. (2021). 3D human pose and shape estimation through collaborative learning and multi-view model-fitting. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, vol. 2021, pp. 1888–1897. Waikoloa, USA.
65. Fang, Q., Shuai, Q., Dong, J., Bao, H., Zhou, X. (2021). Reconstructing 3D human pose by watching humans in the mirror. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp, 12814–12823.
66. He, Y., Yan, R., Fragkiadaki, K., Yu, S. I. (2020). Epipolar transformer for multi-view human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1036–1037. Seattle, WA, USA.

67. Chen, L., Ai, H., Chen, R., Zhuang, Z., Liu, S. (2020). Cross-view tracking for multi-human 3D pose estimation at over 100 fps. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3279–3288. Seattle, WA, USA.
68. Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W. (2019). Cross view fusion for 3D human pose estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4342–4351. Long Beach, CA, USA.
69. Zhang, T., Huang, B., Wang, Y. (2020). Object-occluded human shape and pose estimation from a single color image. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA.
70. Zhen, J., Fang, Q., Sun, J., Liu, W., Jiang, W. et al. (2020). Smap: Single-shot multi-person absolute 3D pose estimation. *European Conference on Computer Vision*, pp. 550–566. Springer, Cham.
71. Wang, C., Li, J., Liu, W., Qian, C., Lu, C. (2020). Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3D pose estimation. *European Conference on Computer Vision*, pp. 242–259. Springer, Cham.
72. Tu, H., Wang, C., Zeng, W. (2020). *VoxelPose: Towards multi-camera 3D human pose estimation in wild Environment.* Springer, Cham.
73. Chen, C. H., Tyagi, A., Agrawal, A., Drover, D., Stojanov, S. et al. (2019). Unsupervised 3D pose estimation with geometric self-supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5714–5724. Long Beach, CA, USA.
74. Rhodin, H., Spörri, J., Katircioglu, I., Constantin, V., Meyer, F. et al. (2018). Learning monocular 3D human pose estimation from multi-view images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8437–8446. Salt Lake City, UT, USA.
75. Chen, X., Lin, K. Y., Liu, W., Qian, C., Lin, L. (2019). Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10895–10904. Long Beach, CA, USA.
76. Kocabas, M., Karagoz, S., Akbas, E. (2019). Self supervised learning of 3D human pose using multi-view geometry. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2019, pp. 1077–1086. Long Beach, CA, USA.
77. Cao, Z., Simon, T., Wei, S. E., Sheikh, S. Y., (2017). Realtime multi-person 2D pose estimation using part affinity fields. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2017, pp. 7291–7299. Honolulu, HI, USA.
78. Novotny, D., Ravi, N., Graham, B., Neverova, N., Vedaldi, A. (2019). C3DPO: Canonical 3D pose networks for non-rigid structure from motion. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7688–7697. Seoul, South Korea.
79. Kong, C., Lucey, S. (2019). Deep non-rigid structure from motion. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, vol. 2019, pp. 1558–1567. Seoul, South Korea.
80. Kocabas, M., Karagoz, S., Akbas, E. (2019). Selfsupervised learning of 3D human pose using multi-view geometry. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2019, pp. 1077–1086. Long Beach, CA, USA.
81. Zhou, B., Cui, Q., Wei, X. S., Chen, Z. M. (2020). BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9719–9728. Seattle, WA, USA.
82. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q. et al. (2020). Srnet: Improving generalization in 3D human pose estimation with a split-and-recombine approach. *European Conference on Computer Vision*, pp. 507–523. Springer, Cham.
83. Xie, R., Wang, C., Wang, Y. (2020). MetaFuse: A pre-trained fusion model for human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2020, pp. 13686–13695. Seattle, WA, USA, IEEE.

84. Wandt, B., Rosenhahn, B. (2019). RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2019, pp. 7782–7791. Long Beach, CA, USA.
85. Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., Black, M. J. (2019). AMASS: Archive of motion capture as surface shapes. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5442–5451. Seoul, South Korea.
86. Li, S., Ke, L., Pratama, K., Tai, Y. W., Tang, C. K. et al. (2020). Cascaded deep monocular 3D human pose estimation with evolutionary training data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6173–6183. Seattle, WA, USA.
87. Pavlakos, G., Zhou, X., Daniilidis, K. (2018). Ordinal depth supervision for 3D human pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA.
88. Martinez, J., Hossain, R., Romero, J., Little, J. J. (2017). A simple yet effective baseline for 3D human pose estimation. *ICCV, 2017,* 2640–2649. DOI 10.1109/ICCV.2017.288.
89. Wang, J., Huang, S., Wang, X., Tao, D. (2019). Not all parts are created equal: 3D pose estimation by modeling bi-directional dependencies of body parts. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, vol. 2019, pp. 7771–7780. Seoul, South Korea.
90. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M. (2019). 3D human pose estimation in video with temporal convolutions and semi-supervised training. *CVPR, 2019,* 7753–7762. DOI 10.1109/CVPR41558.2019.
91. Kanazawa, A., Black, M. J., Jacobs, D. W., Malik, J. (2018). End-to-end recovery of human shape and pose. *CVPR, 2018,* 7122–7131. DOI 10.1109/CVPR.2018.00744.
92. Luvizon, D. C., Picard, D., Tabia, H. (2018). 2D/3D pose estimation and action recognition using multitask deep learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2018, pp. 5137–5146. Salt Lake City, UT, USA.
93. Xu, J., Yu, Z., Ni, B., Yang, J., Yang, X. et al. (2020). Deep kinematics analysis for monocular 3D human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 899–908. Seattle, WA, USA.
94. Hossain, M., Little, J. J. (2017). *Exploiting temporal information for 3D pose estimation.* Springer, Cham.
95. Lee, K., Lee, I., Lee, S. (2018). Propagating LSTM: 3D pose estimation based on joint interdependency. *European Conference on Computer Vision (ECCV)*, pp. 119–135. Munich, Germany. DOI 10.1007/978-3-030-01234-2.
96. Hossain, M. R. I., Little, J. J. (2018). Exploiting temporal information for 3D human pose estimation. *ECCV, 2018,* 68–84. DOI 10.1007/978-3-030-01249-6.
97. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T. et al. (2019). Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, vol. 2019, pp. 2272–2281. Seoul, South Korea.
98. Ma, X., Su, J., Wang, C. (2021). Context modeling in 3D human pose estimation: A unified perspective. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2021, pp. 6238–6247.
99. Levie, R., Monti, F., Bresson, X., Bronstein, M. M. (2017). Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing, 67(1),* pp. 97–109. DOI 10.1109/TSP.2018.2879624.
100. Li, R., Wang, S., Zhu, F., Huang, J. (2018). Adaptive graph convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32. New Orleans, Lousiana, USA.
101. Bresson, X., Laurent, T. (2017). Residual gated graph convnets. arXiv:1711.07553.

102. Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T. et al. (2015). Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems, 2015,* 28.
103. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C. (2014). Human3.6 M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 36(7),* 1325–1339. DOI 10.1109/TPAMI.2013.248.
104. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P. (2016). Direct prediction of 3D body poses from motion compensated sequences. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2019, pp. 991–1000. Las Vegas, NV, USA. DOI 10.1109/CVPR.2016.113.
105. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K. G., Daniilidis, K. (2016). Sparseness meets deepness: 3D human pose estimation from monocular video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4966–4975. Las Vegas, NV, USA.
106. Du, Y., Wong, Y., Liu, Y., Han, F., Gui, Y. et al. (2016). Marker-less 3D human motion capture with monocular image sequence and height-maps. *European Conference on Computer Vision*, pp. 20–36. Springer, Cham.
107. Chen C, R. D. (2017). Human pose estimation 2D pose estimation matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2017, pp. 7035–7043. Honolulu, HI, USA.
108. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y. (2016). Deep kinematic pose regression. *European Conference on Computer Vision*, pp. 186–201. Springer, Cham.
109. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H. et al. (2018). 3D human pose estimation in the wild by adversarial learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2018, pp. 5255–5264. Salt Lake City, UT, USA.
110. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y. et al. (2019). 3D hand shape and pose estimation from a single RGB image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2019, pp. 10833–10842. Long Beach, CA, USA.
111. Liu, J., Ding, H., Shahroudy, A., Duan, L. Y., Jiang, X. et al. (2019). Feature boosting network for 3D pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2),* 494–501. DOI 10.1109/TPAMI.34.
112. Ci, H., Ma, X., Wang, C., Wang, Y. (2020). Locally connected network for monocular 3D human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3),* 1429–1442.
113. Schmidtke, L., Vlontzos, A., Ellershaw, S., Lukens, A., Arichi, T. et al. (2021). Unsupervised human pose estimation through transforming shape templates. *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021,* 2484–2494. DOI 10.1109/CVPR46437.2021.00251.
114. Kundu, J. N., Seth, S., Rahul, M. V., Rakesh, M., Radhakrishnan, V. B. et al. (2020). Kinematic-structure-preserved representation for unsupervised 3D human pose estimation. *AAAI Conference on Artificial Intelligence, 34(7),* 11312–11319. DOI 10.1609/aaai.v34i07.6792.
115. Kundu, J. N., Patravali, J., Radhakrishnan, V. B. (2020). Unsupervised cross-dataset adaptation via probabilistic amodal 3D human pose completion. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, vol. 2020, pp. 469–478. Snowmass village, Colorado, USA.
116. Kundu, J. N., Seth, S., Jampani, V., Rakesh, M., Babu, R. V. et al. (2020). Self-supervised 3D human pose estimation via part guided novel image synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2020, pp. 6152–6162. Seattle, WA, USA.
117. Zhang, L., Sun, L., Li, W., Zhang, J., Cai, W. et al. (2021). A joint Bayesian framework based on partial least squares discriminant analysis for finger vein recognition. *IEEE Sensors Journal, 22(1),* 785–794.
118. Fabbri, M., Lanzi, F., Calderara, S., Alletto, S., Cucchiara, R. (2020). Compressed volumetric heatmaps for multi-person 3D pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7204–7213. Seattle, WA, USA.

119. Johnson, J., Karpathy, A., Li, F. F. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440.
120. Kocabas, M., Karagoz, S., Akbas, E. (2019). Self-supervised learning of 3D human pose using multi-view geometry. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2019, pp. 1077–1086. Long Beach, CA, USA.
121. Lipton, Z. C., Berkowitz, J., Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. computer science. arXiv:1506.00019.
122. Shuai, B., Berneshawi, A. G., Modolo, D., Tighe, J. (2020). Multi-object tracking with siamese track-RCNN. arXiv:2004.07786.
123. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, 34(6),* 1–16. DOI 10.1145/2816795.2818013.
124. Zhang, H., Ouyang, H., Liu, S., Qi, X., Shen, X. et al. (2019). Human pose estimation with spatial contextual information. arXiv:190101760.
125. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C. (2020). Distribution-aware coordinate representation for human pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2020, pp. 7093–7102. Seattle, WA, USA.
126. Pham, H. H., Salmane, H., Khoudour, L., Crouzil, A., Velastin, S. A. et al. (2020). A unified deep framework for joint 3D pose estimation and action recognition from a single RGB camera. *Sensors, 20(7),* 1825. DOI 10.3390/s20071825.
127. Lin, J., Lee L. H. G. (2019). Trajectory space factorization for deep video-based 3D human pose estimation. arXiv:1908.08289.
128. Cai, W., Zhai, B., Liu, Y., Liu, R., Ning, X. (2021). Quadratic polynomial guided fuzzy C-means and dual attention mechanism for medical image segmentation. *Displays, 70,* 102106. DOI 10.1016/j.displa.2021.102106.
129. Mehta, D., Sotnychenko, O., Mueller, F. (2020). XNect: Real-time multi-person 3D motion capture with a single RGB camera. *ACM Transactions on Graphics, 39(4)*. DOI 10.1145/3386569.3392410.
130. Yan, S., Xiong, Y., Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Journal on Image and Video Processing, 78*.
131. Kocabas, M., Athanasiou, N., Black, M. J. (2020). VIBE: Video inference for human body pose and shape estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* vol. 2020, pp. 5253–5263. Seattle, WA, USA.
132. Habermann, M., Xu, W., Zollhofer, M., Pons-Moll, G., Theobalt, C. (2020). Deepcap: Monocular human performance capture using weak supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2020, pp. 5052–5063. Seattle, WA, USA.
133. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B. et al. (2015). Panoptic studio:A massively multiview system for social motion capture. *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015, pp. 3334–3342. Santiago, Chile.
134. Robertini, N., Casas, D., Rhodin, H., Seidel, H. P., Theobalt, C. (2016). Model-based outdoor performance capture. *2016 Fourth International Conference on 3D Vision (3DV)*, vol. 2016, pp. 166–175. Stanford, CA, USA.
135. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S. et al. (2018). Single-shot multi-person 3D pose estimation from monocular RGB. *2018 International Conference on 3D Vision (3DV)*, vol. 2018, pp. 120–130. Verona, Italy. IEEE.
136. Joo, H., Simon, T., Sheikh, Y. (2018). Total capture: A 3D deformation model for tracking faces, hands, and bodies. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8320–8329. Salt Lake City, USA.

137. Zhang, W., Liu, Z., Zhou, L., Leung, H., Chan, A. B. (2017). Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3D human pose estimation. *Image and Vision Computing, 61,* 22–39. DOI 10.1016/j.imavis.2017.02.002.

138. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H. S. et al. (2019). Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2019, pp. 10863–10872. Long Beach, CA, USA.

139. Kim, W., Ramanagopal, M. S., Barto, C., Yu, M. Y., Rosaen, K. et al. (2019). Pedx: Benchmark dataset for metric 3-D pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters, 4(2),* 1940–1947. DOI 10.1109/LRA.2019.2896705.

140. Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C. et al. (2020). Deep learning-based human pose estimation: A survey. arXiv:2012.13392.