





A Review of Human Pose Estimation Methods in Markerless Motion Capture

Huaiming Ji¹, Li Wang¹, Yuwei Zhang¹, Zhi Li¹ and Chenglong Wei²

¹ Faculty of Mechanical Engineering, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

² School of Mechanical Engineering, Shandong University, Jinan 250014, China

Corresponding author: Li Wang, wli@qlu.edu.cn

Abstract. Human pose estimation aims at detecting human joint points from input data such as images and videos, or building a human body model for motion analysis. However, due to the ambiguities of human occlusion, depth blurring and the lack of training data, the accuracy of motion capture is still far from satisfactory. This paper reviews the advances of human pose estimation methods in markerless motion capture since 2019. We propose three types of representations for the human body, and detect that a unified volumetric model provides more detailed motion representation. We introduce datasets and evaluation metrics widely used for 2D and 3D pose estimation. Comparisons and discussions are conducted on different model frameworks for human pose estimation based on accuracy, robustness, and speed, summarizing the strengths and weaknesses of various methods. We discover that pose estimation methods based on the Transformer framework exhibit better accuracy and robustness, while kinematic and physical knowledge greatly assist in solving 3D pose estimation. Additionally, lightweight methods are often overlooked in research. In conclusion, this paper serves as a guide for researchers interested in the field and assists newcomers in selecting and developing human pose estimation methods.

Keywords: motion capture, deep learning, computer vision, pose estimation, virtual reality

DOI: <https://doi.org/10.14733/cadaps.2024.392-423>

1 INTRODUCTION

As a popular computer vision task, markerless motion capture aims to extract body part positions and pose from markerless images or videos utilizing human pose estimation (HPE) methods. Markerless motion capture shows broad application prospects in different fields. For example, in game development[100], it can be used to generate realistic character animations and enhance the immersion of games. In the field of motion medicine[7, 11, 43, 44, 92], markerless motion capture can assist in rehabilitation training to help patients regain normal movement functions. In virtual

reality technology, it can be used for body interaction, allowing users to interact with virtual environments in a natural way. In industrial design[5, 45], it can assist designers in optimizing the human-computer interaction of products, improving user experience and work efficiency.

Nowadays, the most commonly used motion capture methods include marker-based motion capture, inertial motion capture and markless motion capture. Marker-based motion capture methods use infrared cameras and markers to identify body postures, but this method requires trained personnel to manually place markers, and it is prone to errors, usually with an error close to 10mm or 10°[18]. Additionally, when markers are placed close to each other, they can easily occlude each other, making motion tracking difficult. To address this issue, one approach is to increase the number of cameras capturing the markers from different angles, but this increases computational costs and complexity[82]. Another commonly used motion capture method is inertial motion capture, which uses inertial measurement units (IMUs) to track the positions of body joints. Inertial motion capture provides quantitative, reliable, and easy-to-collect motion capture data. However, this method also has limitations as IMUs themselves have drift and are susceptible to external magnetic fields[5, 29].

Markerless motion capture uses highly trained neural networks to estimate the joint center positions of data in motion images or videos. It reduces dependence on professionals and improves the reliability of data by eliminating the need for physical markers. Currently, there are some successful commercial markerless motion capture systems available, such as Theia3D and VisionPose. The current research trend of markerless motion capture technology mainly focuses on improving accuracy and stability while reducing hardware costs and complexity. Some research directions include the combination of multimodal sensors, such as depth cameras and inertial sensors [27, 32, 41, 129, 134], to obtain more accurate and stable pose estimation results. Additionally, researchers are also working on addressing pose estimation challenges in multi-person scenarios, such as interference and conflicts between poses.

In recent years, many scholars have discussed the usage of HPE in markerless motion capture from different applications. Colyer et al. [18] studied the applications in biomechanics and discussed the difference between the markerless motion capture methods and the marker-based motion capture ones. They pointed out that the current markerless motion capture methods have not yet achieved the accuracy required by most analysis in motion science, but the recent methods have greatly improved accuracy. Mathis et al. [68] reviewed deep learning-based markerless motion capture methods, and detailed the process and applications of markerless motion capture. However, this survey focused more on animal motion capture methods and did not investigate human capture methods. Therefore, this survey focuses on markerless motion capture methods after 2019, with a focus on the core of markerless motion capture methods - HPE methods.

The review searched for relevant literature on markerless motion capture through different combinations of keywords like "motion capture", "markerless", "human pose estimation", "2D" and "3D" on Google Scholar, IEEE Xplore, and Science Direct. References from the selected articles were also obtained. The review selected English journal articles or conference papers related to markerless motion capture systems between 2019 and 2023, aiming to provide useful information for researchers interested in this field, explore the application of markerless motion capture in the real world, and discuss viable directions and challenges that need to be overcome in this technology.

This paper focuses on reviewing the HPE methods for markerless motion capture, including not only 2D HPE, but also 3D HPE. We initiate our discussion in Section 2 by exploring various representations of the human body. Subsequently, we introduce a range of existing human body datasets as well as corresponding evaluation metrics. Section 4 is dedicated to discussing the equipment employed in capturing HPE. In Section 5, we delve into the details of prior 2D and 3D methodologies and highlight the lightweight-based methods, which have been relatively underemphasized in prior reviews. Section 6 involves a critical analysis of recent HPE methods, focusing on their accuracy, robustness, and speed. Finally, in Section 7, we outline potential areas of future research in this field.

2 HUMAN BODY REPRESENTATION

Human body model is a critical aspect of HPE, which aims to extract relevant keypoints and features from input data. Human body is usually represented in three common types: skeletal model, planar model, volume model, as shown in Figure 1.

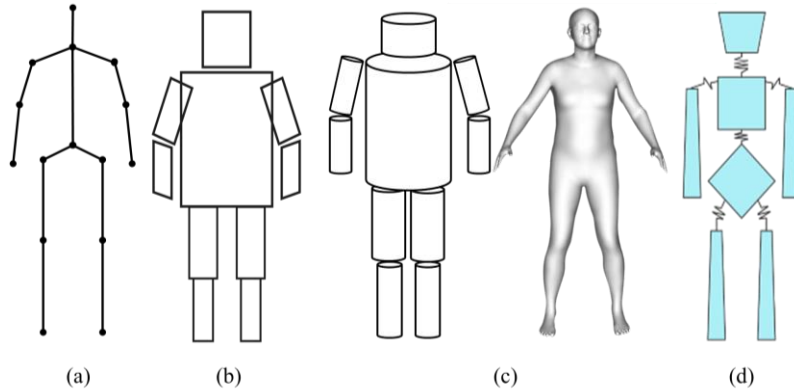


Figure 1: Human body model. (a) Skeleton model where simple graphics such as dotted lines are used to represent the shape of human body. (b) Cardboard model which regards the body as an organic connection of different blocks. (c) Volume model which represents human body by using a collection of simple geometry. (d) Pictorial Structure Model introduced by Felzenszwalb and Huttenlocher [21].

2.1 Skeleton Model

Skeleton model is a representation where the skeleton joints are connected with each other, as shown in Figure 1(a). The parameters are defined as the lengths of adjacent joints and the rotation angles. Felzenszwalb and Huttenlocher [21] introduce the Pictorial Structure Model (PSM), as shown in Figure 1(d). PSM is a widely used graph model that has been successfully applied in 2D HPE and 3D HPE. The main idea is to optimize the joint positions by considering both their appearance and spatial relationships. Recursive Pictorial Structure Model (RPSM)[79] recursively discretizes the ground space around each joint position into a finer-grained grid, which is more suited to recover 3D pose from multi-view 2D pose heatmap. Ma et al. [64] proposed ContextPose by taking advantages of the PSM and Graph Neural Network (GNN), allowing the usage of limb length constraints for 3D HPE.

2.2 Planar Model

In addition to skeletal models, planar models are often used to represent the appearance of the human body. The plane model commonly represents the human body using rectangles that are similar to the human silhouette, as shown in Figure 1(b). A common example is the Cardboard model[40], which contains one limb and eight half-limbs. Each body part is represented by an average of RGB colour. The other commonly used planar model is the Active Shape Models (ASMs)[19].

2.3 Volume Model

The skeletal model can only represent human movement and cannot express shape and texture information. In contrast, volume models use geometric shapes to represent model components, such as cylinders, cones, spheres and other geometric shapes, as shown in Figure 1(c). Volume models can be categorized into rigid body models and non-rigid body models based on their deformation characteristics. Rigid body models refer to models that do not undergo shape deformation when

subjected to external forces, such as cylinder models and skeletal models. Non-rigid body models, on the other hand, allow for deformation. Given the non-rigid nature of the human body, recent focus in research has been on more complex non-rigid body models to provide a more accurate description of human behaviour. A representation of non-rigid body models is the triangular mesh, which is usually derived from high-definition 3D scans. Loper et al. [60] proposed a learn-based method, Skinned Multi-person Linear Model (SMPL), which combines human shape, position-related shape change together. The parameters consist of resting pose templates, blending weights, pose-related blending shapes, identity-related blending shapes, and regressors from vertex to joint position. Unlike previous models, the SMPL can simulate human skin textures and dynamic soft-tissue deformations. Osman et al. [73] proposed STAR to formalize sparse correction blend shapes, reducing the parameters by 80%. STAR extended pose correction formulation by correcting pose and shape regression. This change makes the deformation more realistic and consistent with health-care applications. Saito et al. [86] proposed SCANimate, which can learn detailed clothing mannequins from raw scans.

HPE can be divided into body pose, face pose and hand pose estimation. However, SMPL can only be used for body estimation. Some researchers try to combine the three models together to achieve whole-body pose estimation. Based on the SMPL-H[84], Pavlakos et al. [75] extended the fully articulated hand and face models, which inherits the advantages of the SMPL. However, the representations of facial expression, hand and body are isolated. GHUM & GHUM(ite)[111] is capable of driving facial expressions, hand and full body movements. It consists of a medium-resolution of GHUM with 10,168 vertices and low-resolution of GHUM(ite) with 3,194 vertices.

The human body is a complex organism composed of a torso, limbs, and a head, encompassing both kinematic information and surface texture. In the previous context, we introduced three types of human body models, where skeletal models and volumetric models represented by SMPL have gained attention in research. Particularly in recent years, the SMPL model has become a widely used model for human representation in various datasets. In 3D HPE, utilizing the architecture of a human body model can help infer occluded torso or body joints, leading to improved robustness. The model needs to adhere to kinematic knowledge and general physics principles to reduce unrealistic inferences. Furthermore, models like GHUM and SMPL-X integrate human body models to provide more detailed expressions of human motion. However, volumetric models have a large number of parameters, so simplifying the model while maintaining its fidelity to real human bodies is one of the focuses of current research.

3 DATASETS AND EVALUATION METRICS

As human movement gestures change a lot in different environments, it is hard to create a generic dataset to meet every application requirement. Thus, many researchers often choose datasets that are appropriate to their task. In the following, we introduce several existing datasets and evaluation metrics that have been commonly used over the last few years. According to the different dimensions of human pose information included in the dataset, we divide the dataset into 2D and 3D datasets.

3.1 2D Datasets

Max Planck Institute Informatics (MPII) Human Pose [2] is the first large-scale benchmark dataset for HPE, in which images are downloaded from YouTube. It contains 25,000 images covering 410 human activities such as dancing and running.

Microsoft Common Objects in COntext (MSCOCO) [58] was created in 2014 for object detection, keypoint detection and instance segmentation. In the 2017 version, the train and val sets change from 83,000/41,000 to 118,000/5,000. Also, they introduce a new unannotated dataset of 123,000 images, labelled with 17 keypoints of the human body, as shown in Figure 2. COCO-WholeBody is a variant of COCO [38] and contains 133 marked points (68 on the face, 42 on the hands and 23 on the body and feet).

The Pose Track [1] is a 2D video-based body dataset, containing approximately 1356 video sequences, 46,000 annotated video frames and 276,000 body pose annotations, as shown in Figure 3. It is mainly used for multi-person pose estimation, where each person has a unique track ID and is tagged with 15 body keypoints. AI Challenger Human Keypoint Detection (AIC-HKD) [108] is the largest 2D human pose dataset. It contains 300,000 high-resolution images with multiple people and various poses by Internet search engines, in which 210,000 images are used for training. The CrowdPose [53] is a HPE dataset for dense crowds. A total of 20,000 high-quality images were captured from 80,000 people. The Human-in-Events (HiEve) [59] is a large video-based dataset, consisting of 32 video sequences from nine different scenes with 49,820 frames, 1,302,481 bounding box, 2687 trajectory annotations, 1,099,357 human pose annotations and 56,643 action annotations.



Figure 2: Examples from the COCO dataset [58].



Figure 3: Examples from the Pose Track dataset [1].

Dataset Name	Year	Image/Video	Single-Person/Multi-Person	Joints	Number of images			Evaluation metrics
					Train	Val	Test	
MPII[2]	2014	Image	Single	16	29k	-	12k	PCPm/PCKh
		Image	Multiple	16	3.8k	-	1.7k	mAP
COCO2016	2016	Image	Multiple	17	45k	22k	80k	AP
COCO2017	2017	Image	Multiple	17	64k	2.7k	40k	AP
COCO-WholeBody [38]	2020	Image	Multiple	133		200k		mAP & mAR, OKS
AIC-HKD[108]	2017	Image	Multiple	14	210k	30k	60k	AP
PoseTrack[1]	2017	Video	Multiple	15	292	50	208	mAP
CrowdPose [53]	2019	Image	Multiple	14	10k	2k	8k	mAP
HiEve[59]	2020	Both	Multiple	14	19video	-	13video	AP

Table 1: Datasets for 2D pose estimation.

3.2 3D Datasets

HumanEva [91] contains over 250,000 images acquired by a marker motion capture system. Human3.6M [34] contains 3.6 million human poses and is organized into 15 training scenarios. The dataset was captured using four digital cameras, including one time-of-flight sensor and ten motion cameras, as shown in Figure 4.



Figure 4: Examples from the Human3.6M dataset[91].



Figure 5: Examples of the CMU Panoptic dataset[39].

MPI-INF-3DHP [69] is a 3D HPE dataset that contains complex outdoor scenes. It consists of over 1.3 million frames which were captured from 14 cameras. CMU Panoptic dataset [39] (as shown in Figure 5) uses a large camera system with 480 lenses to capture people's interactions in various social activities. Garau et al. [25] applied PanopTOP to the Panoptic dataset and generated a new dataset, named PanopTOP31K.

3D Poses in the Wild (3DPW) [96] is a large 3D dataset for HPE, using 9-10 inertial sensors to track objects. Over 51,000 frames of the accurate 3D pose were captured, covering movements from walking, walking upstairs, drinking coffee to riding in a car.

3DPeople [78] is the first dataset for HPE with clothes, containing approximately 2 million images from 40 men and women with 70 actions. The data is acquired using four cameras and annotated with RGB colours, 3D skeletons, body parts, fabric segmentation masks, depth maps, optical flow and camera parameters. AMASS [66] contains fifteen sub-datasets and their own recorded human data that were mapped to SMPL [60]. HUMBI [121] is a large multi-view dataset, which captures 772 different objects through 107 simultaneous HD cameras. SMPLy [50] is a new dataset containing 742 objects from 567 scenes, yielding 24,428 images covering indoor and outdoor scenes.



Figure 6: Examples from the AGORA dataset[74].

Avatars in Geography Optimized for Regression Analysis (AGORA) [74] includes 4240 scans covering more than 350 different objects, in which 1051 scans were used for testing (3387 images), 2930 scans for training (14,529 images), and 259 for validation (1225 images), as shown in Figure 6. It contains 3D poses and shapes of children and extends SMPL-X space to capture the differences between children and adults.

DensePose COCO [30] is a large dataset for dense HPE, containing 50,000 real markers with over 5 million actual annotations, as shown in Figure 7. It was constructed with an annotation pipeline so as to achieve image-to-surface correspondence.



Figure 7: Examples of the DensePose COCO dataset [30].

Dataset Name	Year	Capture system	Environment	Joints	Number of frames/videos			Evaluation metrics
					Train	Val	Test	
HumanEva-I[91]	2010	Marker-based MoCap	Lab environment	15	≈6.8k	≈6.8k	≈24k	MPJPE
Human3.6M[34]	2014	Marker-based MoCap	Lab environment	17	≈1.5M	≈0.6M	≈1.5M	MPJPE
CMU Panoptic[39]	2016	Markerless MoCap	Lab environment	15	65 videos(5.5 hours)			3DPCK
MPI-INF-3DHP[69]	2017	Markerless MoCap	Indoor and outdoor	15	≈1.3 M			3DPCK
3DPW[96]	2018	Hand-held cameras with IMUs	Indoor and outdoor	18	60 videos(≈51k frames)			MPJPE&MPJAE
DensePose COCO[30]	2018	-	Outdoor	UV	50k			RCP
3DPeople[78]	2019	-	Indoor and outdoor	mesh models	2.5M			-
AMASS[66]	2019	Marker-based MoCap	Indoor and outdoor	SMPL,SMPL-X	2420.86 minutes			-
HUMBI[121]	2020	Markerless MoCap	Lab environment	SMPL	93k+17.3M+24M+26M			AUC
SMPLY[50]	2020	Markerless MoCap	Indoor and outdoor	SMPL	567sequences (24428frames)			MPJPE&3DPCK
3DOH50K[125]	2020	Markerless Mocap	Indoor	SMPL	50310	-	1290	MPJPE
AGORA[74]	2021	-	Outdoor	SMPL, SMPL-X	14k	1.2k	3k	MPJPE&MVE
Self contact[71]	2021	-	Indoor	SMPL-X	8752 self-contact, 16752 no self-contact and 9491 unclear poses			MPJPE/MV2VE
MPSD[72]	2021	-	Indoor and outdoor	volumetric voxel	450k			CD, P2S, IoU

Table 2: Datasets for 3D pose estimation. UV refers to texture coordinates used to map textures (such as images or patterns) onto the surface of a three-dimensional model. In this context, U represents the horizontal coordinate, and V represents the vertical coordinate.

3D Occlusion Human 50K (3DOH50K) [125] is constructed for the occlusion problem and contains 51,600 images captured in real scenes with six cameras. It provides 2D and 3D annotations and SMPL parameters for the generated meshes.

Self Contact dataset [71] is a self-contact dataset proposed by Lea Mvller et al. Many existing datasets deliberately avoid physical occlusion and interaction. In contrast, this dataset contains self-exposure grid data from six subjects.

Multiple People Synthetic (MPSD)[72] is a single-image multi-person dataset containing 3D human models, RGB images, depth maps, instance segmentation and 6D degrees-of-freedom spatial locations.

Previous datasets seldom contain children and people with complex body shapes, which might not meet the requirements in medical applications [57]. Most of them focus on simple movements, with little attention given to interacting activities engaged in body contact. Handling occlusion and making accurate representation are the key problems for multi-person motion capture. In Table 2, we have summarized commonly used 3D datasets as well as the latest datasets. In recent years, researchers have started to focus on the construction of outdoor datasets, which often consist of real or synthetic images of outdoor scenes. For example, AGORA and AMASS datasets are two such examples. It is evident that the latest datasets tend to use volumetric models such as SMPL or SMPL-X to represent the human body structure. The use of full-body models like SMPL-X enables better representation of detailed human motion.

3.3 Evaluation Metrics

Currently, numerous metrics have been proposed to evaluate the accuracy of pose estimation methods. In the following, we introduce the most commonly used metrics. The following metrics mentioned in the text are only used to measure the accuracy of pose estimation methods on the

corresponding datasets, and do not include evaluation metrics for markerless motion capture in practical applications.

Percentage of Correct Parts (PCP)[22] measures the percentage of correctly predicted parts, with higher value indicating better performance.

Percentage of Correct Key Points (PCK)[116] measures the percentage of correctly predicted keypoints. A predicted body joint is defined as correct if the distance from the actual joint is within a certain threshold.

$$PCK = \frac{\sum_i \delta\left(\frac{d_i}{S} \leq T\right)}{\sum_i 1} \quad (3.1)$$

where d_i indicates the distance between the i th predicted keypoint and the valid keypoint. When $\frac{d_i}{S}$ is less than or equal to the threshold T , the function $\delta()$ returns 1. In 3D pose estimation, the deformed body of PCK is often used, known as 3DPCK.

Percentage of Detected Joints (PDJ)[94] is commonly used to evaluate the 2D pose estimation models. The distance between the predicted joint and the actual joint is considered to be correct if it is beyond a certain fraction of the trunk diameter. Toshev and Szegedy [94] suggested that PDJ could be used as an alternative to the PCP metric to address the problem of short limbs.

Average Precision (AP) is a generalized measure to calculate the performance degradation of a model [58]. AP measure is an index to measure the accuracy of keypoints detection according to precision and recall. Mean Average Precision (mAP) is the mean of average precision over all classes. It has been applied to the evaluations of MPII[2] and PoseTrack[1]. Average Recall (AR) is another metric used in the COCO keypoint evaluation. AP, AR, and their variants are reported based on an analogous similarity measure: object keypoint similarity (OKS) which plays the same role as the Intersection over Union (IoU).

Mean Per Joint Position Error (MPJPE)[34] calculates the mean Euclidean distance error between the actual joint coordinate and the measured joint coordinate in the Human3.6M[34].

Geodesic point similarity (GPS)[30] is used to measure object keypoint similarity.

$$GPS_j = \frac{1}{|P_j|} \sum_{p \in P_j} \exp\left(\frac{-g(i_p, \hat{i}_p)^2}{2k^2}\right) \quad (3.2)$$

where P_j represents the set of actual points annotated on the person instance, i_p is the vertex value estimated by the model at point p , and \hat{i}_p is the actual point of p .

Pose Structure Score (PSS) [47] is used to evaluate the structural soundness between the predicted pose and the actual pose.

$$PSS_{p,q} = \delta(C(p), C(q)) \quad (3.3)$$

Where

$$C(p) = \arg \min_k \|p - \mu_k\|_2^2, \delta(i, j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (3.4)$$

Here, p is the predicted pose and q is the actual pose.




4 CAPTURING EQUIPMENT

The RGB camera is one of the most commonly used camera types in motion capture. However, it has inherent limitations in capturing 3D information as it cannot provide depth information. To overcome this limitation, depth cameras have emerged as a viable alternative and have been

adopted by many motion capture researchers and users. With the development of depth-sensing technology and machine learning, easy-to-use equipment for 3D motion analysis have been commercially available. Microsoft Kinect V1 (also known as Xbox 360 Kinect) was released in 2010 for capturing body movement in video games. It calculates 20 keypoints coordinates (shown in Table 3) using a depth map obtained from a depth sensor[118]. The difference between depth cameras and traditional cameras for markerless motion capture is whether it produces depth images[18]. The depth information is helpful to alleviate the problems of shadows, poor lighting conditions and cluttered environments[18]. Generating depth images relies on the technology of structured light (SL) and time of flight (ToF). The Kinect V1 uses an infrared structured light 2D camera to project deformable scattering patterns onto the scene. Kinect V2 used ToF to measure distances [17]. In addition, Azure Kinect is a new type of camera with the highest depth camera resolution. In Table 3, we summarize the three Kinect depth cameras.

Although a single depth camera can improve the performance of human capture, it is still less effective in dealing with occlusions among objects. For complicated tasks, several works have used multiple Kinect sensors to complete body scans. Bortolini et al. [6] used four sets of Kinect V2s to obtain human joint angles and movement paths for ergonomic analysis. Jiang et al. [37] used two Kinect sensors to collect human skeletal information and make data fusion from joint angles. Large errors might be produced if the Kinect sensors are not placed in the recommended positions [77]. Another approach to improve the robustness of Kinect in complex environment is to correct data errors by reconstructing the unreliable part of Kinect poses [90]. Shum et al. [90] proposed an optimization method to construct a motion database. They calculated the reliability values of the pose joints and used them as weights to extract a set of kinematically similar poses. The local principal component is further used to optimize the poses. Plantard et al. [76] proposed a method that uses filtered pose maps to represent the intrinsic relationships between poses.

Existing research works mainly focus on images captured from conventional RGB and RGB-D cameras. The development of event cameras [24] offers new opportunities for human pose capture. As an emerging bionic imaging sensor, event cameras differ in many ways from conventional frame-based cameras. Instead of capturing images at a fixed frequency, event cameras asynchronously measure the change of brightness and refer to this change as an event.

	<i>Kinect V1</i>	<i>Kinect V2</i>	<i>Azure Kinect</i>
			
Released date	June 2010	July 2014	June 2019
Color camera resolution	1280×720px@12fps 640×480px@30fps	1920×1080px@30fps	3840×2160px@30fps
Depth camera resolution	320×240@30fps	512×424@30fps	NFOV unbinned—640 × 576px@ 30 fps NFOV binned—320 × 288px @ 30 fps WFOV unbinned—1024 × 1024px @ 15 fps WFOV binned—512 × 512px @ 30 fps
Depth FOV	57°H×43°V	70.6°H×60°V	NFOV unbinned—75°×65° NFOV binned—75°×65° WFOV unbinned—120°× 120° WFOV binned—120°×120°
Optimal measuring range	0.4-4.0 m	0.5-4.5 m	0.5-3.86 m
Depth sensing technology	SL	ToF	ToF
Body tracking technology	-	Random Forests model based on depth images	Deep Neural Network model based on depth images

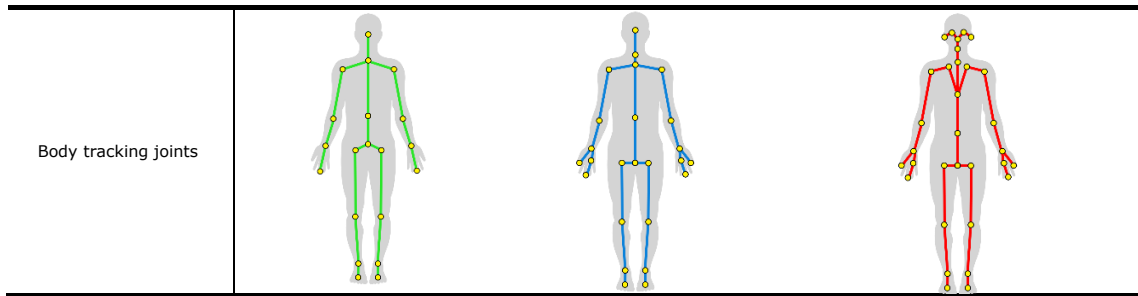


Table 3: Comparisons of three Kinect depth cameras.

The event, expressed as a triplet (x, t, p) , measures the significant change with respect to pixel x , time t , and the polar state p [114]. Compared to the conventional cameras, event cameras have attractive features: high temporal resolution (milliseconds), high dynamic range (140 dB \sim 60 dB), low power consumption, and high pixel bandwidth (kHz level). As a result, event cameras can capture very fast motion without suffering motion blur. DHP19 [9] is an early dataset that uses event cameras for pose estimation. In DHP19, a convolutional neural network (CNN) is designed to estimate 2D human poses and detect 2D human joints from a stream of events. Unfortunately, the high temporal resolution results in very sparse measurements within each frame interval because the changes in brightness from frame to frame are very subtle, and because the event stream encodes only temporal intensity changes, it is difficult to initialize tracking and prevent drift. Event Cap[114] is the first approach to capturing high-speed human motion using a single event camera. It utilized a pre-trained CNN-based human detection module to obtain an initial estimate and took event streams to reconstruct high-frequency motion details. They proposed an optimization algorithm based on hybrid asynchronous batch processing to address the low signal-to-noise ratio, drift, and initialization difficulties of event cameras. Event HPE[135] inferred optical flow from events through unsupervised learning and relied on optical flow features to estimate 3D body shapes. Their method needs to provide or detect the starting pose and shape on the first frame of grayscale image, but avoids the input grayscale image stream like EventCap. Due to the high temporal resolution, low latency, high dynamic range, and low power consumption, event cameras have been applied in a number of computer vision tasks, including camera pose estimation, feature tracking, optical flow, multiview stereo, gesture recognition, motion deblurring and so on.

In addition to the two cameras mentioned above, recent works used fisheye cameras[99, 127] for gait analysis by combining motion capture systems, pressure pads, and other aids[23] to capture human movement.

5 METHODS IN DETAILS

As an important part of markerless motion capture, HPE methods have been developed in recent years. Similar to the classification method for datasets in Section 3, HPE methods can be divided into 2D and 3D pose estimation methods. Additionally, since pose estimation methods in videos often involve time, we summarize them as a separate section.

5.1 2D Pose Estimation

Recent methods have achieved great success with the development of deep learning, it is still challenging to handle images with complex environments and clothing. To address these challenges, previous work classified HPE methods into single-person and multi-person pose estimation according to the number of people in the image. In this section, we introduce the two types of methods in detail.

5.1.1 Single-person pose estimation

The regression-based methods usually directly map the input image to the body joint coordinates through the fully-connected (FC) prediction layer, which is more straightforward than the heatmap-based methods, as shown in Figure 8. But at present, there are only a few methods that use this design. This is due to the lack of spatial and contextual information, which makes the learning of human pose model challenging due to the intrinsic visual ambiguity in joint location [122]. DeepPose [94] is the first regression-based method that uses cascaded deep neural networks to learn keypoints from images. Mao et al. [67] proposed a transformer-based network, which takes the feature map of the CNN as input and outputs the human joint coordinates in a formulaic way. They avoided the destruction of the spatial structure of the convolutional feature map caused by reducing the network parameters of the FC layer using global average pooling. Li et al. [54] provided two solutions for pose estimation: a two-stage approach and a sequential process. They develop an encoder to generate contextual features and passed them to the decoder to detect human body in a bounding box. The cropped image is further fed into a network to obtain the final keypoints.

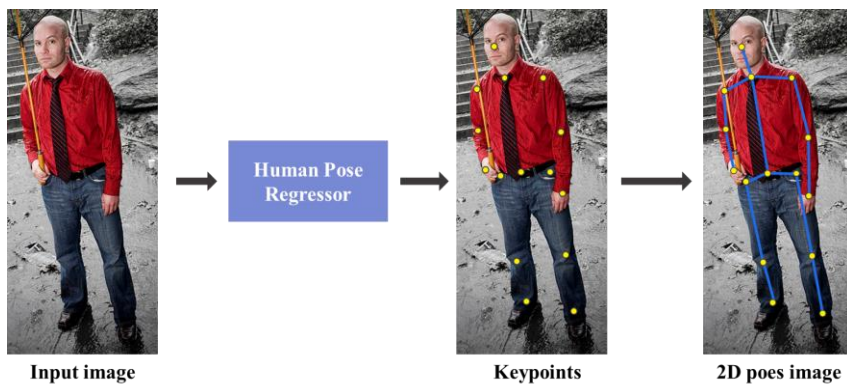


Figure 8: Regression-based methods [131].

HPE is essentially a regression problem. Regression-based methods have achieved good results due to their ability to learn nonlinear feature representations. These methods can obtain continuous output through end-to-end training.

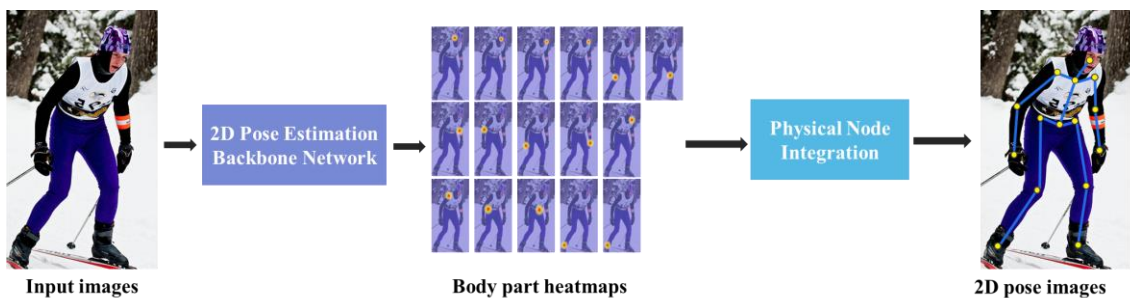


Figure 9: Pipeline of Heatmap-based method [131].

Due to its high accuracy performance, the heatmap-based method has become a standard label representation method for HPE (shown in Figure 9). These methods generate a 2D Gaussian probability distribution centered on the ground-truth joint position, which avoids false positives.

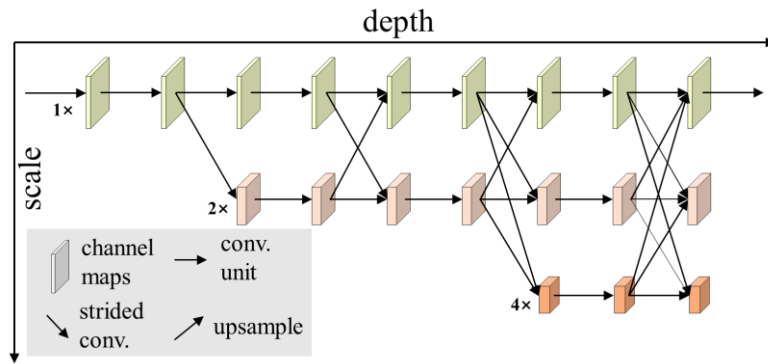


Figure 10: HRNet architecture[93]. Feature maps at different resolutions are exhibited parallelly with downsample or upsample connection.

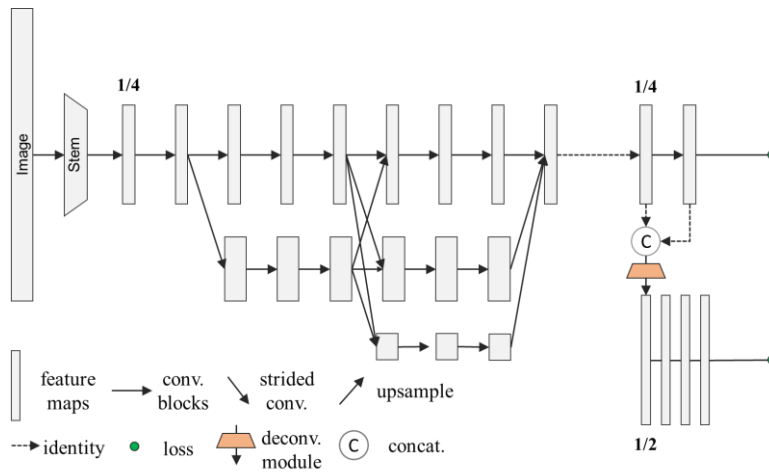


Figure 11: HigherHRNet architecture [13].

Previous high-resolution image processing needs to downsample high-resolution bounding box to low-resolution images, and then upsample back to the original resolution. The Simple Baselines[110] utilizes three back-evolution layers after the backbone network to obtain the final predicted heatmap. It is worth noting that quantisation error can be introduced during the above resolution reduction. Zhang et al. [122] propose a Distribution-Aware Representation of Keypoints (DARK) for more accurate localization of joints in low-resolution images. Through the distribution approximation based on Taylor expansion, the distribution information of heatmap activation is considered comprehensively. In the process of coordinate decoding, the unbiased heat map is used to mitigate the quantization error. Li et al. [57] suggested to use horizontal and vertical coordinates separately for keypoint localization. Instead, HRNet[93] is a representative heatmap-based method that uses multi-stage architectures. Repeated multi-scale fusion allows the high-resolution to low-resolution network to continuously exchange information in parallel. To handle small people in images, HigherHRNet[13] (shown in Figure 11) constructs a high-resolution feature pyramid and generates high-resolution feature map by assigning different resolution of training targets to the feature pyramids. Zhang et al. [128] proposed a compound loss function to measure the similarity between the generated high-resolution images and the real high-resolution images. But converting heatmaps

to joint coordinates requires post-processing, which may not be differentiable, making the framework not end-to-end learnable.

5.1.2 Multi-person pose estimation

Existing works for multi-person pose estimation can be classified into top-down methods (shown in Figure 12) and bottom-up methods (shown in Figure 13). Next, we introduce the problems, advantages and development of the two methods in recent years.

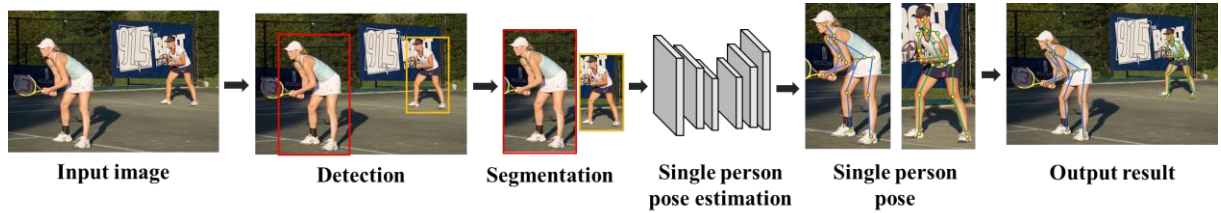


Figure 12: Pipeline of the Top-Down method.



Figure 13: Pipeline of the Bottom-Up method.

The top-down method consists of two stages: human detection and individual pose estimation. The human body detector in the first stage is greatly affected by the environment, and it will be greatly affected when people get close to each other and are blocked. It directly affects the next stage of pose estimation. The second stage uses the existing single-person pose estimator. The top-down method is closely related to the number of people in the image, and each person's posture is estimated separately. In addition, the accuracy and complexity of the estimation depend on the individual pose estimator. Qiu et al. [80] generated initial pose through a heatmap-based method, and then used an Image-Guided Progressive GCN (IGP-GCN) and a cascaded feature adaptation module to optimize the final pose. Li et al. [53] proposed joint candidate Single-Person Pose Estimation (SPPE) to output target and interference candidates, and then performed joint association to obtain joint connection graph.

The bottom-up method has two main steps including keypoints detection and association of body keypoints. To handle images with limited resolution, Kreiss et al. [49] proposed a two-headed neural network, in which the Part Intensity Field (PIF) is used to predict the confidence score, position, and size of joints, and the Part Association Field (PAF) is used to obtain the relationship between joints. Based on the predicted keypoints heatmap and component heatmap, Li et al. [52] obtained candidate keypoints by suppressing nonmaximum probability points.

Method		COCO		MPII		LSP		CrowdPose	#Params	GFLOPs	
		AR	AP	PCKh@0.5	PCP	PCKh@0.5	AP				
Single-person pose estimation	Regression-based methods	TFPose[67]		72.4	90.4					20.4	
		DeepPose[94]				61					
		PRTR[54]		80.2	73.3	89.5				57.2M	37.8
		DARK[122]		81.1	76.2	90.6				63.6M	32.9

Heatmap-based methods	SimDR[57]	81.2	75.9		66.3M	14.6	
	HRNet[93]	81.2	76.3	92.3	63.6M	32.9	
	HigherHRNet[13]	74.9	70.5		67.6	63.8M	154.3
	Zhang et al.[128]			78.9	79.6		
	CAL[98]	82.8	78.2			110.2M	47.2
Top-down methods	MSPN[56]	83.1	78.1	92.6			
	Simple Baseline[110]	79.0	73.7		60.8	68.6M	35.69
	SPPE[53]				66.0		
Multi-person pose estimation	Mask R-CNN[31]		63.1				
	SAHG[120]	74.7	67.5				
	SWAHR[62]		73.8			63.8M	154.6
	PIFPAF[49]		66.7				
	Simple Pose[52]	68.1					
Bottom-up methods	OpenPose[10]		61.8				

Table 4: Comparison of accuracy of 2D pose estimation methods in various datasets. Different methods use different data sets, choose the most commonly used COCO, MPII, LSP, CrowdPose datasets to use different evaluation indicators to measure each data set. #Params and GFLOPs are calculated for the pose estimation network. If it is not pointed out, there is no data.

In the standard heatmap generation, different keypoints are covered by Gaussian kernels with the same standard deviation, which means that different keypoints are supervised by the same constructed heatmap. Yu et al. [120] incorporated scale information into heatmap generation for keypoints and proposed a scale-aware heatmap generator (SAHG). To alleviate the problem of unbalanced keypoint detection at small scales, they adjusted the contribution values among the joints by a weight distribution loss function. Luo et al. [62] propose a scale adaptive heatmap regression (SAHR), which can learn to adjust the standard deviation for each keypoint by itself. The SAHR may aggravate the imbalance between fore-background samples, which potentially restricts the improvements of SAHR. Thus, they introduced weight-adaptive heatmap regression (WAHR) to reduce the loss of easy samples.

The bottom-up method shows higher robustness than the top-down method. The bottom-up method usually uses the method based on heatmap, and the change of character scale in the image will significantly affect the performance of pose estimation. In most cases, they can't predict the exact posture of smaller people.

To sum up, with the rapid development of deep learning, the performance of 2D HPE has been significantly improved. In recent years, stronger and more robust network models have enhanced the performance of 2D HPE methods, such as HRNet, DARK, SimplePose, OpenPose and so on. The HPE methods mentioned above are sorted out in Table 4, and the accuracy, the number of parameters and the complexity of the algorithm are compared respectively. We can see that the heatmap-based method has more accurate results than the regression-based method, but the computational complexity and the number of parameters is too large. In the multi-person pose estimation, the top-down method is better than the bottom-up method. As a whole, the single-person HPE method achieves better results than the multi-person method. This is due to the existence of mutual occlusion in the multi-person scene, although the top-down human body detector may not be able to recognize most of the overlapping methods, but for the bottom-up method, it is more difficult to obtain the position of the human body node. In addition, it is worth noting that the heatmap-based method has become a standard HPE method because of its excellent performance, but this is accompanied by an increase in the amount of calculation. It is difficult for these methods to run efficiently on mobile devices (mobile phones, tablet computer, etc.). Therefore, how to reduce the amount of calculation and improve the computational efficiency is one of the challenges of the existing HPE methods.

5.2 3D Pose Estimation

3D HPE aims at generating 3D spatial coordinates of human joints, which is more difficult than the task of 2D pose estimation. Similarly, previous methods can be categorized into single-person methods and multi-person methods. When performing 3D HPE from monocular RGB image, the lack of depth information for human keypoints often requires the incorporation of multiple-view images for pose estimation. By leveraging RGB image data from multiple viewpoints, a more accurate 3D human body model can be reconstructed, leading to more precise HPE. In addition to using multi-view images, other sensors can be utilized to obtain 3D human pose information. For example, depth cameras can capture depth data to provide more accurate 3D HPE results.

5.2.1 Single-person pose estimation

Single-person 3D HPE methods can generally be categorized into three types: direct 3D pose generation from images, 2D pose estimation followed by 3D pose mapping, and model-based approaches.

The Direct method is to train deep CNN end-to-end to estimate 3D human poses directly from the input images, as shown in Figure 14. Zhou et al. [133] used Part-Centric Heatmap Triplets (HEMIlets) as intermediate states. They first acquired HEMIlets from input image via ConvNet, and then regressed volumetric joint heatmap. The direct method benefits from the rich information contained in the image, such as the front and back direction of the limb. However, it will also be affected by many factors, such as background, lighting, clothing and so on. Networks trained on one dataset cannot be well extended to other datasets with different environments[51].

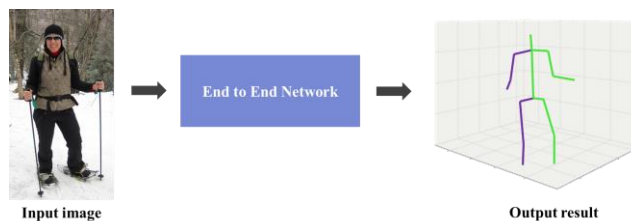


Figure 14: Pipeline of Direct methods.

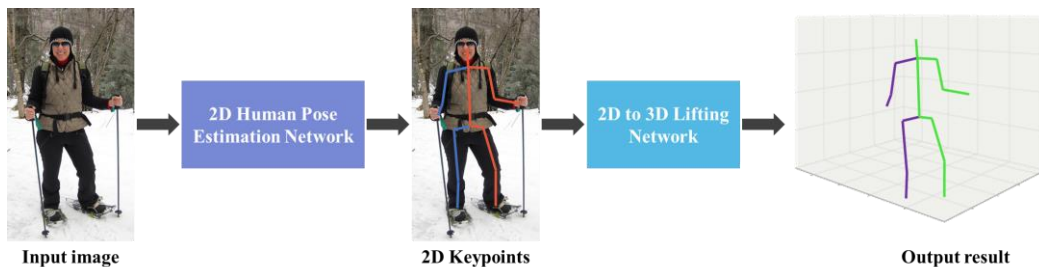


Figure 15: Pipeline of Lifting-based methods [131].

The Lifting-based method first generates 2D keypoints from input image, and then lift 2D pose to 3D one, as shown in Figure 15. Estimating 3D pose from 2D representation is essentially an ill-posed problem, because there may be multiple valid 3D poses for a single 2D pose, which makes it difficult to infer the only valid solution, especially in the case of severe occlusion [51, 97, 105]. Li and Lee [51] introduced the mixture density networks (MDN) into the process of 3D HPE. They suggested using the negative log-likelihood of minimizing Gaussian multimodal mixing to estimate multiple assumptions from 2D to 3D poses. Therefore, they used the mixing coefficients and variances of Gaussian kernels output by MDN to represent the uncertainty of each 3D pose. Cai et al. [8] exploited

the temporal and spatial relationships of 3D HPE through GCN. They designed non-uniform graph convolution to apply different weights for different neighborhoods of a joint. GCN can achieve good results in 3D HPE because graph convolution has good feature extraction ability for graph structure data[51]. However, the features extracted by the existing architecture are too simple to use multi-scale and multi-level features, which limits the expression ability of the model. Xu and Takano [115] proposed a Graph Stacked Hourglass Networks to extract multi-scale features to enable high-accuracy 3D HPE.

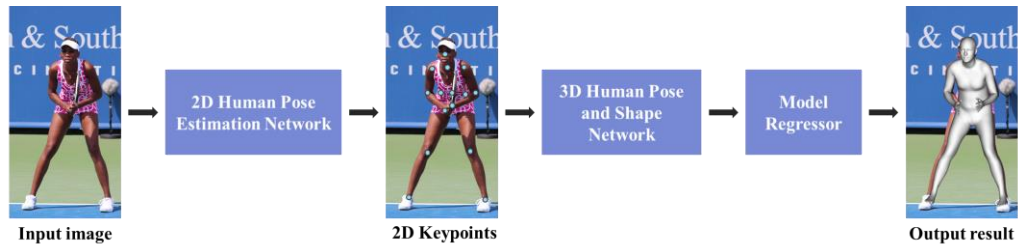


Figure 16: Pipeline of the Model-based method.

Model-based methods focus on estimating the parameters of regression grid models such as SMPL, as shown in Figure 16. Madadi et al. [65] regressed SMPL parameters by using 3D articulated points and sparse human surface marker points, thus avoiding the artefacts produced by direct regression. Kolotouros et al. [48] regressed the parameters of SMPL by deep learning and used them to initialize the iterative fitting routine. Afterward, these parameters are used for supervision to reduce the regression and optimization cycles. Xiang et al. [109] proposed 3D Part Orientation Fields (POFs) to encode 3D orientation of body parts. Rong et al. [85] performed this task by running the face, hand, and 3D pose regression modules independently. Then, they combine the three outputs through an integrated module. The depth regression of PyMAF proposed by Zhang et al. [123] used multi-scale spatial features to achieve high-level and detailed results. Model-based methods are ill-conditioned and suboptimal due to loss of depth information or ambiguity of joint orientation and shape parameters. However, 3D joints and surface markers can better deal with this problem[65].

The direct use of human models for structural representation is often affected by noise and artifacts. Taking advantage of prior knowledge can improve accuracy and efficiency to some extent. By constraining the human model with kinematic knowledge, many works have obtained better performance for 3D HPE. Xu et al. [112] suggest incorporating kinematic regularisation into the deep model. They not only optimize 3D key points but also perform kinematic structure correction for noisy 2D inputs. Skeletor[35] uses temporal continuity and body prior to reducing the noises in 3D HPE. Weng and Yeung [107] combine a set of physical constraints and prior knowledge in a joint optimization process. MotioNet[89] uses a Deep Neural Network (DNN) to decompose 2D joint position sequences into skeletal length-dependent and 3D joint rotation sequences, which are then fed into forwarding kinematic layers to output 3D positions. Santesteban et al. [88] added soft tissue dynamics to the parametric model. In the real world, human motion involves spatial and structural kinematic principles: the human skeletal structure exists in three-dimensional space and generates keypoints by projecting onto a two-dimensional plane. The positions of these keypoints should comply with the geometric constraints of perspective projection. The distance between two adjacent joints remains constant throughout the entire motion sequence. The estimated 3D trajectory formed by the same joint should be smooth and continuous, implying that there should be reasonable transitions between the positions of adjacent joints at different time points, rather than abrupt jumps or discontinuities.

5.2.2 Multi-person pose estimation

We classify multi-person 3D pose estimation into top-down methods (shown in Figure 17) and bottom-up methods (shown in Figure 18). Top-down methods first detect edges, and then estimate

each person's 3D pose. In contrast, bottom-up methods first generate body joint positions and depth maps, and then correlate the individual joints based on the depth maps.

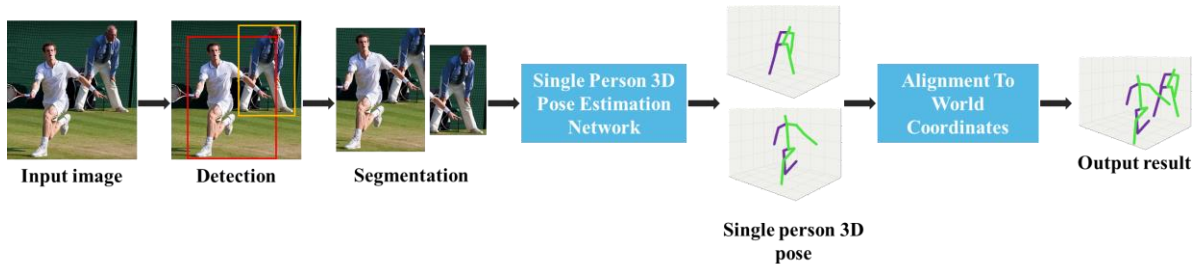


Figure 17: Pipeline of the 3D multi-person top-down method [131].

Similar to the 2D domain, the top-down method uses the person detector to crop the area containing one person, and then executes the 3D single person pose estimator. LCR-Net[83] is the first method that estimates multi-player 3D human pose from a single image, where pose suggestions are obtained by locating K anchor poses on the region of interest. PandaNet[4] utilized pose-aware anchor selection to reduce ambiguous anchors. Jiang et al. [36] added a new branch on the R-CNN pipeline to generate SMPL parameters. In order to punish the intersection between SMPL models, they introduced internetworking loss to punish the intersection of the reconstructed population, and introduced another Loss function to punish the difference between the reconstructed human depth order and the actual depth order.

The bottom-up method is more difficult than the former, but the results are better. Li et al. [57] proposed a two-branch framework that uses both top-down and bottom-up networks. The final 3D pose is obtained by feeding the output of the two networks into an integrated network.

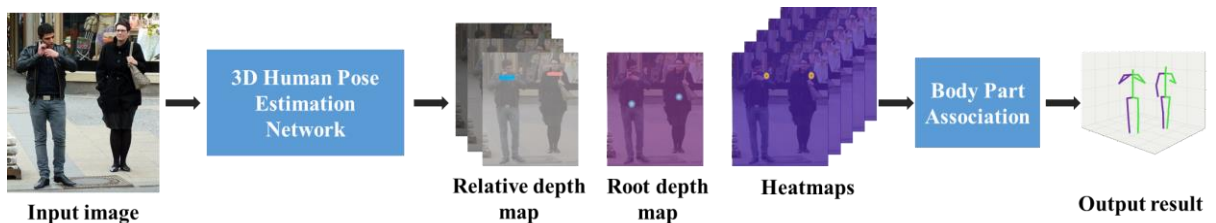


Figure 18: Pipeline of the 3D multi-person bottom-up method [131].

They also proposed a two-person pose discriminator to handle natural interactions between multiple people. Zhang et al. [126] proposed VoxelPose, which is a part of 2D HPE information provided by different cameras through multi-branch networks, which is finally integrated into 3D HPE. Mehta et al. [70] used spatio-temporal skeleton model to predict 2D and 3D poses and generated coherent skeleton joint angles in real-time. In the first stage, they only deal with the analysis of visible joints and their parent joints, which reduces computational overhead. In the second stage, they use whole-body context and other intermediate information to reason and obscure joints. In the first stage, they designed a new architecture-SelecsLSNet. This allows the entire method to work in real time at a speed of 25-30fps on a single consumer GPU.

5.2.3 Pose estimation from multiple views

The location information from a monocular image is known to be limited. Recent works intend to obtain the positional relationship of the same target from multiple cameras.

Existing multi-view HPE methods are typically divided into three stages. The first stage is to detect human keypoints in independent 2D views. The second stage involves grouping the

corresponding poses of the same person in different views based on appearance and geometric cues. The third stage is to estimate the 3D poses of each individual using standard methods such as triangulation or PSM. In multi-view settings, the primary challenge is to find the corresponding relationships of keypoints across different views and associate them with the corresponding individuals. Qiu et al. [79] proposed the use of RPSM to recover 3D poses from multi-view 2D poses. Compared to PSM, this method achieved a 50% reduction in error. Dong et al. [20] enhanced the cross-view consistency with appearance features. They utilized a person's Re-ID network to extract appearance features for each individual. These features were then used to calculate appearance-based distances. They also formulated a convex optimization problem to solve for the optimal correspondence matrix, and employed rank constraints to enforce cycle consistency. Remelli et al. [81] also adopted the idea of explicit fusion. They transformed the latent features with known camera extrinsic characteristics into a canonical 3D space and then stacked the transformed features from multiple views together for joint inference. The methods mentioned earlier explicitly establish correspondences between joints in the same or different views, there are also several methods that indirectly handle joint correspondences through other strategies. AdaFuse[130] utilizes a pose estimation network to generate 2D heatmaps for each view. It establishes point-to-line correspondences between two views using epipolar geometry. The sparse representation of the heatmaps is then explored to implicitly determine point-to-point correspondences. Background pixels in individual views are influenced by epipolar lines in other views. To address this, AdaFus applies the SoftMax operator to eliminate responses and obtain the most reliable joint locations in the ground truth joint positions. VoxelPose[95] projects 2D pose heatmaps from multiple views into a 3D common space and performs 3D pose estimation on candidate voxel representations for each individual in the 3D space. By doing so, VoxelPose implicitly establishes associations between joints within a single view and across different views, effectively reducing the impact of erroneously established cross-view correspondences. However, it should be noted that this method involves performing 3D convolutions on all voxel positions, which can be computationally intensive.

Method		Human3.6M		MPI-INF-3DHP		3DPW		Campus	Shelf
		Protocol #1	Protocol #2	PCK	AUC	MPJPE	PA-MPJPE	PCP	PCP
Single person pose estimation	Direct prediction	HEMIlets[133]	39.9	32.1	75.3	38			
		Wehrbein et al. [105]	44.3	32.4	84.3				
	Lifting methods	GraphSH[115]	35.8		80.1	45.8			
		RepNet[97]	89.9	65.1	82.5	58.5			
		Cai et al. [8]	48.8	39.0					
	Li and Lee [51]	52.7	42.6	67.9					
		SMPLR[65]	62.6	52.2					
	Model-based methods	SPIN[48]	62.5				96.9	59.2	
		Frankmocap[85]					94.3	60	
		HMR[42]	87.9	56.8	72.93		130	76.7	
		POFs[109]	58.3						
		PyMAF[123]					92.8	58.9	
		HKMR[26]	59.6	67.74					
		Pose2Mesh	64.9				88.9	56.3	
Multi-person pose estimation	Top-down methods	LCR-Net[83]	53.5	61.2	70.6				
		PandaNet[4]			72				
		Jiang et al. [36]		52.7	69.12				
	Bottom-up methods	VoxelTrack[126]						96.7	97.1
		Xnet[70]	63.6		82.8	45.3	103.3	60.7	
Human pose estimation from multiple views	Dong et al. [20]							96.3	96.9
	AdaFuse[130]	19.5							
	VoxelPose[95]							96.7	97.0
	Remelli et al. [81]	34.2							

Table 5: Accuracy Comparison of the 3D pose estimation methods. The methods in this table are applicable only to RGB images.

5.2.4 Pose estimation from other sources

Although the ordinary RGB camera is the most common device for 3D HPE, depth cameras and other sensors can also achieve the goal of 3D HPE. In this section, we provide a brief introduction to these alternative devices.

As discussed earlier, the depth images generated by depth cameras can provide depth information of key points, which alleviates the depth blur problem of key points in 3D HPE. Wang et al. [104] designed a dual-input network that takes RGB images and depth information as inputs, and generate feature streams through feature map generator. Bashirov et al. [3] converted the depth information to the SMPL pose parameters using the SMPL-X model. Since depth information is not effective in representing smaller body parts, they used RGB images to obtain the pose information.

Recently, an important trend in HPE is fusing image and IMU data to achieve more realistic pose estimation [27, 32, 41, 129, 134]. The underlying logic of this approach is to build a parametric 3D human body model and optimize its parameters to minimize the discrepancies between the model and both image and IMU data.

In summary, the results of the 3D HPE methods mentioned above are summarized in Table 5. According to Table 5, we can see that the accuracy of lifting methods is higher than that of model-based methods, which is because lifting methods adopts the strategy of upgrading from 2D to 3D. The progress of 2D HPE methods greatly improves the performance of 3D HPE. However, the key point location information obtained from a single RGB image is limited, and obtaining the key point from a single image is an ill-posed problem. The multi-view 3D HPE method achieves the best results because multiple views can alleviate the occlusion problem and provide depth information of the node, but the integration of multiple views increases the calculation cost. Although the existing 3D HPE method has achieved good results on Human3.6M, the error rate on outdoor data set 3DPW has increased significantly. This is because the outdoor environment is more complex, and factors such as light, object occlusion, crowd occlusion and clothing seriously affect the algorithm performance. Therefore, model generalization is a challenge faced by 3D HPE method. In addition, the use of physical knowledge and kinematic knowledge to constrain the human body model can help the HPE method to reason out the blocked body parts or reduce many unreasonable positions of the human body model.

5.3 Pose Estimation from Video

The methods of HPE from video aim at establishing correlations of all joints in time and space. Wang et al. [102] created a spatial-temporal tube and used a Clip Tracking Network to clip all video frames. Yang et al. [117] used the trajectory of the acquired pose from previous frames as input via a GNN. They produced the final pose by aggregating the predicted and detected poses in the same frame. ChallenCap[62] acquired an initial noisy skeletal motion map from monocular video. Then, they used a motion optimizer with 3D prior knowledge and 2D contour information to optimize the skeletal motion. To address the temporal consistency problem and reduce the feature dependence of human motion, Choi et al. [14] proposed a temporally consistent mesh recovery system (TCMR) to remove the residual link between static and temporal features. The SMPL parameters were predicted by using temporal features in the PoseForecast. However, the trajectory of each joint is different throughout the spacetime and needs to be learned separately.

Owing to their high efficiency, attention mechanisms have been widely used in computer vision tasks such as object detection, attentive mechanisms for pose estimation from videos have attracted increasing research interest. Attention mechanisms can be divided into spatial, channel, layer, and mixed attentions. Kocabas et al. [46] proposed VIBE to distinguish real human motion and motion generated by temporal pose. However, adversarial training at the sequence level could produce kinematically plausible motion sequences without actual 3D labels. Wang et al. [101] proposed spatial-temporal mesh attention convolution (MAC) and used weakly supervised learning for HPE. Additionally, Li et al. [55] proposed a Transformer-based method, named MHFORER, to learn the Spatio-temporal representation of multiple pose hypotheses. In PoseFormer [132], the authors

designed a temporal structure to capture the correlation between human joints in time. The attention maps computed by the self-attentive mechanism module might not be stable. To solve this problem, Wei et al. [106] proposed a motion continuity attention (MoCA) module that can better adapt to different temporal action contents. Zhu et al. [134]'s method achieved the best performance, mainly by using the Dual-stream Spatiotemporal Transform (DSTformer) as the motion encoder in the process of 2D to 3D pose estimation. The motion encoder not only learns the 2D pose with added noise, but also utilizes the long-range relationship among skeleton keypoints. In Table 6, we have summarized the methods discussed in this section, and it is observed that the methods utilizing Transformer achieve higher accuracy, particularly on the 3DPW dataset, compared to the methods mentioned earlier. This improvement can be attributed to the attention mechanism employed in Transformer.

Method	Human3.6M		MPI-INF-3DHP		3DPW	
	MPJPE		PCK	AUC	MPJPE	PA-MPJPE
Human Pose Estimation in Video	TCMR[14]	62.3			95.0	55.8
	VIBE[46]	65.9			93.5	56.5
	MHFormer[55]	43.0	150			
	PoseFormer[132]	44.3	88.6	56.4		
	MixSTE[124]	39.8				
	MotionBERT[134]	37.5			68.8	40.6
	MPS-Net[106]	47.4			84.3	52.1

Table 6: Accuracy of human posture estimation from video.

5.4 Lightweight Methods

In recent years, with the popularity of mobile devices and the advancement of virtual reality (VR) and augmented reality (AR) technologies, mobile VR and AR have become hot topics of interest in both academia and industry. Many state-of-the-art methods for HPE take advantage of high-resolution networks, which make the networks more powerful. However, the number of required parameters and computational complexity multiplies, making it difficult for deep network models to be applied to mobile devices. Recently, many research efforts have been put into developing lightweight networks to improve efficiency and accuracy. The lightweight network design method is divided into two categories: lightweight network structure design and model compression. Based on MobileNetV2 [87], Choi et al. [15] proposed MobileHumanPose, a single-stage model for mobile devices, to solve the fine-tuning problem. Wang et al. [103] used deconvhead to remove the redundant refinement of the high-resolution branch on MobileNetV2, and then performed scale-aware multi-resolution fusion through a single branch, and used the neural architecture to search and optimize the model and use a large convolution kernel. Yu et al. [119] proposed Lite-HRNet, which combines the shuffle module in ShuffleNet [63] and the high-resolution design pattern in HRNet together, achieve feature fusion of channels and spaces through attention mechanisms. The best AP of Lite-HRNet in COCO dataset is 70.4%, and the number of parameters is only 1.8M. Compared with HRNet-W48, although AP is reduced by 5.1%, the number of parameters is reduced by 61.8 M. Luo et al. [61] proposed FastNet by using a modified asymmetric bottleneck module as the basic component. In addition, many researchers[12, 113] have proposed methods based on Neural Architecture Search (NAS). Among them, Xu et al. [113] proposed ViPNAS which allows joint optimization of spatial and temporal layer search through a single pass framework. The method based on neural architecture search achieves better results of other lightweight methods, but their own number of parameters is several times that of other methods.

In Table 7, we summarized the results of lightweight HPE methods mentioned earlier. By comparing with Table 3, the lightweight model architectures have reduced the parameter count by over 90%, resulting in a significant decrease in computational complexity. For example, Lite-HRNet has a GFLOPs of 0.70, while HRNet-W48[93] has a GFLOPs of 32.9, reducing the computational complexity by 97.87%. On average, the accuracy of lightweight models decreased by around 5%,

with the NAS-based methods showing the best performance and having the highest number of parameters compared to other lightweight methods.

Methods	COCO		Human3.6M		#Params	GFLOPs
	AR	AP	Protocol #1	Protocol #2		
Lite-HRNet[119]	76.2	70.4			1.8M	0.70
FastNet[61]	77.1	71.4			9.0M	3.2
MobileHumanPose [15]			35.2	51.4	4.07M	5.49
LitePose[103]		40.6			1.7M	
ViPNAS[113]	81.2	74.7			16.3M	1.44

Table 7: Accuracy comparisons of lightweight network architectures. #Params and GFLOPs are calculated for the pose estimation network.

6 DISCUSSIONS

In the above sections, we have introduced deep learning-based methods, human representation models and capture devices for HPE. In the following, we further discuss the performances of these methods from three perspectives: accuracy, speed, and robustness.

Accuracy. The accuracy of 3D HPE methods has been improved by 70 mm over the last decade with the improvements of 2D HPE. Usually, the multi-view methods produce more accurate results than the monocular image ones. At present, the most advanced 3D HPE method[134] using monocular images has realized 16.9mm average MPJPE on Human3.6M. In their method, geometric prior knowledge is used for multi-hypothesis aggregation. The multi-view HPE method reaches 15.6mm average MPJPE[16]. The multi-view method can provide more information about the position of the keypoints, which is helpful to solve the position ambiguity of the keypoints of 3D HPE in a single image. Although the information from multi-views reduces self-occlusions and extra-occlusions, the usage of multiple cameras is much more expensive.

As heatmap-based methods usually take high-resolution images as inputs, the prediction accuracy will be greatly reduced if low-resolution images are used as inputs. How to balance accuracy and network computational complexity has become one of the concerns of heatmap-based methods. By using different Gaussian kernels, the accurate location of each key point is realized. For example, Yu et al. [120]'s method and Luo et al. [62]'s method continuously learn and adjust the standard deviation of the Gaussian kernel for each keypoint and achieve 69.4% AP and 72.0% AP, respectively, on the COCO test-dev dataset.

Vision Transformer shows excellent strength in HPE, and the reason comes from its own attention mechanism. Attention mechanism can pay more attention to global, local and contextual information, which not only improves the performance of the model, but also reduces the amount of calculation. The state-of-the-art model MotionBERT(Finetune) [134] has achieved 16.9mm average MPJPE in Human3.6M. Compared with the monocular image method, the MPJPE is reduced by 2.7mm on average, and is better than the multi-view method. Table 5 lists several of the best results achieved using Transformer's pose estimation method.

Robustness. There are certain challenges in extending 3D HPE to outdoor environments. This is because most of the existing large-scale motion capture datasets are collected indoors, such as Human3.6M and Total Capture. As a result, the performance on outdoor datasets is poorer, for example, MotionBERT [134] achieves the best performance on the Human3.6M dataset but only achieves 76.9mm MPJPE on the 3DPW dataset. To overcome this discrepancy, it is necessary to collect more outdoor data and use data augmentation techniques to improve the model's generalization ability.

To improve the performance of the model in outdoor environments, the following methods can be considered: firstly, utilize existing outdoor datasets such as AGORA and DensePose to acquire diverse outdoor scenes and a variety of human pose data. This will help improve the model's generalization ability. Secondly, by using data augmentation techniques [28], introduce training data with different backgrounds and human poses to increase the model's ability to recognize occlusions.

Occlusion is an important challenge in monocular human detection. To address this issue, image context understanding and prior knowledge of body structure can be utilized to infer the positions of occluded body parts. For example, by leveraging joint constraints and the assumption of body part coherence for pose estimation, the accuracy of estimating occluded regions can be improved.

Additionally, multi-view methods are commonly used solutions in handling occlusion problems. By using multiple cameras or viewpoints to observe the target from different angles and jointly analyzing images from multiple views, object occlusion can be reduced, thereby improving the accuracy and robustness of HPE.

Speed. Previous methods have used complex networks to achieve high precision, but the predictions require a large amount of computation. Some recent lightweight networks allow algorithms to run on low-performance devices, which enables affordable markerless motion capture by mobile phone cameras. Deployed on low-end devices (iPhone7 with CPU), [33] can run faster than 20 fps, while high-end devices (iPhoneXS with NPU) have throughput above 160 fps. It is difficult for other large network architectures to perform faster than 10 fps on CPU and GPU. The Xnect[70] runs at 27 fps on laptops equipped with Intel i7-8780H and 1080-MaxQ. The existing low-end devices basically meet the computing requirements of the lightweight method and can meet some entertainment needs.

7 CONCLUSIONS AND FUTURE DEVELOPMENTS

In this paper, we review the recent developments in human pose estimation methods for markerless motion capture. According to the number of people in the input image, we provide single-person and multi-person pose estimation methods in different dimensions. In addition, we also focus on video-based and lightweight methods. We also compare and discuss different models and methods in terms of robustness, accuracy, and speed, analysing their advantages and disadvantages. Furthermore, we review current datasets and evaluation metrics from various dimensions and compile commonly used capture devices in human pose estimation. Thus, this review serves as a guide for researchers interested in this field and as a reference for studying existing models and developing new methods. So far, there are still several problems that need to be solved in the future. In the following, we propose possible directions for future works.

A unified overall model deserves to be explored in the future. Existing models of faces, hands, and bodies are isolated from each other and therefore unable to take full advantage of the large-scale data analysis and model-building processes that are emerging in the context of deep learning. Full-body models help capture and deeply analyse subtle interactions in virtual reality as well as physical 3D Spaces.

Attention mechanism plays an important role in 3D pose estimation. Transformer gets attention because of its excellent performance in posture regression, and researchers apply attention mechanism to CNN. However, the Transformer method has a large amount of computation and needs a simple and efficient model structure.

Most approaches ignore the close interaction between humans and 3D scenes. Reducing body intersection and occlusions by adding knowledge of physical movements will be an interesting direction in the future.

Faster and lighter network architectures need to be explored to reduce the cost of computation. At the same time, the accuracy of pose estimation should be guaranteed. Thus, many applications will be benefitted.

ACKNOWLEDGEMENTS

This research is funded by the National Natural Science Foundation of China (grant number 61772293), the Shandong Province Key Research and Development Plan (grant number 2020CXGC011004).

Li Wang, <https://orcid.org/0000-0002-4773-3936>

Yu-Wei Zhang, <http://orcid.org/0000-0001-6566-5714>

REFERENCES

- [1] Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; Schiele, B.: PoseTrack: A benchmark for human pose estimation and tracking, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 4654-4663. <https://doi.org/10.1109/CVPR.2018.00542>
- [2] Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, 3686-3693. <https://doi.org/10.1109/CVPR.2014.471>
- [3] Bashirov, R.; Ianina, A.; Isakov, K.; Kononenko, Y.; Strizhkova, V.; Lempitsky, V.; Vakhitov, A.: Real-time RGBD-based Extended Body Pose Estimation, 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, 2806-2815. <https://doi.org/10.1109/WACV48630.2021.00285>
- [4] Benzine, A.; Chabot, F.; Luvison, B.; Pham, Q. C.; Achard, C.: Pandanet: Anchor-based single-shot multi-person 3d pose estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 6855-6864. <https://doi.org/10.1109/CVPR42600.2020.00689>
- [5] Bortolini, M.; Faccio, M.; Gamberi, M.; Pilati, F.: Motion Analysis System (MAS) for production and ergonomics assessment in the manufacturing processes, Computers & Industrial Engineering, 139, 2020, 105485. <https://doi.org/10.1016/j.cie.2018.10.046>
- [6] Bortolini, M.; Gamberi, M.; Pilati, F.; Regattieri, A.: Automatic assessment of the ergonomic risk for manual manufacturing and assembly activities through optical motion capture technology, Procedia CIRP, 72, 2018, 81-86. <https://doi.org/10.1016/j.procir.2018.03.198>
- [7] Bridgeman, L.; Volino, M.; Guillemaut, J.-Y.; Hilton, A.: Multi-person 3d pose estimation and tracking in sports, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, 2487-2496. <https://doi.org/10.1109/CVPRW.2019.00304>
- [8] Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.-J.; Yuan, J.; Thalmann, N. M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 2272-2281. <https://doi.org/10.1109/ICCV.2019.00236>
- [9] Calabrese, E.; Taverni, G.; Awai Easthope, C.; Skriabine, S.; Corradi, F.; Longinotti, L.; Eng, K.; Delbruck, T.: Dhp19: Dynamic vision sensor 3d human pose dataset, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, 1695-1704. <https://doi.org/10.1109/CVPRW.2019.00217>
- [10] Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S. E.; Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 2021, 172-186. <https://doi.org/10.1109/TPAMI.2019.2929257>
- [11] Chen, X.; Pang, A.; Yang, W.; Ma, Y.; Xu, L.; Yu, J.: SportsCap: Monocular 3D human motion capture and fine-grained understanding in challenging sports videos, 129(10), 2021, 2846-2864. <https://doi.org/10.1007/s11263-021-01486-4>
- [12] Chen, Z.; Huang, Y.; Yu, H.; Xue, B.; Han, K.; Guo, Y.; Wang, L.: Towards part-aware monocular 3d human pose estimation: An architecture search approach, Computer Vision—ECCV 2020, 2020, 715-732. https://doi.org/10.1007/978-3-030-58580-8_42

- [13] Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T. S.; Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 5385-5394. <https://doi.org/10.1109/CVPR42600.2020.00543>
- [14] Choi, H.; Moon, G.; Chang, J. Y.; Lee, K. M.: Beyond static features for temporally consistent 3d human pose and shape from a video, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 1964-1973. <https://doi.org/10.1109/CVPR46437.2021.00200>
- [15] Choi, S.; Choi, S.; Kim, C.: MobileHumanPose: Toward real-time 3D human pose estimation in mobile devices, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, 2328-2338. <https://doi.org/10.1109/CVPRW53098.2021.00265>
- [16] Chun, S.; Park, S.; Chang, J. Y.: Representation learning of vertex heatmaps for 3D human mesh reconstruction from multi-view images, arXiv, abs/2306.16615, 2023, <https://doi.org/10.48550/arXiv.2306.16615>
- [17] Clark, R. A.; Mentiplay, B. F.; Hough, E.; Pua, Y. H. J. G.; posture: Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives, Gait & Posture, 68, 2019, 193-200. <https://doi.org/10.1016/j.gaitpost.2018.11.029>
- [18] Colyer, S. L.; Evans, M.; Cosker, D. P.; Salo, A. I.: A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System, Sports Med Open, 4(1), 2018, 24. <https://doi.org/10.1186/s40798-018-0139-y>
- [19] Cootes, T. F.; Taylor, C. J.; Cooper, D. H.; Graham, J.: Active shape models-their training and application, Computer Vision and Image Understanding, 61(1), 1995, 38-59. <https://doi.org/10.1006/cviu.1995.1004>
- [20] Dong, J.; Jiang, W.; Huang, Q.; Bao, H.; Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 7784-7793. <https://doi.org/10.1109/CVPR.2019.00798>
- [21] Felzenszwalb, P. F.; Huttenlocher, D. P.: Efficient belief propagation for early vision, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004 CVPR 2004, 70(1), 2006, 41-54. <https://doi.org/10.1109/CVPR.2004.1315041>
- [22] Ferrari, V.; Marin-Jimenez, M.; Zisserman, A.: Progressive search space reduction for human pose estimation, 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, 1-8. <https://doi.org/10.1109/CVPR.2008.4587468>
- [23] Fürst, M.; Gupta, S. T.; Schuster, R.; Wasenmüller, O.; Stricker, D.: HPERL: 3d human pose estimation from RGB and lidar, 2020 25th International Conference on Pattern Recognition (ICPR), 2021, 7321-7327. <https://doi.org/10.1109/ICPR48806.2021.9412785>
- [24] Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.: Event-based vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1), 2022, 154-180. <https://doi.org/10.1109/TPAMI.2020.3008413>
- [25] Garau, N.; Martinelli, G.; Bródka, P.; Bisagno, N.; Conci, N.: PanopTOP: a framework for generating viewpoint-invariant human pose estimation datasets, 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, 234-242. <https://doi.org/10.1109/ICCVW54120.2021.00031>
- [26] Georgakis, G.; Li, R.; Karanam, S.; Chen, T.; Koščeká, J.; Wu, Z.: Hierarchical kinematic human mesh recovery, European Conference on Computer Vision, 2020, 768-784. https://doi.org/10.1007/978-3-030-58520-4_45
- [27] Gilbert, A.; Trumble, M.; Malleson, C.; Hilton, A.; Collomosse, J.: Fusing Visual and Inertial Sensors with Semantics for 3D Human Pose Estimation, International Journal of Computer Vision, 127(4), 2019, 381-397. <https://doi.org/10.1007/s11263-018-1118-y>

- [28] Gong, K.; Zhang, J.; Feng, J.: PoseAug: A Differentiable Pose Augmentation Framework for 3D Human Pose Estimation, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 8571-8580. <https://doi.org/10.1109/CVPR46437.2021.00847>
- [29] Gu, C.; Lin, W.; He, X.; Zhang, L.; Zhang, M.: IMU-based Mocap system for rehabilitation applications: A systematic review, *Biomimetic Intelligence and Robotics*, 3(2), 2023, 100097. <https://doi.org/10.1016/j.birob.2023.100097>
- [30] Güler, R. A.; Neverova, N.; Kokkinos, I.: Densepose: Dense human pose estimation in the wild, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 7297-7306. <https://doi.org/10.1109/CVPR.2018.00762>
- [31] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.: Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV), 2017 of Conference, 2980-2988. <https://doi.org/10.1109/ICCV.2017.322>
- [32] Huang, F.; Zeng, A.; Liu, M.; Lai, Q.; Xu, Q.: Deepfuse: An imu-aware network for real-time 3d human pose estimation from multi-view image, 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, 418-427. <https://doi.org/10.1109/WACV45572.2020.9093526>
- [33] Hwang, D.-H.; Kim, S.; Monet, N.; Koike, H.; Bae, S.: Lightweight 3d human pose estimation network training using teacher-student learning, 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, 468-477. <https://doi.org/10.1109/WACV45572.2020.9093595>
- [34] Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C.: Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 2014, 1325-1339. <https://doi.org/10.1109/TPAMI.2013.248>
- [35] Jiang, T.; Camgoz, N. C.; Bowden, R.: Skeletor: Skeletal transformers for robust body-pose estimation, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, 3389-3397. <https://doi.org/10.1109/CVPRW53098.2021.00378>
- [36] Jiang, W.; Kolotouros, N.; Pavlakos, G.; Zhou, X.; Daniilidis, K.: Coherent reconstruction of multiple humans from a single image, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 5579-5588. <https://doi.org/10.1109/CVPR42600.2020.00562>
- [37] Jiang, Y.; Song, K.; Wang, J.: Action recognition based on fusion skeleton of two kinect sensors, 2020 International Conference on Culture-oriented Science & Technology (ICCST), 2020, 240-244. <https://doi.org/10.1109/ICCST50977.2020.00052>
- [38] Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P.: Whole-body human pose estimation in the wild, *European Conference on Computer Vision*, 2020, 196-214. https://doi.org/10.1007/978-3-030-58545-7_12
- [39] Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture, 2015 IEEE International Conference on Computer Vision (ICCV), 2015, 3334-3342. <https://doi.org/10.1109/ICCV.2015.381>
- [40] Ju, S. X.; Black, M. J.; Yacoob, Y.: Cardboard people: A parameterized model of articulated image motion, *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 1996, 38-44. <https://doi.org/10.1109/AFGR.1996.557241>
- [41] Jung, S.; Park, S.; Lee, K.: Pose tracking of moving sensor using monocular camera and IMU sensor, *KSII Transactions on Internet and Information Systems*, 15(8), 2021, 3011-3024. <http://doi.org/10.3837/tiis.2021.08.017>
- [42] Kanazawa, A.; Black, M. J.; Jacobs, D. W.; Malik, J.: End-to-end recovery of human shape and pose, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 7122-7131. <https://doi.org/10.1109/CVPR.2018.00744>

- [43] Kanko, R. M.; Laende, E.; Selbie, W. S.; Deluzio, K. J.: Inter-session repeatability of markerless motion capture gait kinematics, *Journal of Biomechanics*, 121, 2021, 110422. <https://doi.org/10.1016/j.jbiomech.2021.110422>
- [44] Kanko, R. M.; Laende, E. K.; Davis, E. M.; Selbie, W. S.; Deluzio, K. J.: Concurrent assessment of gait kinematics using marker-based and markerless motion capture, *Journal of Biomechanics*, 127, 2021, 110665. <https://doi.org/10.1016/j.jbiomech.2021.110665>
- [45] Kim, W.; Sung, J.; Saakes, D.; Huang, C.; Xiong, S.: Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose), *International Journal of Industrial Ergonomics*, 84, 2021, 103164. <https://doi.org/10.1016/j.ergon.2021.103164>
- [46] Kocabas, M.; Athanasiou, N.; Black, M. J.: Vibe: Video inference for human body pose and shape estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 5253-5263. <https://doi.org/10.1109/CVPR42600.2020.00530>
- [47] Kocabas, M.; Karagoz, S.; Akbas, E.: Self-supervised learning of 3d human pose using multi-view geometry, 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 2019, 1077-1086. <https://doi.org/10.1109/FG52635.2021.9667074>
- [48] Kolotouros, N.; Pavlakos, G.; Black, M. J.; Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 2252-2261. <https://doi.org/10.1109/ICCV.2019.00234>
- [49] Kreiss, S.; Bertoni, L.; Alahi, A.: Pifpaf: Composite fields for human pose estimation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 11969-11978. <https://doi.org/10.1109/CVPR.2019.01225>
- [50] Leroy, V.; Weinzaepfel, P.; Brégier, R.; Combaluzier, H.; Rogez, G.: SPLY benchmarking 3d human pose estimation in the wild, 2020 International Conference on 3D Vision (3DV), 2020, 301-310. <https://doi.org/10.1109/3DV50981.2020.00040>
- [51] Li, C.; Lee, G. H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 9879-9887. <https://doi.org/10.1109/CVPR.2019.01012>
- [52] Li, J.; Su, W.; Wang, Z.: Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation, *Proceedings of the AAAI conference on artificial intelligence*, 34(07), 2020, 11354-11361. <https://doi.org/10.1609/aaai.v34i07.6797>
- [53] Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 10855-10864. <https://doi.org/10.1109/CVPR.2019.01112>
- [54] Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; Tu, Z.: Pose recognition with cascade transformers, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 1944-1953. <https://doi.org/10.1109/CVPR46437.2021.00198>
- [55] Li, W.; Liu, H.; Tang, H.; Wang, P.; Van Gool, L.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 13137-13146. <https://doi.org/10.1109/CVPR52688.2022.01280>
- [56] Li, W.; Wang, Z.; Yin, B.; Peng, Q.; Du, Y.; Xiao, T.; Yu, G.; Lu, H.; Wei, Y.; Sun, J.: Rethinking on multi-stage networks for human pose estimation, 2019, <https://doi.org/10.48550/arXiv.1901.00148>
- [57] Li, Y.; Yang, S.; Zhang, S.; Wang, Z.; Yang, W.; Xia, S.-T.; Zhou, E.: Is 2D Heatmap Representation Even Necessary for Human Pose Estimation?, 2021, <https://doi.org/10.48550/arXiv.2107.03332>
- [58] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. L.: Microsoft coco: Common objects in context, *European conference on computer vision*, 2014, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48
- [59] Lin, W.; Liu, H.; Liu, S.; Li, Y.; Qian, R.; Wang, T.; Xu, N.; Xiong, H.; Qi, G.-J.; Sebe, N.: Human in events: A large-scale benchmark for human-centric video analysis in complex events, 2020, <https://doi.org/10.48550/arXiv.2005.04490>

- [60] Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M. J.: SMPL: A Skinned Multi-Person Linear Model, *Acm T Graphic*, 34(6), 2015, 1-16. <http://doi.acm.org/10.1145/2816795.2818013>
- [61] Luo, Y.; Ou, Z.; Wan, T.; Guo, J.-M.: FastNet: Fast high-resolution network for human pose estimation, *Image and Vision Computing*, 119, 2022, 104390. <https://doi.org/10.1016/j.imavis.2022.104390>
- [62] Luo, Z.; Wang, Z.; Huang, Y.; Wang, L.; Tan, T.; Zhou, E.: Rethinking the heatmap regression for bottom-up human pose estimation, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 13259-13268. <https://doi.org/10.1109/CVPR46437.2021.01306>
- [63] Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design, *European conference on computer vision (ECCV)*, 2018, 122-138. https://doi.org/10.1007/978-3-030-01264-9_8
- [64] Ma, X.; Su, J.; Wang, C.; Ci, H.; Wang, Y.: Context modeling in 3d human pose estimation: A unified perspective, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 6238-6247. <https://doi.org/10.1109/CVPR46437.2021.00617>
- [65] Madadi, M.; Bertiche, H.; Escalera, S.: SMPLR: Deep learning based SMPL reverse for 3D human pose and shape recovery, *Pattern Recognition*, 106, 2020, <https://doi.org/10.1016/j.patcog.2020.107472>
- [66] Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; Black, M. J.: AMASS: Archive of motion capture as surface shapes, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 5442-5451. <https://doi.org/10.1109/ICCV.2019.00554>
- [67] Mao, W.; Ge, Y.; Shen, C.; Tian, Z.; Wang, X.; Wang, Z. J. a. p. a.: Tfpose: Direct human pose estimation with transformers, 2021, <https://doi.org/10.48550/arXiv.2103.15320>
- [68] Mathis, A.; Schneider, S.; Lauer, J.; Mathis, M. W.: A Primer on Motion Capture with Deep Learning: Principles, Pitfalls, and Perspectives, *Neuron*, 108(1), 2020, 44-65. <https://doi.org/10.1016/j.neuron.2020.09.017>
- [69] Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision, 2017 international conference on 3D vision (3DV), 2017, 506-516. <https://doi.org/10.1109/3DV.2017.00064>
- [70] Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Elgharib, M.; Fua, P.; Seidel, H.-P.; Rhodin, H.; Pons-Moll, G.; Theobalt, C.: XNect: Real-time multi-person 3D motion capture with a single RGB camera, *Acm Transactions On Graphics (TOG)*, 39(4), 2020, 82: 81-82: 17. <https://doi.org/10.1145/3386569.3392410>.
- [71] Muller, L.; Osman, A. A.; Tang, S.; Huang, C.-H. P.; Black, M. J.: On self-contact and human pose, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 9985-9994. <https://doi.org/10.1109/CVPR46437.2021.00986>
- [72] Mustafa, A.; Caliskan, A.; Agapito, L.; Hilton, A.: Multi-person implicit reconstruction from a single image, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 14469-14478. <https://doi.org/10.1109/CVPR46437.2021.01424>
- [73] Osman, A. A.; Bolkart, T.; Black, M. J.: Star: Sparse trained articulated human body regressor, *European Conference on Computer Vision*, 2020, 598-613. https://doi.org/10.1007/978-3-030-58539-6_36
- [74] Patel, P.; Huang, C.-H. P.; Tesch, J.; Hoffmann, D. T.; Tripathi, S.; Black, M. J.: AGORA: Avatars in geography optimized for regression analysis, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 13468-13478. <https://doi.org/10.1109/CVPR46437.2021.01326>
- [75] Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; Black, M. J.: Expressive body capture: 3d hands, face, and body from a single image, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 10967-10977. <https://doi.org/10.1109/CVPR.2019.01123>

- [76] Plantard, P.; H Shum, H. P.; Multon, F.: Filtered pose graph for efficient kinect pose reconstruction, *Multimed Tools Appl*, 76(3), 2017, 4291-4312. <https://doi.org/10.1007/s11042-016-3546-4>
- [77] Plantard, P.; Shum, H. P.; Le Pierres, A.-S.; Multon, F.: Validation of an ergonomic assessment method using Kinect data in real workplace conditions, *Applied Ergonomics*, 65, 2017, 562-569. <https://doi.org/10.1016/j.apergo.2016.10.015>
- [78] Pumarola, A.; Sanchez-Riera, J.; Choi, G.; Sanfeliu, A.; Moreno-Noguer, F.: 3dpeople: Modeling the geometry of dressed humans, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 2242-2251. <https://doi.org/10.1109/ICCV.2019.00233>
- [79] Qiu, H.; Wang, C.; Wang, J.; Wang, N.; Zeng, W.: Cross view fusion for 3d human pose estimation, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 4342-4351. <https://doi.org/10.1109/ICCV.2019.00444>
- [80] Qiu, L.; Zhang, X.; Li, Y.; Li, G.; Wu, X.; Xiong, Z.; Han, X.; Cui, S.: Peeking into occluded joints: A novel framework for crowd pose estimation, *European Conference on Computer Vision*, 2020, 488-504. https://doi.org/10.1007/978-3-030-58529-7_29
- [81] Remelli, E.; Han, S.; Honari, S.; Fua, P.; Wang, R.: Lightweight multi-view 3d pose estimation through camera-disentangled representation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 6039-6048. <https://doi.org/10.1109/CVPR42600.2020.00608>
- [82] Richards, J. G.: The measurement of human motion: A comparison of commercially available systems, *Human Movement Science*, 18(5), 1999, 589-602. [https://doi.org/10.1016/S0167-9457\(99\)00023-8](https://doi.org/10.1016/S0167-9457(99)00023-8)
- [83] Rogez, G.; Weinzaepfel, P.; Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5), 2020, 1146-1161. <https://doi.org/10.1109/TPAMI.2019.2892985>
- [84] Romero, J.; Tzionas, D.; Black, M. J.: Embodied hands: Modeling and capturing hands and bodies together, *ACM Trans Graph*, 36(6), 2022, 1-17. <https://doi.org/10.1145/3130800.3130883>
- [85] Rong, Y.; Shiratori, T.; Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration, 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, 1749-1759. <https://doi.org/10.1109/ICCVW54120.2021.00201>
- [86] Saito, S.; Yang, J.; Ma, Q.; Black, M. J.: SCANimate: Weakly supervised learning of skinned clothed avatar networks, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 2885-2896. <https://doi.org/10.1109/CVPR46437.2021.00291>
- [87] Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 4510-4520. <https://doi.org/10.1109/CVPR.2018.00474>
- [88] Santesteban, I.; Garces, E.; Otaduy, M. A.; Casas, D.: SoftSMPL: Data - driven Modeling of Nonlinear Soft - tissue Dynamics for Parametric Humans, *Computer Graphics Forum*, 39(2), 2020, 65-75. <https://doi.org/10.1111/cgf.13912>
- [89] Shi, M.; Aberman, K.; Aristidou, A.; Komura, T.; Lischinski, D.; Cohen-Or, D.; Chen, B.: Motionet: 3d human motion reconstruction from monocular video with skeleton consistency, *Acm T Graphic*, 40(1), 2020, 1-15. <https://doi.org/10.1145/3407659>
- [90] Shum, H. P.; Ho, E. S.; Jiang, Y.; Takagi, S.: Real-time posture reconstruction for Microsoft Kinect, *IEEE Transactions on Cybernetics*, 43(5), 2013, 1357-1369. <https://doi.org/10.1109/TCYB.2013.2275945>
- [91] Sigal, L.; Balan, A. O.; Black, M. J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, 87(1), 2010, 4-27. <https://doi.org/10.1007/s11263-009-0273-6>
- [92] Sonnenfeld, J. J.; Crutchfield, C. R.; Swindell, H. W.; Schwarz, W. J.; Trofa, D. P.; Ahmad, C. S.; Lynch, T. S.: An Analysis of In Vivo Hip Kinematics in Elite Baseball Batters Using a

- Markerless Motion-Capture System, *Arthrosc Sports Med Rehabil*, 3(3), 2021, e909-e917. <https://doi.org/10.1016/j.asmr.2021.03.006>
- [93] Sun, K.; Xiao, B.; Liu, D.; Wang, J.: Deep high-resolution representation learning for human pose estimation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 5686-5696. <https://doi.org/10.1109/CVPR.2019.00584>
- [94] Toshev, A.; Szegedy, C.: Deeppose: Human pose estimation via deep neural networks, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, 1653-1660. <https://doi.org/10.1109/CVPR.2014.214>
- [95] Tu, H.; Wang, C.; Zeng, W.: Voxelpose: Towards multi-camera 3d human pose estimation in wild environment, *European Conference on Computer Vision*, 2020, 197-212. https://doi.org/10.1007/978-3-030-58452-8_12
- [96] von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera, *European Conference on Computer Vision (ECCV)*, 2018, 601-617. https://doi.org/10.1007/978-3-030-01249-6_37
- [97] Wandt, B.; Rosenhahn, B.: Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 7774-7783. <https://doi.org/10.1109/CVPR.2019.00797>
- [98] Wang, C.; Zhang, F.; Zhu, X.; Ge, S. S.: Low-resolution human pose estimation, *Pattern Recognition*, 126, 2022, 108579. <https://doi.org/10.1016/j.patcog.2022.108579>
- [99] Wang, J.; Liu, L.; Xu, W.; Sarkar, K.; Theobalt, C.: Estimating egocentric 3d human pose in global space, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 11480-11489. <https://doi.org/10.1109/ICCV48922.2021.01130>
- [100] Wang, J.; Qiu, K.; Peng, H.; Fu, J.; Zhu, J.: Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance, *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, 374-382. <https://doi.org/10.1145/3343031.3350910>
- [101] Wang, K.; Xie, J.; Zhang, G.; Liu, L.; Yang, J.: Sequential 3D human pose and shape estimation from point clouds, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 7273-7282. <https://doi.org/10.1109/CVPR42600.2020.00730>
- [102] Wang, M.; Tighe, J.; Modolo, D.: Combining detection and tracking for human pose estimation in videos, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 11088-11096. <https://doi.org/10.1109/CVPR42600.2020.01110>
- [103] Wang, Y.; Li, M.; Cai, H.; Chen, W.-M.; Han, S.: Lite Pose: Efficient Architecture Design for 2D Human Pose Estimation, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 13116-13126. <https://doi.org/10.1109/CVPR52688.2022.01278>
- [104] Wang, Z.; Lu, Y.; Ni, W.; Song, L.: An RGB-D Based Approach for Human Pose Estimation, 2021 International Conference on Networking Systems of AI (INSAI), 2021 of Conference, 166-170. <https://doi.org/10.1109/INSAI54028.2021.00039>
- [105] Wehrbein, T.; Rudolph, M.; Rosenhahn, B.; Wandt, B.: Probabilistic monocular 3d human pose estimation with normalizing flows, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 11199-11208. <https://doi.org/10.1109/ICCV48922.2021.01101>
- [106] Wei, W.-L.; Lin, J.-C.; Liu, T.-L.; Liao, H.-Y. M.: Capturing Humans in Motion: Temporal-Attentive 3D Human Pose and Shape Estimation from Monocular Video, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 13201-13210. <https://doi.org/10.1109/CVPR52688.2022.01286>
- [107] Weng, Z.; Yeung, S.: Holistic 3d human and scene mesh estimation from single view images, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 334-343. <https://doi.org/10.1109/CVPR46437.2021.00040>
- [108] Wu, J.; Zheng, H.; Zhao, B.; Li, Y.; Yan, B.; Liang, R.; Wang, W.; Zhou, S.; Lin, G.; Fu, Y.: Ai challenger: A large-scale dataset for going deeper in image understanding, 2019 IEEE International Conference on Multimedia and Expo (ICME), 2017, <https://doi.org/10.1109/ICME.2019.00256>

- [109] Xiang, D.; Joo, H.; Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 10957-10966. <https://doi.org/10.1109/CVPR.2019.01122>
- [110] Xiao, B.; Wu, H.; Wei, Y.: Simple baselines for human pose estimation and tracking, European conference on computer vision (ECCV), 2018, 472-487. https://doi.org/10.1007/978-3-030-01231-1_29
- [111] Xu, H.; Bazavan, E. G.; Zangir, A.; Freeman, W. T.; Sukthankar, R.; Sminchisescu, C.: GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 6183-6192. <https://doi.org/10.1109/CVPR42600.2020.00622>
- [112] Xu, J.; Yu, Z.; Ni, B.; Yang, J.; Yang, X.; Zhang, W.: Deep kinematics analysis for monocular 3d human pose estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 896-905. <https://doi.org/10.1109/CVPR42600.2020.00098>
- [113] Xu, L.; Guan, Y.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Ouyang, W.; Wang, X.: Vipnas: Efficient video pose estimation via neural architecture search, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 16067-16076. <https://doi.org/10.1109/CVPR46437.2021.01581>
- [114] Xu, L.; Xu, W.; Golyanik, V.; Habermann, M.; Fang, L.; Theobalt, C.: Eventcap: Monocular 3d capture of high-speed human motions using an event camera, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 4967-4977. <https://doi.org/10.1109/CVPR42600.2020.00502>
- [115] Xu, T.; Takano, W.: Graph stacked hourglass networks for 3d human pose estimation, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 16100-16109. <https://doi.org/10.1109/CVPR46437.2021.01584>
- [116] Yang, Y.; Ramanan, D.: Articulated human detection with flexible mixtures of parts, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), 2013, 2878-2890. <https://doi.org/10.1109/TPAMI.2012.261>
- [117] Yang, Y.; Ren, Z.; Li, H.; Zhou, C.; Wang, X.; Hua, G.: Learning dynamics via graph neural networks for human pose estimation and tracking, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 8070-8080. <https://doi.org/10.1109/CVPR46437.2021.00798>
- [118] Yeung, L.-F.; Yang, Z.; Cheng, K. C.-C.; Du, D.; Tong, R. K.-Y.: Effects of camera viewing angles on tracking kinematic gait patterns using Azure Kinect, Kinect v2 and Orbbec Astra Pro v2, Gait & Posture, 87, 2021, 19-26. <https://doi.org/10.1016/j.gaitpost.2021.04.005>
- [119] Yu, C.; Xiao, B.; Gao, C.; Yuan, L.; Zhang, L.; Sang, N.; Wang, J.: Lite-hrnet: A lightweight high-resolution network, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 10435-10445. <https://doi.org/10.1109/CVPR46437.2021.01030>
- [120] Yu, H.; Du, C.; Yu, L.: Scale-aware heatmap representation for human pose estimation, Pattern Recognition Letters, 154, 2022, 1-6. <https://doi.org/10.1016/j.patrec.2021.12.018>
- [121] Yu, Z.; Yoon, J. S.; Lee, I. K.; Venkatesh, P.; Park, J.; Yu, J.; Park, H. S.: Humbi: A large multiview dataset of human body expressions, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 2987-2997. <https://doi.org/10.1109/CVPR42600.2020.00306>
- [122] Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C.: Distribution-aware coordinate representation for human pose estimation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 7091-7100. <https://doi.org/10.1109/CVPR42600.2020.00712>
- [123] Zhang, H.; Tian, Y.; Zhou, X.; Ouyang, W.; Liu, Y.; Wang, L.; Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 11426-11436. <https://doi.org/10.1109/ICCV48922.2021.01125>
- [124] Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video, 2022 IEEE/CVF Conference on Computer Vision and

- Pattern Recognition (CVPR), 2022, 13222-13232.
<https://doi.org/10.1109/CVPR52688.2022.01288>
- [125] Zhang, T.; Huang, B.; Wang, Y.: Object-occluded human shape and pose estimation from a single color image, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 7374-7383. <https://doi.org/10.1109/CVPR42600.2020.00740>
- [126] Zhang, Y.; Wang, C.; Wang, X.; Liu, W.; Zeng, W. J. I. T. o. P. A.; Intelligence, M.: Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), 2022, 2613-2626.
<https://doi.org/10.1109/TPAMI.2022.3163709>
- [127] Zhang, Y.; You, S.; Gevers, T.: Automatic calibration of the fisheye camera for egocentric 3d human pose estimation from a single image, 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, 1772-1781.
<https://doi.org/10.1109/WACV48630.2021.00181>
- [128] Zhang, Z.; Wan, L.; Xu, W.; Wang, S.: Estimating a 2D pose from a tiny person image with super-resolution reconstruction, Computers & Electrical Engineering, 93, 2021, 107192.
<https://doi.org/10.1016/j.compeleceng.2021.107192>
- [129] Zhang, Z.; Wang, C.; Qin, W.; Zeng, W.: Fusing wearable imus with multi-view images for human pose estimation: A geometric approach, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 2197-2206.
<https://doi.org/10.1109/CVPR42600.2020.00227>
- [130] Zhang, Z.; Wang, C.; Qiu, W.; Qin, W.; Zeng, W.: Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild, International Journal of Computer Vision, 129(3), 2021, 703-718. <https://doi.org/10.1007/s11263-020-01398-9>
- [131] Zheng, C.; Wu, W.; Yang, T.; Zhu, S.; Chen, C.; Liu, R.; Shen, J.; Kehtarnavaz, N.; Shah, M.: Deep Learning-Based Human Pose Estimation: A Survey, ArXiv, abs/2012.13392, 2019,
<https://doi.org/10.48550/arXiv.2012.13392>
- [132] Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; Ding, Z.: 3d human pose estimation with spatial and temporal transformers, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 11656-11665. <https://doi.org/10.1109/ICCV48922.2021.01145>
- [133] Zhou, K.; Han, X.; Jiang, N.; Jia, K.; Lu, J.: HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 2344-2353. <https://doi.org/10.1109/ICCV.2019.00243>
- [134] Zhu, W.; Ma, X.; Liu, Z.; Liu, L.; Wu, W.; Wang, Y.: Learning Human Motion Representations: A Unified Perspective, 2022, <https://doi.org/10.48550/arXiv.2210.06551>
- [135] Zou, S.; Guo, C.; Zuo, X.; Wang, S.; Wang, P.; Hu, X.; Chen, S.; Gong, M.; Cheng, L.: EventHPE: Event-based 3D Human Pose and Shape Estimation, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 10976-10985.
<https://doi.org/10.1109/ICCV48922.2021.01081>