Review of models for estimating 3D human pose using deep learning

Sani Salisu[1], Kamaluddeen Usman Danyaro[2], Maged Nasser[2], Israa M. Hayder[3] and Hussain A. Younis[4]

[1] Department of Information Technology, Federal University Dutse, Dutse, Jigawa, Nigeria
[2] Computer & Information Sciences Department, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia
[3] Department of Computer Systems Techniques, Qurna Technique Institute, Southern Technical University, Basrah, Iraq
[4] College of Education for Women, University of Basrah, Basrah, Iraq

ABSTRACT

Human pose estimation (HPE) is designed to detect and localize various parts of the human body and represent them as a kinematic structure based on input data like images and videos. Three-dimensional (3D) HPE involves determining the positions of articulated joints in 3D space. Given its wide-ranging applications, HPE has become one of the fastest-growing areas in computer vision and artificial intelligence. This review highlights the latest advances in 3D deep-learning-based HPE models, addressing the major challenges such as accuracy, real-time performance, and data constraints. We assess the most widely used datasets and evaluation metrics, providing a comparison of leading algorithms in terms of precision and computational efficiency in tabular form. The review identifies key applications of HPE in industries like healthcare, security, and entertainment. Our findings suggest that while deep learning models have made significant strides, challenges in handling occlusion, real-time estimation, and generalization remain. This study also outlines future research directions, offering a roadmap for both new and experienced researchers to further develop 3D HPE models using deep learning.

Subjects Artificial Intelligence, Computer Vision, Data Mining and Machine Learning, Multimedia, Neural Networks
Keywords Human pose estimation, Deep learning, 3D image, Survey, Neural network, Review

Submitted 8 April 2024
Accepted 12 November 2024
Published 4 February 2025

Corresponding author
Sani Salisu, sani.salisu@fud.edu.ng

Academic editor
Ivan Miguel Pires

Additional Information and Declarations can be found on page 22

DOI **10.7717/peerj-cs.2574**

INTRODUCTION

Human pose estimation (HPE) means identifying the pose of human body segments and key points or joints in images, videos, and real-time environments. It involves tracking, detecting, and grouping the semantic key points of a given object for solving human problems such as clinical and rehabilitation solutions among others. HPE offers geometric and motion information about the human body which has been applied to a wide variety of applications such as human-computer interaction, motion analysis, augmented reality (AR), virtual reality (VR), healthcare, action recognition, animation, *etc*… and is used in several areas of human endeavours which include gaming industries, movies industries, entertainments, academic, professional research, *etc*. Different methods have been proposed for 3D human pose estimation, including silhouette contours (*Mondal, Ghosh & Ghosh, 2013*), edge-based histograms (*Mori & Malik, 2002*), pictorial structures (PS) (*Andriluka, Roth & Schiele, 2009*; *Dantone et al., 2013*) and deformable part models

How to cite this article Salisu S, Danyaro KU, Nasser M, Hayder IM, Younis HA. 2025. Review of models for estimating 3D human pose using deep learning. **PeerJ Comput. Sci. 11:e2574 DOI 10.7717/peerj-cs.2574**

(DPMs) (*Felzenszwalb et al., 2010*), continued to build appearance models for each key points separately. Due to the features complexity of 3D human pose estimation such as viewpoint invariant 3D feature maps (*Haque et al., 2016*), histograms of 3D joint locations, multifractal spectrum, and volumetric attention models, traditional models (*Barajas, Dávalos-Viveros & Gordillo, 2013*; *Chan, Koh & Lee, 2013*; *Dinh et al., 2013*; *Handrich & Al-Hamadi, 2013*; *Xul et al., 2013*; *Belagiannis et al., 2014*; *Jung & Kim, 2014*; *Liang et al., 2014*; *Zhu et al., 2014*) are not able to perform accurately in HPE research.

Despite the achievement in solving issues related to pose estimation, HPE methods are facing challenges in detecting, capturing, and extracting the significant key points of the human body. Such challenges include occlusion (self-occlusion, inter-person occlusion, and out-of-frame occlusion), limited data (limited annotation, limited variation of pose, limited number of pose), bad input data (blurry, low resolution, low light, low contrast, small scale, noisy), domain gap, camera-centric, crowd scenes, speed, complex pose, *etc*. HPE has recently attracted increasing attention in the computer vision community in facing those challenges. A large number of deep learning-based models have been developed by enhancing the existing model to deal with those challenges facing both 2D (*Li et al., 2021*; *Liu et al., 2022*) and 3D (*Saini et al., 2022*; *Wu et al., 2022*) pose estimation. An article review is one of the most significant and effective approaches for guiding future researchers about the state-of-the-art of any scientific domain.

However, most of the existing surveys and literature reviews on human pose estimation focused on 2D HPE (*Jingtian et al., 2020*; *Munea et al., 2020*; *Ulku & Akagündüz, 2022*) while survey and literature reviews on 3D models HPE are limited. Out of the few surveys and comprehensive literature reviews on 3D HPE (*Ji et al., 2020*; *Wang et al., 2021*; *Toshpulatov et al., 2022*; *Tian et al., 2023*; *Azam & Desai, 2024*), none of these focus on the progress of the state-of-the-art deep learning-based 3D human pose estimation models. For example, *Azam & Desai (2024)* aim to estimate human body poses and develop body representations from a first-person camera perspective, focusing solely on egocentric human pose estimation and neglecting other 3D pose estimation issues. *Ji et al. (2020)* concentrates on monocular 3D images, disregarding those captured using binocular or stereo vision systems. *Zhang et al. (2021)* extensively reviews deep learning supervision models but only covers research articles from 2013 to 2021, missing recent advancements. A general survey of both 2D and 3D human pose estimation, including classical and deep learning approaches, is presented (*Dubey & Dixit, 2023*), making it challenging for researchers to find specific methodologies. *Wang et al. (2021)* published a review on 3D deep learning-based human estimation, but it heavily emphasizes 3D pose estimation datasets Furthermore, the literature reviews some existing 3D HPE researchers and presents a description of their methods and model architectures.

RELATED WORK

Wang et al. (2023) reviews deep learning methods for 3D pose estimation, summarizes their pros and cons, and examines benchmark datasets for comparison and analysis, offering insights to guide future model and algorithm designs. A novel method to extract 3D information from 2D images without 3D pose supervision using 2D pose annotations and perspective knowledge to generate relative depth of joints was proposed (Qiu et al., 2023). The authors introduced a 2D pose dataset (MCPC) and a weakly-supervised pre-training (WSP) strategy for depth prediction. WSP improves depth prediction and generalization for 3D human pose estimation, achieving state-of-the-art results on benchmark datasets. In Pavlakos et al. (2019) a 3D model of body pose, hand pose and facial expression from a single monocular image using Skinned Multi-Person Linear (SMPL)-X was computed. The authors improved upon the SMPLify approach by detecting 2D features for the face, hands, and feet, training a new neural network pose prior, defining a fast interpenetration penalty, and automatically detecting gender. Their newly implemented SMPLify-X significantly speeds up fitting SMPL-X to images. Sun et al. (2020) address the issue of monocular 3D human pose estimation with deep learning. To overcome occlusion problems inherent in single-view methods, they proposed an end-to-end network that generates multi-view 2D poses from single-view 2D poses, uses data augmentation for multi-view 2D pose annotations, and employs a graph convolutional network to infer 3D poses.

Clemente et al. (2024) explores the feasibility of a model for 3D HPE from monocular 2D videos (MediaPipe Pose) in a physiotherapy context, by comparing its performance to ground truth measurements. MediaPipe Pose was investigated in eight exercises typically performed in musculoskeletal physiotherapy sessions, where the range of motion (ROM) of the human joints was the evaluated parameter. This model showed the best performance for shoulder abduction, shoulder press, elbow flexion, and squat exercises. Results have shown that their model has achieved a higher performance. He et al. (2024) proposed a novel approach for telerehabilitation based on deep learning 3D human pose estimation. Their approach aims to evaluate the effectiveness and practicality of the telerehabilitation method over a 12-week experiment through a randomized controlled trial on older adults with sarcopenia, this study compared the training effects of an AI-based remote training group using deep learning-based 3D human pose estimation technology with those of a face-to-face traditional training group and a general remote training group.

HUMAN BODY MODELLING

Because humans vary in their shapes and sizes, modelling the human body is a crucial aspect of HPE. To determine the posture of an individual, their body needs to meet the specific criteria necessary for a particular task to establish and describe the human body's pose. HPE frequently employs five distinct categories of human body models that include the kinematic-base model, planner model, volumetric model, SMPL-based model and Surface-based model (Wang et al., 2021; Salisu et al., 2023).

Skeletal-based model

This model is commonly referred to as a kinematic-based model or a stick figure, characterized by its uncomplicated and adaptable representation of the human body's structure. It finds frequent application in both 2D (Cao et al., 2017) and 3D HPE (Mehta et al., 2018). This model primarily captures the positions of joints and limb orientations to depict the human body's skeletal structure, facilitating the detection of connections between different body parts. This human skeleton model is conceptualized as a tree-like structure, encompassing numerous key points in the human body. It establishes connections between neighbouring joints through edges. The fusion of a convolutional neural network pose regressor and kinematic skeletal fitting allows for the real-time capture of a comprehensive 3D skeletal pose in a given environment using only a single RGB camera (*Terreran, Barcellona & Ghidoni, 2023*). While the kinematic model offers flexibility in graph representation, it has constraints in effectively representing texture and shape information.

Contour-based model

The contour-model, also known as the planer model stands in contrast to the kinematic-based model. In the contour-based model, essential points are approximated with rectangular shapes or object boundaries. This model is primarily employed to represent the outline and structure of the human body. The planar model is a common choice in classical HPE methods (*Jiang, 2010*), such as those that utilize techniques like cardboard (*Freifeld et al., 2010*) mode and active shape modelling to capture the human body's structure and silhouette distortions through principal component analysis (PCA). In this model, researchers commonly depict body parts using rectangles that approximate the contours of the human body.

Many scholars employ this model to address issues related to the relationships between various human body parts. For instance, *Toshpulatov et al. (2022)* applied the planar model to capture the connections between different body parts, and their findings indicate that this model effectively represents the shape and appearance of the human body. *Ju, Black & Yacoob (1996)* proposed that an individual can be portrayed as a collection of interconnected planar patches. Their study demonstrated that limbs can be represented by these planar patches, offering a useful approach for tracking human legs across extended image sequences. Additionally, *Black & Yacoob (1995)* illustrated that a planar model can accurately approximate the motion of a human head, providing a succinct depiction of optical flow within a specific region.

SMPL-based model

The Skinned Multi-Person Linear (SMPL) model, as introduced by *Loper et al. (2015)*, serves as a tool for predicting 3D human body joint locations, as described by *Bogo et al. (2016)*. This model represents the human skin as a mesh with 6,890 vertices, which can be adjusted through shape and pose parameters. Shape parameters are employed to capture aspects like body proportions, height, and weight, while pose parameters account for the specific deformations of the body. By learning and optimizing these shape and body parameters, one can estimate the 3D positions of the body's pose. Several researchers attempted to address the issue of SMPL-based by utilising single or combined models.

Other researchers shifted their attention to an enhanced version of the SMPL model, recognizing its limitations, such as computational complexity and the absence of facial and hand landmarks. Some of these researchers sought to overcome these constraints. For instance, *Xiu et al. (2022)* introduced an iterative refinement of SMPL parameters during

3D reconstruction. Similarly, *Lin, Wang & Liu (2021a, 2021b)* put forward as methods for reconstructing 3D human pose and mesh from a single image, without depending on any parametric mesh model like SMPL.

Surface-based model

A more recent human body model known as DensePose proposed by *Güler, Neverova & Kokkinos (2018)* has been introduced in response to the limitations of sparse image key points in comprehensively describing the human body's condition. To overcome this limitation, a fresh dataset called DensePose-COCO has been created. This dataset establishes detailed correspondences between image pixels and a surface-based representation of the human body, enhancing the ability to capture the human body's state.

Volumetric model

Volume-based models are employed to depict the silhouette and pose of a three-dimensional object using a geometric mesh. Traditional geometric meshes used for modelling human body parts included shapes like cylinders and cones. In contrast, modern volume-based models are recognised by their mesh representations derived from 3D scans. These models represent the body as a 3D volume, often using voxels (3D pixels) or implicit functions to capture both the outer shape and potential internal structures. Volumetric models give a complete spatial representation of the body (*Salisu et al., 2023*). Among the most widely used volumetric models for 3D pose estimation are the Stitched Puppet Model (SPM) and the Unified Deformation Model (UDM) (*Trumble et al., 2017*), as well as models like Frankenstein & Adam and the Generic Human Model (GHUM) along with the Low-Resolution Generic Human Model (GHUML) (*Xu et al., 2020*).

METHODOLOGY

On the topic of "3D human pose estimation using deep learning," an extensive study was conducted, investigating key research questions and systematically searching and organizing the relevant literature. Our research methodology adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol (*Moher et al., 2009*) to outline the data sources, search strategies, and criteria for literature inclusion.

Research questions

The survey articles aim to address the following questions

RQ1 How 3D human pose estimation evolved and developed.
RQ2 How are neural networks applied to different 3D pose estimation task.
RQ3 What areas of human endeavour where human pose estimation is applicable.
RQ4 What are the current challenges of 3D HPE.
RQ5 What are the future research directions for 3D human pose estimation.

Research strategy and data sources

This review utilized automatic and manual search methods to ensure optimal outcomes. Four databases (Scopus, Web of Science, Google Scholar, and IEEE) were employed to locate pertinent articles or research within the 3D human pose estimation (3DHPE) domain and its applications. Specific search queries, comprising various keywords and their combinations were employed to identify relevant publications from 2014 to 2024. These queries included terms such as “3D human pose estimation,” “3D human body model,” “3D datasets and evaluation metrics,” “3D HPE application,” and “deep-learning based 3DHPE.” To ensure impartial coverage, identical queries were executed across all four databases.

Inclusion/exclusion criteria

Given the objective of conducting a comprehensive review tailored to the study’s requirements, slight variations in search strategies were adopted, considering the unique search capabilities of each selected database. Initially, title and abstract searches were conducted in IEEE and Scopus, whereas full-text searches were performed in Web of Science and Google Scholar. Retrieved articles underwent scrutiny based on their abstracts and titles to determine inclusion or exclusion eligibility. Articles lacking sufficient or relevant information were excluded.

Subsequently, the full text of screened papers was meticulously examined to ascertain their relevance for inclusion or exclusion. Additionally, references cited within selected articles were identified and utilized to retrieve additional relevant papers for the study. To ensure the generation of clean and standardized documents, devoid of noise and duplicates, supplementary selection and rejection criteria were applied. These criteria stipulated that articles must be written in English, and published in English journals or conferences between the years 2014-2024. Furthermore, the articles were required to employ deep-learning techniques rather than classical methods.

RESULTS

The literature search identified 601 articles on 3D human pose estimation. After removing 299 duplicates, 302 unique articles remained. Screening the titles and abstracts reduced this number to 97. Following a full-text review, 67 additional articles were excluded, leaving 30 relevant studies for inclusion in the review. Figure 1 illustrates the stages of the search from the input query, the inclusion and exclusion criteria to the final inclusion stage. A total of three studies ($n = 3$) were published in 2024, and eight studies ($n = 8$) were published in 2023. The remaining studies were distributed as follows: seven ($n = 7$) in 2022, six ($n = 6$) in 2021, three ($n = 3$) in 2020, two ($n = 2$) in 2019, and one ($n = 1$) in 2017. Figure 2 illustrates the growing research interest in the field of 3D human pose estimation (3DHPE). A critical evaluation was conducted on all 30 contributions that met the filtering criteria. In 3DHPE research, the choice of dataset plays a crucial role. Some researchers used benchmark datasets (*Yin, Lv & Shao, 2023*), while others employed self-prepared datasets (*Niu et al., 2024*; *Yan et al., 2024*) or a combination of both

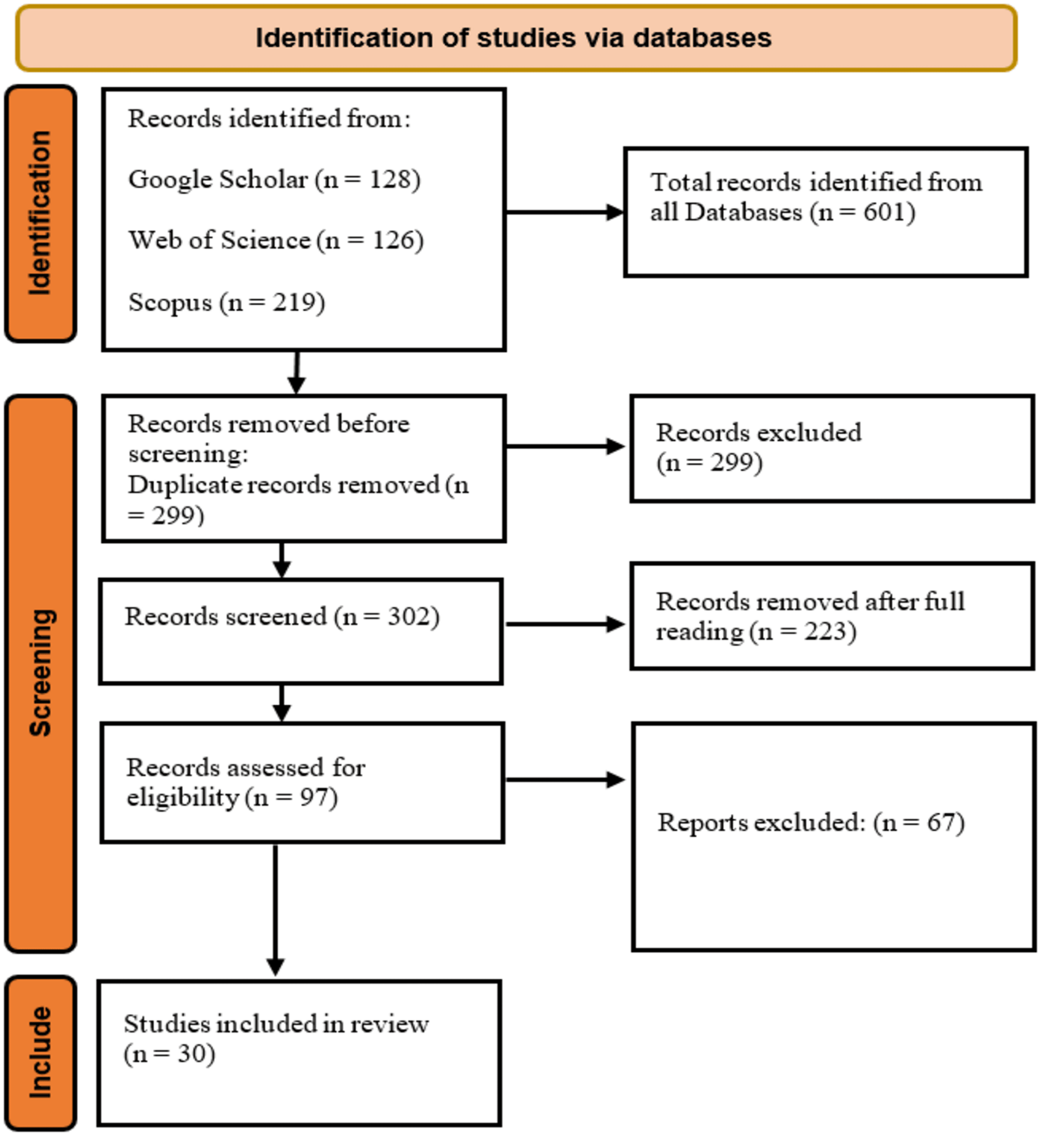
Identification of studies via databases
Identification
Screening
Include
Records identified from:
Google Scholar (n = 128)
Web of Science (n = 126)
Scopus (n = 219)
Total records identified from all Databases (n = 601)
Records removed before screening:
Duplicate records removed (n = 299)
Records excluded (n = 299)
Records screened (n = 302)
Records removed after full reading (n = 223)
Records assessed for eligibility (n = 97)
Reports excluded: (n = 67)
Studies included in review (n = 30)

Figure 1 Article selection process using PRISMA protocol. Direct component sources. https://creativecommons.org/licenses/by/4.0/. Full-size DOI: 10.7717/peerj-cs.2574/fig-1

(*Xi et al., 2024*). Human 3.6M emerged as the most frequently used dataset among the identified studies. A summary of the critical evaluation of all 30 contributions is presented in Table 1.

RQ1: HOW 3D HUMAN POSE ESTIMATION EVOLVED AND DEVELOPED

The fundamental process of HPE comprises three main phases. Firstly, it involves identifying key points/joints in the human body, such as the knee, ankle, shoulder, head, arms, and hands. This initial stage is crucial for pinpointing the specific locations of these key points. The choice of human pose dataset format plays a significant role in gathering and recognizing key points stored in selected 2D datasets. It is important to note that the

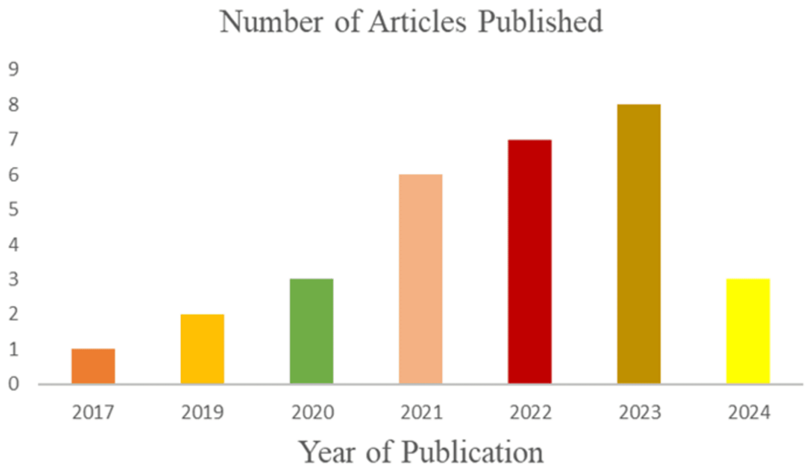

Figure 2 **The number of published articles from 2017 to 2024.** The growing research interest in the field of 3D human pose estimation (3DHPE). Full-size DOI: 10.7717/peerj-cs.2574/fig-2

Table 1 **The summary of the current deep learning-based article from Scopus.** Summarizes the literature of 30 of the most recently published articles in Scopus, illustrating the study, the model utilized, the type of dataset and a summary of the aim, achievement made and the limitation of each study.

Study	Model	Dataset	Comments
Xi et al. (2024)	TCN, T-MHSA	Human 3.6M and MPI-INF-3DHP	**Aim:** to enhance the precision of pose estimation. **Achievement:** effectiveness in handling real-time and complex sports scenarios. **Limitation:** scalability, the robustness of the approach across various sports.
Zhang et al. (2023c)	P2P-MeshNet	Self-prepared (FreeMocap)	**Aim:** to estimate the body joint rotations directly **Achievement:** potential application prospects of the method **Limitation:** Could be further evaluated on different datasets for generalisation.
Yan et al. (2024)	IP +DL + EA	Self-prepared	**Aim:** Provide a reliable and efficient solution for 3D human pose estimation and awkward posture detection in a construction environment. **Achievement:** Contributes to the proactive management of worker health and safety. **Limitation:** Not applicable to health care monitoring data
Martini et al. (2022)	MAEVE platform	MSCOCO	**Aim:** Implement a low-cost, real-time and usable platform for 3DPE that guarantees accuracy. **Achievement:** Real-time performance and higher accuracy. **Limitation:** Future improvements could focus on enhancing adaptability across various HPE scenarios and patient populations to strengthen its clinical utility
Ran et al. (2023)	De-occlusion multi-task learning network	3DPW, 3DPW-OCC, 3DOH, and Human 3.6 M	**Aim:** To de-noise the feature for mesh parameter regression **Achievement:** Competitive performance on a non-occlusion dataset **Limitation:** Real-world applicability and processing speed remain potential areas for future work

Table 1 (continued)

Study	Model	Dataset	Comments
Vukicevic et al. (2021)	VIBE	Self-prepared	**Aim:** To utilize IoT force sensors and IP cameras to detect unsafe P&P acts timely and objectively. **Achievemen**t: Excellent covenant with motion sensors and a high potential for checking and improving the safety of the P&P workplace. **Limitation:** Not applicable to P&P datasets
Bigalke et al. (2023)	Domain adaptation	Self-prepared and SLP dataset	**Aim:** Implement a model from a labeled source to a shifted unlabeled target domain **Achievement:** Outperformed the SOTA method (baseline and gap) **Limitation:** Tested on SLP and MVIBP datasets only
Yin, Lv & Shao (2023)	COG	Human 3.6M	**Aim:** To design a multi-branch network based on the human center of gravity **Achievement**: higher efficiency and validity **Limitation:** Need further exploration into integrating additional contextual features
Xu et al. (2022)	RSC-Net	Human 3.6M and MPI-INF-3DHP	**Aim:** To implement a method that can deal with the issues of low-resolution input images or video. **Achievement:** Accurate learning of 3D body pose and shape across different resolutions with one single model. **Limitation:** Future work might focus on further refining the robustness of cross-view matching in complex real-world settings.
Mehta et al. (2017)	VNect	MPI-INF-3DHP and Human 3.6M	**Aim:** To develop a real-time model that captures the full body 3D skeletal human pose in a steady, temporally reliable manner using a single RGB camera. **Achievement:** Better applicability than RGB-D solutions. **Limitation:** Not capable of estimating 3D poses from different camera views
Li et al. (2020a)	Dual–Stage pipeline	3D human pose datasets, Human 3.6M and MPI INF-3DHP.	**Aim:** To develop a self-supervised model to avoid manual annotations of 3D poses. **Achievement:** more effectiveness compared to the current weekly-supervised model. **Limitation:** Relying on the completeness of the shape prior provided by RGBD-PIFu.
Saini et al. (2022)	UAVs + AirPose	Self-prepared	**Aim:** To develop a new model that estimates human pose and shape using images captured by multiple irrelevantly uncelebrated flying cameras **Achievement:** 3D HPE system for unstructured, uncontrolled and outdoor environments **Limitation:** Not robust to various lighting conditions and diversifying training data.
Yang et al. (2022)	PoseMoNet	Human 3.6M and HumanEva-I	**Aim:** To develop an elf-projection mechanism that cogently conserves human motion kinematics. **Achievement:** A competitive advantage compared to SOTA **Limitation:** Limited mechanisms to handle more varied and complex motions.
Gao, Yang & Li (2022)	GroupPoseNet	InterHands2.6M	**Aim:** To develop a 3D PE model that can differentiate between two hand shapes and poses from a single RGB image. **Achievement:** Higher efficiency in the overall result. **Limitation:** Lack of generalization capabilities for the real-world environment.

(Continued)

Table 1 (continued)

Study	Model	Dataset	Comments
Kourbane & Genc (2022)	Two-stage GCN-based	STB and RHD	**Aim:** To develop a model that learns per-pose relationship constraints in estimating 3D hand pose. **Achievement:** Outperforms SOTA in terms of accurate 3D hand pose estimation. **Limitation:** Future research could explore optimizing computational demands and validating model generalizability across more diverse, real-world scenarios.
Pavlakos et al. (2019)	SMPLify-X	EHF and SMPL+H	**Aim:** Simplify the analysis of human actions and emotions estimation. **Achievement:** Higher speed than the SOTA methods. **Limitation:** lack of a dataset of in-the-wild SMPL-X fits, which restricts the current ability to learn a regressor that can directly estimate SMPL-X parameters from RGB images.
Mehrizi et al. (2019)	DNN + Hourglass network	Self-prepared	**Aim:** This study aims to develop and validate a DNN-based model for 3D pose estimation during lifting **Achievement:** Higher accuracy even with the shortcomings of marker-based motion systems. **Limitation:** The number and position of cameras were not explored, the study focused on lifting tasks only and markers on the body may alter the natural appearance of the body.
Zhu et al. (2023)	MHPT	MoVi	**Aim:** To develop a deep-learning human pose technique for clinical gait analysis. **Achievement:** pose estimations have been improved significantly. **Limitation:** The system needs further validation across diverse patient groups and conditions to assess its reliability fully in clinical practice.
Kong & Kang (2021)	3DMPPE	Human 3.6M	**Aim:** To develop a model to reduce computation cost and processing time. **Achievement:** Reducing the processing time and the performance of the model. **Limitation:** Lack of optimisation and adaptability to diverse scenarios.
Zou et al. (2021)	EventHPE	Self-prepared and DHP19	**Aim:** To develop a stage deep learning model for accurate pose estimation. **Achievement:** Effectiveness of the new model. **Limitation:** Could enhance its real-time performance and adaptability to diverse environments.
Šajina & Ivašić-Kos (2022)	Tracking	Self-prepared	**Aim:** To develop a new model, that will enhance the 3D sequences of poses generated from the designed testing model. **Achievement:** Revealing the drawbacks of other methods in the field of HPE. **Limitation:** Lacks of advanced occlusion handling techniques to improve reliability across various sports contexts.
Zhang et al. (2023a)	PoseAug	Human 3.6M	**Aim:** To develop a model that will cater for the variation of 2D and 3D pose pairs. **Achievement:** higher improvement on both frame and video-based 3D HPE. **Limitation:** The model needs Further testing to confirm its generalizability.
Zhou, Dong & EI Saddik (2020)	DGCNN	ITOP and EVAL	**Aim:** To develop a model that will solve the issues of 3DHPE using depth images. **Achievement:** Higher accuracy than SOTA models **Limitation:** Not applicable to a fast and occluded movement.

Table 1 (continued)

Study	Model	Dataset	Comments
Retsinas, Efthymiou & Maragos (2023)	Template mushroom model	Self-prepared	**Aim:** To develop a deep learning model that solves the annotation problem and estimates their pose on 3D data. **Achievement:** more effectiveness compared to SOTA models. **Limitation:** Further research could enhance the approach's robustness through domain adaptation and extensive testing on real-world data.
Ding et al. (2021)	Kinematic Constrained Learning	Self-prepared	**Aim:** To develop a learning model for predicting skeletal key points from observed radar data. **Achievement:** The fusion of kinematic constraints with the learning of 3D skeletal reconstruction. **Limitation:** The model's performance should be validated in varied environments
Ying & Zhao (2021)	3D learning module	MHAD and SURREAL	**Aim:** To develop a model that can estimate 3D human pose from RGB-D images. **Achievement:** SOTA performance on the stated datasets. **Limitation:** further validation in diverse settings and improvements in computational efficiency would be necessary for broader real-world applicability.
Niu et al. (2024)	SPCT+TRP	Self-prepared	**Aim:** To attack the intrinsic difficulties of the classical model **Achievement:** Improves accuracy and robustness in challenging conditions. **Limitation:** Lack of broader testing and validation across different settings
Rapczyński et al. (2021)	Scale normalization+ OpenPose	HumanEva-I, Human 3.6M, and Panoptic Studio,	**Aim:** To develop a model that will tackle the issue of both cross and dataset generalisation. **Achievement:** Improvements in cross-dataset and in-dataset generalisation. **Limitation:** Manual parameterisation for each new dataset.
Manesco, Berretti & Marana (2023)	Domain Unified Approach	SURREAL and Human 3.6M	**Aim:** solving pose misalignment problems on a cross-dataset scenario. **Achievement:** showing significant improvements in cross-domain accuracy by leveraging the domain adaptation technique **Limitation:** Future directions might involve testing across more diverse datasets and addressing computational efficiency for real-time use
Sun et al. (2020)	End-to-end 3DPEN	Human 3.6M	**Aim:** to address the issue of not using single-view images directly in multi-view methods. **Achievement:** higher effectiveness and performance improvement **Limitation:** Not validated on complex poses and diverse datasets.

output of body key points from the same image may differ based on the dataset format and platform employed.

Moving on to the second stage, pose estimation entails grouping the localized key points to form valid human pose configurations, thereby determining pairs of organs in the human body. Various researchers have experimented with different techniques for connecting key point candidates in this stage.

The third and final stage involves estimating a 3D pose based on the previously determined 2D key points. This is achieved by combining sample 2D frames captured at different times through a temporal procedure known as a temporal convolution neural network.

3D human pose estimation

In recent years, machine learning approaches have significantly transformed pose estimation, with deep learning (DL) methods making notable strides in enhancing its performance. While substantial progress has been made in 2D HPE, the task of 3D HPE remains challenging. Many existing studies address 3D HPE using monocular images or videos, presenting an ill-posed and inverse problem due to the loss of one dimension in the projection from 2D to 3D. However, when multiple views or additional sensors like Inertial Measuring Units (IMU) are utilized, 3D HPE becomes a well-posed problem that can benefit from information fusion techniques. Below is the description of 3D HPE using various sensors, including camera sensors, IMU sensors, point cloud and depth sensors, and radiofrequency device sensors.

3D HPE from digital RGB images and videos

3D HPE using deep learning from digital RGB images and video can be classified into three categories: single-view single-person 3D HPE, single-view multi-person 3D HPE, and multi-view 3D HPE.

A single-view single-person 3D HPE: Methods for single-person 3D HPE can be categorized into two groups, namely, skeleton-only and human mesh recovery (HMR). This classification is based on whether the goal is to reconstruct a 3D human skeleton or to recover a 3D human mesh using a human body model.

Skeleton-only: Methods falling into the skeleton-only category produce 3D human joint estimates as their ultimate output. These approaches do not utilize human body models for reconstructing a 3D human mesh representation. Within this category, these methods can be further subcategorized into direct estimation approaches and 2D to 3D lifting approaches.

Regression-based estimation

Regression-based estimation methods deduce the 3D human pose directly from 2D images, without the need for an intermediate step of estimating a 2D pose representation. This study (*Liang, Sun & Wei, 2018*) introduced a regression approach that considers the data's structure. Instead of relying on a joint-based representation, they opted for a more stable bone-based description. They defined a compositional loss by leveraging the 3D bone structure, using the bone-based description to encode long-range relations between the bones. A volumetric approach to transform the challenging non-linear 3D coordinate regression task into a handier form within a discretized space was presented (*Pavlakos et al., 2017*; *Pavlakos, Zhou & Daniilidis, 2018*). A convolutional network was employed to predict voxel likelihoods for each joint in the volume. They utilized ordinal depth relations of human joints to mitigate the requirement for precise 3D ground truth pose information.

2D to 3Dlifting

Inspired by the recent achievements in 2D human pose estimation (HPE), the popularity of 2D to 3D lifting approaches has risen. These methods involve inferring the 3D human pose from an interim estimation of the 2D human pose. During the initial phase, pre-existing 2D HPE models are utilized to predict the 2D pose. Subsequently, in the second stage, the process of 2D to 3D lifting is applied to derive a 3D pose. A fully connected residual network for the regression of 3D joint locations, relying on the provided 2D joint locations was introduced by *Tralic et al. (2013)*. While this method achieved state-of-the-art results during its time, its susceptibility to failure stemmed from the reconstruction ambiguity resulting from excessive dependence on the 2D pose detector. The most optimal 3D pose was identified in *Jahangiri & Yuille (2017)*, *Li & Lee (2019)*, *Sharma et al. (2019)* by initially producing various 3D pose possibilities and subsequently utilising ranking networks to optimise the output.

Human body recovery

Human body recovery (HBR) methods also known as human mesh recovery integrate parametric body models, as outlined in Section 2 (human body modelling), to reconstruct the human mesh. The 3D pose is then acquired by utilizing the joint regression matrix defined by the model. Common examples of HBR are SMPL-based volumetric-based and surface-based models.

Single-view multi-person 3D HPE

Multi-person 3D HPE from monocular RGB images or videos has a more compound and demanding challenge, involving the identification of the number of individuals, their positions, and poses, followed by the grouping of their localized body key points and finally estimating the 3D pose. To address these challenges, multi-person pose estimation can be categorized into two techniques: Top-down and Bottom-up.

Top-down techniques

Top-down approaches in 3D multi-person HPE initially engage in human detection to identify each person. Subsequently, for every detected individual, the absolute root coordinate and 3D root-relative pose are estimated through 3D pose networks. This process involves inputting the image into the human detection network, cropping the detected humans, aligning them to the world coordinate, and finally estimating the 3D root-relative pose.

Some researchers like *Zou et al. (2023)* introduce Snipper, a cohesive framework designed to execute multi-person 3D pose estimation, tracking, and motion forecasting concomitantly within a single stage. Their approach incorporates a proficient yet robust deformable attention mechanism, enabling the aggregation of spatiotemporal information from the video snippet. Leveraging this deformable attention mechanism, they train a video transformer to capture spatiotemporal features from the multi-frame snippet and generate informative pose features for multi-person pose queries. Ultimately, these pose queries are processed to predict both multi-person pose trajectories and future motions in a single shot.

Bottom-up techniques

Differing from top-down strategies, bottom-up approaches initially generate the locations of all body joints and depth maps. Subsequently, these methods associate body parts with each individual based on the root depth and relative depth of each part. Grouping the human body joints belonging to each person is one of the major challenges of the bottom-up technique. For example, *Xiao et al. (2023)* presented a refined method for body representation and a streamlined single-stage multi-person pose regression network named AdaptivePose++. The innovative body representation can adequately capture diverse pose information and efficiently model the connection between a human instance and its associated key points within a single forward pass.

Multi-view multi-person 3D HPE

Multi-view 3D HPE encounters difficulties in handling partial occlusion when operating in a single-view setting. A viable approach to address this challenge involves estimating the 3D human pose from various viewpoints. This is because the obscured portions in one view might be visible in other views. To achieve 3D pose reconstruction from multiple perspectives, it is crucial to resolve the association of corresponding locations across different cameras. Many researchers attempted to overcome the issue of partial occlusion associated with single-view 3D HPE (*Wang et al., 2021b*; *Liu, Wu & He, 2022*; *Xu & Kitani, 2022*; *Gerats, Wolterink & Broeders, 2023*; *Silva et al., 2023*). Generally, multi-view settings are primarily employed for multi-person pose estimation.

3D HPE from digital image with imu sensors

Among the renounced sensors capable of tracking the orientation and acceleration of the human body, Wearable inertial measurement units (IMUs) have proven to be among the best. It can monitor the orientation and acceleration of various human body parts by capturing movements without being hindered by object occlusions or clothing obstructions. Many researchers (*Huang et al., 2020b*; *Zhang et al., 2020*; *Liao et al., 2023a*, *2023b*; *Zhao et al., 2023a*) proposed IMU-based pipelines to reconstruct 3D human poses to improve pose accuracy.

Cloud and depth sensors

One of the hurdles encountered in 3D human pose estimation is the ambiguity related to depth. In recent years, there has been an increasing interest among computer vision researchers in employing depth sensors due to their precision and cost-effectiveness. Numerous studies have suggested using depth images for the estimation of 3D human poses (*Yu et al., 2018*; *Xiong et al., 2019*; *Zhou, Bhatnagar & Pons-Moll, 2020*; *Wang et al., 2023*). Additionally, research indicates that utilizing points can exhibit excellent performance in recovering 3D human mesh and other models of human pose (*Jiang, Cai & Zheng, 2019*; *Wang et al., 2020*; *Gu et al., 2022*; *Hermes, Bigalke & Heinrich, 2023*).

Radio frequency sensors

Radiofrequency sensing devices have demonstrated an efficient and promising result in estimating 3D human poses. Researchers like *Zhao et al. (2019)*, and *Xie et al. (2023)* utilised radio frequency techniques in their work and promising output was obtained. The significant advantage of employing an RF-based sensing system lies in its capacity to move through walls and rebound off human bodies within the WiFi range without the need for carrying wireless transmitters. This approach provides a major benefit. Additionally, privacy can be maintained as the data is non-visual. However, it is worth noting that RF signals exhibit a relatively lower spatial resolution when compared to visual camera images, and RF systems have been demonstrated to yield coarse 3D pose estimations (*Zhao et al., 2018*; *Xie et al., 2022*, *2023*).

Datasets for 3D human pose estimation

There are many datasets for 3D human pose estimation. For this article, only the most widely used deep learning–based 3D human pose estimation datasets are included.

Human 3.6M is the most widely used indoor dataset for 3D HPE from monocular images and videos. This dataset contains 3.6 million 3D human poses with 3D ground truth annotation captured by an accurate marker-based motion capture system. The training and testing images are categorized into subjects S1, S5, S6, and S7 for training, and images of subjects S9 and S11 for testing. Several researchers (*Reddy et al., 2021*; *Zhu et al., 2022*; *Chun, Park & Chang, 2023b*, *2023a*; *Shan et al., 2023*) utilised the Human 3.6M dataset for estimating the human pose.

MuPoTS-3D is another 3D dataset in which 3D poses were captured by a multi-view marker-less motion capture system containing 20 real-world scenes. This dataset contains many images with occlusions, drastic illumination changes, and lens flares. A total of 8,000 frames were collected in the 20 sequences by eight subjects. There are many more datasets for 3D human pose estimation that are not covered due to several reasons including page limitation. Many researchers utilised this dataset in their work. In another research performed by *Haque et al. (2016)*, *Sun et al. (2023)*, *Xing (2023)*, this dataset is also utilized.

3D Poses in the Wild dataset (3DPW) is the first unconstrained dataset in the wild with accurate 3D poses for evaluation. This dataset consists of about 60 video sequences, 3D body scans, and 3D object models out of which 18D models are different clothing. Among all the 3D benchmark datasets, 3DPW is reported to be the first dataset that contains video footage covered by a mobile phone camera. Its unique peculiarities inspired many researchers (*Cho et al., 2023*; *Ma et al., 2023*; *Nam et al., 2023*; *Oreshkin, 2023*; *Zhang et al., 2023b*) in the field of computer vision.

MPI-INF-3DHP, More than 1.3 million frames from 14 cameras were recorded in a green screen studio which allows automatic segmentation and augmentation. The dataset contains many human activities including walking, sitting, complex exercise poses, and dynamic actions. Other 3D HPE datasets consist of either outdoor or indoor scenes but MPI-INF-3DHP is reported to accommodate both complex outdoor and controlled indoor scenes respectively. This dataset is one of the top datasets utilized by many researchers including (*Jiang et al., 2023*; *Mehraban, Adeli & Taati, 2023*; *Yu et al., 2023*; *Zhao et al., 2023b*).

Evaluation metrics for 3D human pose estimation

Evaluating the performances of 3D human pose estimation is quite challenging because there are many features and equipment that need to be considered. Therefore, unlike like in 2D human pose estimation, there limited number of evaluation matric covered in this review.

Mean per joint position error

It is the most frequently used metric to evaluate the performance of 3D HPE. Mean per joint position error (MPJPE) is calculated using the Euclidean distance between the projected 3D joints and the ground truth locations. Some researchers (*Güler, Neverova & Kokkinos, 2018*; *Wandt et al., 2021*) applied this metric to evaluate the performances of their models

Mean average precision

This metric is used to measure the performance of predictions in the 3D dataset. Detection is successful when the predicted 3D body landmark falls within a distance less than the assigned mAP (mean average precision) value from the ground truth. The mAP is utilised as the evaluation metric to determine the unique rate of sample points (*Wang, Chen & Fu, 2022*).

3DPCK

It is a 3D extended version of the percentage of correct keypoints (PCK) metric used in 2D HPE evaluation. An estimated joint is said to be correct if the distance between the estimation and the ground truth is within a certain threshold. Generally, the threshold is set to 150 m. This metric is applied by *Pavlakos et al. (2019)* to evaluate the performance of their model.

Euler angle error

It is a standard practice used to measure the error of 3D pose prediction in the Human 3.5 M dataset the evaluation is done by computing the Euclidean norm predictions and the ground truth Euler angle representations. The metric was applied in the Simulation experiments to verify the effectiveness of the proposed model (*Li et al., 2018*).

RQ2: HOW ARE NEURAL NETWORKS APPLIED TO DIFFERENT 3D POSE ESTIMATION TASK

Research on 3B human pose estimation is gaining more popularity today. Many articles have been published in different journals and book chapters. Some of these articles published in Scopus are reviewed and summarized in this study as follows.

In *Yu et al. (2023)* an effective model known as Global-local Adaptive Graph Convolutional Network (GLA-GCN) was proposed to improve the 3D human pose lifting *via* ground truth data to improve the quality of estimated pose data. Three benchmark datasets were used in the experiments and the output shows that the proposed model

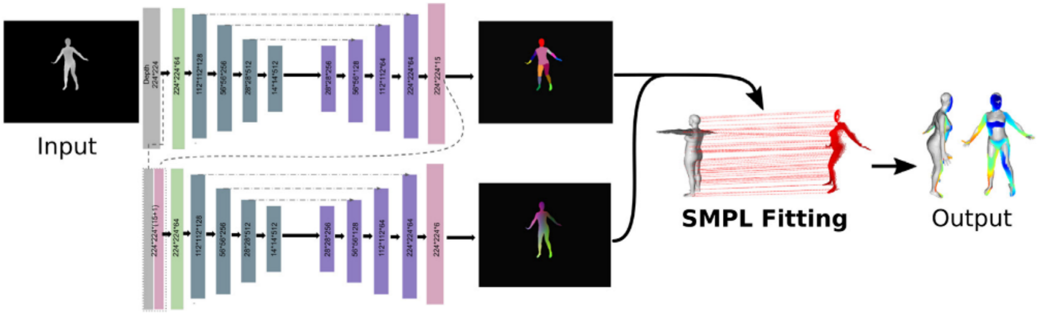

Figure 3 The architecture of the proposed hybrid mode to predict 3D human pose and shape (*Wang et al., 2023*). Direct License. https://s100.copyright.com/CustomerAdmin/PLF.jsp?ref=cef4c52c-cb41-4996-9ed0-5b42de22dfc5. Full-size DOI: 10.7717/peerj-cs.2574/fig-3

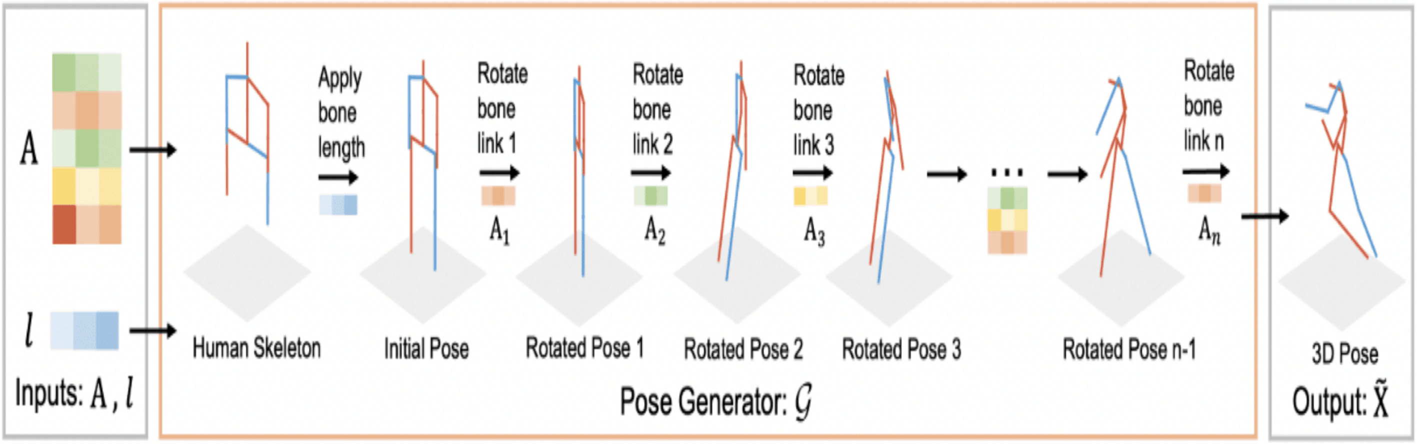

Figure 4 3D human pose generator (*Guan et al., 2023*). Full-size DOI: 10.7717/peerj-cs.2574/fig-4

performed better than the state-of-the-art model. A spatial-temporal mesh attention convolution (MAC) was developed (*Wang et al., 2020*) to predict the 3D coordinates of mesh vertices at the high resolution based on the estimated 3D coordinates and features at the low resolution. The framework was generalised to the real data of human bodies with a weakly supervised fine-tuning method. The developed model achieves higher accuracy in recovering the 3D body model sequence from a sequence of point clouds.

To tackle the shortcomings of estimating the pose and shape of human bodies (*Wang et al., 2023*) developed a hybrid pipeline that put together the strength of both DL-based and sensor-based models. After performing the experiments with four benchmark datasets which include the DFAUST dataset, SURREAL, and some parts of AMASS dataset, the outcome of the experiments shows that this hybrid approach enables us to enhance pose and shape estimation compared to using DL or model fitting separately. The hybrid model is presented in Fig. 3.

A novel human pose generator was developed (*Guan et al., 2023*) to generate diverse 3D human poses. The performance of the proposed model known as PoseGU was evaluated

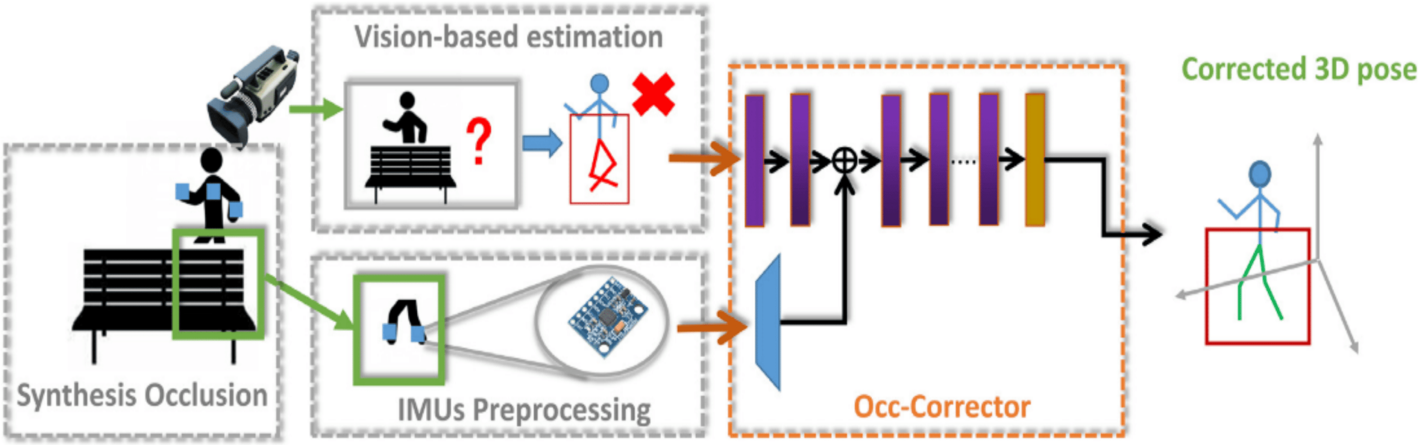

Figure 5 **Occ-Corrector framework (*Zhao et al., 2023a*).** Full-size DOI: 10.7717/peerj-cs.2574/fig-5

with three different benchmark datasets and the experimental analysis shows that it generates 3D poses with improved data diversity and better generalization ability. The framework of the developed module is shown in Fig. 4.

Estimating the 3D pose of humans in a clinical setting requires a simple and accurate system. The current method is attached to many devices which makes it difficult in a clinical environment. A model that attempts to reduce the number of devices and maintain its accuracy is developed by *Zhao et al. (2023a)*. A semantic convolution-based neural network known as Occ-Corrector is developed to deal with occlusion issues that might arise during the estimation of 3D poses from a single camera by integrating with IMU's sensors. Occ-Corrector is presented in Fig. 5.

The summary of the current deep learning-based article from Scopus is shown in Table 1. Table 1, summarizes the literature of 30 of the most recently published articles in Scopus, illustrating the Study, the model utilized, the type of dataset and a summary of the aim achievement made and the limitation of each study.

RQ3 WHAT AREAS OF HUMAN ENDEAVOUR WHERE HUMAN POSE ESTIMATION IS APPLICABLE

Area of application

In terms of real-world applications, 3D human pose estimation provides numerous applications in human endeavours. Here some popular applications of 3D HPE are reviewed and presented.

Business activities

Human pose estimation has proven valuable in both real-world and virtual business settings. The influence of e-commerce trends, particularly in areas like face masks and clothing purchases, has been noteworthy. Traditional depictions of clothing items in images are no longer sufficient to meet customer expectations. Consumers now desire a more dependable representation, wanting to visualize how selected clothes will look on them. Through 3D human pose estimation (3D HPE), it becomes feasible to generate realistic representations of human body regions for virtual fitting rooms, specifically for clothes and face mask analysis. This is achieved through techniques such as clothes parsing (*Saito et al., 2019*; *Yu et al., 2019*) and pose transfer (*Li, Huang & Loy, 2019*), which infer the three-dimensional appearance of an individual wearing specific garments.

Entertainment

In the gaming, film, and animation sectors, 3D HPE plays a pivotal role. While motion capture systems serve as the foundation for these industries, addressing the intricate movements of actors, 3D HPE is increasingly employed as a cost-effective alternative to high-priced motion capture devices. With the help of 3D pose estimation and human mesh recovery, 3D character animation from a single photo is developed by *Weng, Curless & Kemelmacher-Shlizerman (2019)*.

Healthcare

Human pose estimation finds practical application in the fields of human rehabilitation and physiotherapy. It involves tracking human activities during rehabilitation exercises by precisely identifying key points of the patient's movements for effective treatment. *Gu et al. (2019)* proposed a physiotherapy system designed to assess and guide patients at home. Additionally, *Weiming Chen, Guo & Ni (2020)* utilized HPE algorithms for monitoring fall detection, enabling prompt assistance. Consequently, 3D pose estimation techniques have the potential to develop a system for correcting sitting postures. Such a system could monitor user status, provide quantitative human motion information, aid in diagnosing complex diseases, formulate rehabilitation training, and facilitate physical therapy under the guidance of physicians.

Sport activities

In sports, coaches and trainers analyse and monitor athletes' performance. The advancements in 3D HPE have enabled Artificial Intelligence (AI) trainers to provide precise coaching through action detection techniques using just a few camera settings. *Hwang, Park & Kwak (2017)* developed a system, leveraging 3D poses extracted from videos, for performance analysis, rapid response, and improvement. *Zecha et al. (2018)* employed human pose estimation techniques to precisely track and assess athletes' performance in swimming exercises, ensuring accurate metrics at any given time. *Wang et al. (2019)* developed an AI coaching system incorporating a pose estimation module, offering personalised assistance for athletic training.

Security/safety

3D human pose estimation finds application in surveillance, involving the estimation and analysis of a person's poses or activities in a surveillance environment. Many advanced shopping malls and stores have implemented cashier-less systems to identify customers engaging in suspicious behaviour. These establishments aim to identify any irregularities among their customers by employing a hybrid computer vision system that combines camera sensor networks and Internet of Things (IoT) devices with HPE. Human pose estimation plays a crucial role in scenarios where the actual contact between the customer and the product is not visible to the camera. In such cases, the HPE model analyses the positions of customers' hands and heads to determine whether they have taken a product from the shelf or left it in place. This system utilizes pose information for action recognition, tracking, prediction, and detection. Researchers like *Angelini et al. (2018)* have proposed real-time action recognition methods using pose-based algorithms, while human action detection in videos has also been explored (*Cao et al., 2020*).

RQ4: WHAT ARE THE CURRENT CHALLENGES OF 3D HPE

Challenges

No doubt about how powerful deep leaning-based models are, in dealing with different types of computer vision problems specifically human pose estimation. Unlike 2D HPE, 3D HPE models are facing different challenges in the implementation of deep learning–based pipelines.

Occlusion

Is one of the main challenges facing the estimation of 3D human pose in a deep learning-based approach. The act of occluding the target object is done by covering the required key points of that object which automatically affects the accuracy of 3D pose estimation. Occlusion may be self-occlusion, inter-person-occlusion, out-of-frame occlusion, *etc*.

Inadequate data

One of the requirements of any deep learning model to enable accurate learning is the availability of data. Researchers in this field are facing the serious challenge of limited data with limited poses, inadequate pose variations, and inadequate annotation. Deep learning models are data-driven models that require large data with different poses for training and calibration. Thus inadequate data will affect the performance of any deep learning model.

Tampered input data

Data that is being tampered with may affect the accuracy of estimation despite the efficiency of the deep learning model. Some data are tampered with by adding noise, blurring, low or high contrast, and resolution which seriously affect the estimation accuracy.

Complex scene

The complexity of an image particularly in multi-person pose estimation can affect the 2D pose estimation and also the 3D pose estimation. When the 2D estimation is not accurate, estimating the 3D can be affected since the 3D pipeline estimates the 2D key points first before estimating the 3D joints. An example of this is heavy crowd, football games and so on.

RQ5: WHAT ARE THE FUTURE RESEARCH DIRECTIONS FOR 3D HUMAN POSE ESTIMATION

Multi-person reconstruction

In every community setting, individuals frequently engage in activities such as walking, talking, or collaborating in groups, such as families or teams. A compelling avenue for future exploration involves reconstructing groups of people across both spatial and temporal dimensions, thereby unveiling relationships and activities within the targeted group. Additionally, when addressing person matching across various cameras or extended temporal sequences, leveraging the relationships among individuals within a group offers a more stable context that can be utilized to address challenges like occlusions or detection failures. This task can also be integrated with person tracking (*Rajasegaran et al., 2021*) and re-identification (*Lisanti et al., 2017*) to enhance the robustness of reconstruction, especially in crowded scenarios.

Physical constraints

Many current approaches overlook the interaction between humans and 3D scenes. There exist significant constraints in the relationship between humans and scenes, such as the inability of a human subject to occupy the same locations as other objects in the scene simultaneously. By exploring these physical constraints alongside semantic cues, it is possible to enhance the reliability and realism of 3D HPE.

Full body reconstruction

Parametric models such as SMPL and SMPL-X are limited to representing individuals with minimal clothing. To surpass the representational limitations of parametric models, the research community must explore alternative models that offer greater flexibility. Previous studies have employed meshes, point clouds (*Ma et al., 2021*), and implicit fields (*Li et al., 2020b*) to capture detailed clothing deformation. While these approaches can yield reasonable outcomes, the reconstructed surfaces often appear overly smoothed and lack robustness when confronted with novel poses. Addressing these issues involves incorporating diverse types of representations (*Shao et al., 2022*) to harness the modeling capabilities of varied approaches.

Adaptation of HPE domain

In certain scenarios, like the estimation of human pose from images of infants (*Huang et al., 2021*) or collections of artwork (*Madhu et al., 2022*), there is a scarcity of training data accompanied by accurate ground truth annotations. Additionally, the data for these applications showcase distributions distinct from those found in standard pose datasets. HPE methods trained on conventional datasets might struggle to generalize effectively across diverse domains. A recent approach to mitigate this domain gap involves the use of generative adversarial network (GAN)-based learning techniques. However, the effective transfer of human pose knowledge to bridge these domain gaps remains an unexplored challenge.

Generalized metrics

3D human pose estimation finds applications in visual tracking and analysis. Current methods for reconstructing 3D human pose and shape from videos lack smooth and continuous results. One contributing factor is that evaluation metrics like MPJPE do not assess smoothness and the level of realism adequately. There is a need to develop suitable frame-level evaluation metrics that specifically address temporal consistency and motion smoothness.

CONCLUSIONS

This article explores recent advancements in deep learning-based 3D human pose estimation models. First, we provide a brief introduction to human pose estimation and its various types, followed by an in-depth discussion of the evolution, and achievements of these models. Second, we review studies that apply 3D pose estimation models to different tasks, such as hand pose, full-body pose, and human activities, highlighting the main challenges in the field. Third, we explain the diverse applications of 3D human pose estimation across various domains and list all potential neural network modules for addressing 3D HPE problems. Fourth, we offer readers a comprehensive understanding of existing approaches and outline detailed future directions for deep learning-based 3D human pose estimation. Lastly, our review aims to serve as a comprehensive guide for both industry and academic practitioners, providing a significant and rich understanding of various aspects of estimating three-dimensional human poses while advocating for continued advancements in the field.

ACKNOWLEDGEMENTS

We express our genuine gratitude to everyone who offered their support and assistance throughout the research endeavour, contributing to the successful completion of this survey.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This research is funded by the Universiti Teknologi PETRONAS through a Short-Term Internal Research Funding (STIRF) Grant (Grant Number: 015LA0-057) (previously known as Cost Centre: YUTP-FRG 015LC0-373). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Universiti Teknologi PETRONAS through a Short-Term Internal Research Funding (STIRF): 015LA0-057.
Previously known as Cost Centre: YUTP-FRG 015LC0-373.

Competing Interests

The authors declared that there is no competing interest.

Author Contributions

- Sani Salisu conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Kamaluddeen Usman Danyaro conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Maged Nasser conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Israa M. Hayder performed the computation work, prepared figures and/or tables, and approved the final draft.
- Hussain A. Younis conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The article is a literature review.

REFERENCES

Andriluka M, Roth S, Schiele B. 2009. Pictorial structures revisited: people detection and articulated pose estimation. *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR* **2009**:1014–1021 DOI 10.1109/CVPR.2009.5206754.

Angelini F, Fu Z, Long Y, Shao L, Naqvi SM. 2018. ActionXPose: a novel 2D multi-view pose-based algorithm for real-time human action recognition. ArXiv preprint 1–14 DOI 10.48550/arXiv.1810.12126.

Azam MM, Desai K. 2024. A survey on 3D egocentric human pose estimation. ArXiv preprint 1643–1654 DOI 10.48550/arXiv.2403.17893.

Barajas M, Dávalos-Viveros JP, Gordillo JL. 2013. 3D tracking and control of UAV using planar faces and monocular camera. In: Carrasco-Ochoa JA, Martínez-Trinidad JF, Rodríguez JS, di Baja GS, eds. *Pattern Recognition. MCPR 2013. Lecture Notes in Computer Science.* Vol. 7914. Berlin, Heidelberg: Springer, 64–73 DOI 10.1007/978-3-642-38989-4_7.

Belagiannis V, Amin S, Andriluka M, Schiele B, Navab N, Ilic S. 2014. 3D pictorial structures for multiple human pose estimation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1669–1676 DOI 10.1109/CVPR.2014.216.

Bigalke A, Hansen L, Diesel J, Hennigs C, Rostalski P, Heinrich MP. 2023. Anatomy-guided domain adaptation for 3D in-bed human pose estimation. *Medical Image Analysis* **89(7)**:102887 DOI 10.1016/j.media.2023.102887.

Black MJ, Yacoob Y. 1995. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In: *IEEE International Conference on Computer Vision*, 374–381 DOI 10.1109/iccv.1995.466915.

Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ. 2016. Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Leibe B, Matas J, Sebe N, Welling M, eds. *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science.* Vol. 9909. Cham: Springer, 561–578 DOI 10.1007/978-3-319-46454-1_34.

Cao Z, Gao H, Mangalam K, Cai QZ, Vo M, Malik J. 2020. Long-term human motion prediction with scene context. In: Vedaldi A, Bischof H, Brox T, Frahm JM, eds. *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science.* Vol. 12346. Cham: Springer, 387–404 DOI 10.1007/978-3-030-58452-8_23.

Cao Z, Simon T, Wei SE, Sheikh Y. 2017. Realtime multi-person 2D pose estimation using part affinity fields. In: *Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua*, 1302–1310 DOI 10.1109/CVPR.2017.143.

Chan K, Koh C, Lee CSG. 2013. Using action classification for human-pose estimation. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Tokyo, Japan, 1176–1181 DOI 10.1109/IROS.2013.6696499.

Cho H, Cho Y, Ahn J, Kim J. 2023. Implicit 3D human mesh recovery using consistency with pose and shape from unseen-view. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2023-June*, 21148–21158 DOI 10.1109/CVPR52729.2023.02026.

Chun S, Park S, Chang JY. 2023a. Learnable human mesh triangulation for 3D human pose and shape estimation. *Proceedings—2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023* **1**:2849–2858 DOI 10.1109/WACV56688.2023.00287.

Chun S, Park S, Chang JY. 2023b. Representation learning of vertex heatmaps for 3D human mesh reconstruction from multi-view images. *2023b IEEE International Conference on Image Processing (ICIP)*. Kuala Lumpur, Malaysia, 670–674 DOI 10.1109/ICIP49359.2023.10222297.

Clemente C, Chambel G, Silva DCF, Montes AM, Pinto JF, da Silva HP. 2024. Feasibility of 3D body tracking from monocular 2D video feeds in musculoskeletal telerehabilitation. *Sensors* **24(1)**:1–19 DOI 10.3390/s24010206.

Dantone M, Gall J, Leistner C, Van Gool L. 2013. Human pose estimation using body parts dependent joint regressors. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3041–3048 DOI 10.1109/CVPR.2013.391.

Ding W, Cao Z, Zhang J, Chen R, Guo X, Wang G. 2021. Radar-based 3D human skeleton estimation by kinematic constrained learning. *IEEE Sensors Journal* **21(20)**:23174–23184 DOI 10.1109/JSEN.2021.3107361.

Dinh DL, Han HS, Jeon HJ, Lee S, Kim TS. 2013. Principal direction analysis-based real-time 3D human pose reconstruction from a single depth image. In: *ACM International Conference Proceeding Series*, 206–212 DOI 10.1145/2542050.2542071.

Dubey S, Dixit M. 2023. A comprehensive survey on human pose estimation approaches. *Multimedia Systems* **29(1)**:167–195 DOI 10.1007/s00530-022-00980-0.

Felzenszwalb PF, Girshick RB, Mcallester D, Ramanan D. 2010. Object detection with partbase. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32(9)**:1627–1645 DOI 10.1109/TPAMI.2009.167.

Freifeld O, Weiss A, Zuffi S, Black MJ. 2010. Contour people: a parameterized model of 2D articulated human shape. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 639–646 DOI 10.1109/CVPR.2010.5540154.

Gao C, Yang Y, Li W. 2022. 3D interacting hand pose and shape estimation from a single RGB image. *Neurocomputing* **474(4)**:25–36 DOI 10.1016/j.neucom.2021.12.013.

Gerats BGA, Wolterink JM, Broeders IAMJ. 2023. 3D human pose estimation in multi-view operating room videos using differentiable camera projections. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization* **11(4)**:1197–1205 DOI 10.1080/21681163.2022.2155580.

Gu X, Guo Y, Yang GZ, Lo B. 2022. Cross-domain self-supervised complete geometric representation learning for real-scanned point cloud based pathological gait analysis. *IEEE Journal of Biomedical and Health Informatics* **26(3)**:1034–1044 DOI 10.1109/JBHI.2021.3107532.

Gu Y, Pandit S, Saraee E, Nordahl T, Ellis T, Betke M. 2019. Home-based physical therapy with an interactive computer vision system. In: *Proceedings—2019 International Conference on Computer Vision Workshop, ICCVW 2019*, 2619–2628 DOI 10.1109/ICCVW48693.2019.

Guan S, Lu H, Zhu L, Fang G. 2023. PoseGU: 3D human pose estimation with novel human pose generator and unbiased learning. *Computer Vision and Image Understanding* **233(7)**:1–19 DOI 10.1016/j.cviu.2023.103715.

Güler RA, Neverova N, Kokkinos I. 2018. DensePose: dense human pose estimation in the wild. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7297–7306 DOI 10.1109/CVPR.2018.00762.

Handrich S, Al-Hamadi A. 2013. A robust method for human pose estimation based on geodesic distance features. In: *Proceedings—2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, 906–911 DOI 10.1109/SMC.2013.159.

Haque A, Peng B, Luo Z, Alahi A, Yeung S, LF F. 2016. Towards viewpoint invariant 3D human pose estimation. *Eccv* **9905**:398–413 DOI 10.1007/978-3-319-46448-0.

He S, Meng D, Wei M, Guo H, Yang G, Wang Z. 2024. Proposal and validation of a new approach in tele-rehabilitation with 3D human posture estimation: a randomized controlled trial in older individuals with sarcopenia. *BMC Geriatrics* **24(1)**:1–15 DOI 10.1186/s12877-024-05188-7.

Hermes N, Bigalke A, Heinrich MP. 2023. Point cloud-based scene flow estimation on realistically deformable objects: a benchmark of deep learning-based methods. *Journal of Visual Communication and Image Representation* **95(1)**:103893 DOI 10.1016/j.jvcir.2023.103893.

Huang X, Fu N, Liu S, Ostadabbas S. 2021. Invariant representation learning for infant pose estimation with small data. In: *Proceedings—2021 16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021* DOI 10.1109/FG52635.2021.9666956.

Huang F, Zeng A, Liu M, Lai Q, Xu Q. 2020b. DeepFuse: an IMU-Aware network for real-time 3D human pose estimation from multi-view images. In: *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, 418–427 DOI 10.1109/WACV45572.2020.9093526.

Hwang J, Park S, Kwak N. 2017. Athlete pose estimation by a global-local network. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2017-July*, 114–121 DOI 10.1109/CVPRW.2017.20.

Jahangiri E, Yuille AL. 2017. Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections. In: *Proceedings—2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017 2018-Janua*, 805–814 DOI 10.1109/ICCVW.2017.100.

Ji X, Fang Q, Dong J, Shuai Q, Jiang W, Zhou X. 2020. A survey on monocular 3D human pose estimation. *Virtual Reality and Intelligent Hardware* **2(6)**:471–500 DOI 10.1016/j.vrih.2020.04.005.

Jiang H. 2010. *Finding human poses in videos using concurrent matching and segmentation.* Queenstown, New Zealand: ACCV, 228–243.

Jiang H, Cai J, Zheng J. 2019. Skeleton-aware 3D human shape reconstruction from point clouds. In: *Proceedings of the IEEE International Conference on Computer Vision 2019-Octob*, 5430–5440 DOI 10.1109/ICCV.2019.00553.

Jiang Z, Zhou Z, Li L, Chai W, Yang C-Y, Hwang J-N. 2023. Back to optimization: diffusion-based zero-shot 3D human pose estimation. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA, 6130–6140 DOI 10.1109/WACV57701.2024.00603.

Jingtian S, Xue C, Yanan L, Jianwen C. 2020. 2D human pose estimation from monocular images: a survey. In: *2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology, CCET 2020*, 111–121 DOI 10.1109/CCET50901.2020.

Ju SX, Black MJ, Yacoob Y. 1996. Cardboard people: a parameterized model of articulated image motion. In: *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 38–44 DOI 10.1109/afgr.1996.557241.

Jung S, Kim M. 2014. Estimation of a 3D bounding box for a segmented object region in a single image. *IEICE Transactions on Information and Systems* **E97D(11)**:2919–2934 DOI 10.1587/transinf.2014EDP7098.

Kong DH, Kang SJ. 2021. Downsizing heatmap resolution for real-time 3D human pose estimation. In: *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications, ITC-CSCC 2021*, 1–4 DOI 10.1109/ITC-CSCC52171.2021.9501409.

Kourbane I, Genc Y. 2022. A hybrid classification-regression approach for 3D hand pose estimation using graph convolutional networks. *Signal Processing: Image Communication* **101(2)**:116564 DOI 10.1016/j.image.2021.116564.

Li J, Dang P, Li Y, Gu B. 2018. A general Euler angle error model of strapdown inertial navigation systems. *Applied Sciences (Switzerland)* **8(1)**:74 DOI 10.3390/app8010074.

Li Y, Huang C, Loy CC. 2019. Dense intrinsic appearance flow for human pose transfer. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June*, 3688–3697 DOI 10.1109/CVPR.2019.00381.

Li C, Lee GH. 2019. Generating multiple hypotheses for 3D human pose estimation with mixture density network. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June*, 9879–9887 DOI 10.1109/CVPR.2019.01012.

Li Y, Li K, Jiang S, Zhang Z, Huang C, Da Xu RY. 2020a. Geometry-driven self-supervised method for 3D human pose estimation. In: *AAAI, 2020—34th AAAI Conference on Artificial Intelligence*, 11442–11449 DOI 10.1609/aaai.v34i07.6808.

Li Z, Yu T, Pan C, Zheng Z, Liu Y. 2020b. Robust 3D self-portraits in seconds. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1341–1350 DOI 10.1109/CVPR42600.2020.00142.

Li Y, Zhang S, Wang Z, Yang S, Yang W, Xia ST, Zhou E. 2021. TokenPose: learning keypoint tokens for human pose estimation. In: *Proceedings of the IEEE International Conference on Computer Vision*, 11293–11302 DOI 10.1109/ICCV48922.2021.01112.

Liang S, Sun X, Wei Y. 2018. Compositional human pose regression. *Computer Vision and Image Understanding* **176-177**:1–8 DOI 10.1016/j.cviu.2018.10.006.

Liang D, Weng K, Wang C, Liang G, Chen H, Wu X. 2014. A 3D object recognition and pose estimation system using deep learning method. In: *ICIST 2014—Proceedings of 2014 4th IEEE International Conference on Information Science and Technology*, 401–404 DOI 10.1109/ICIST.2014.6920502.

Liao X, Dong J, Song K, Xiao J. 2023a. Three-dimensional human pose estimation from sparse IMUs through Temporal encoder and regression decoder. *Sensors* **23(7)**:1–13 DOI 10.3390/s23073547.

Liao X, Zhuang J, Liu Z, Dong J, Song K, Xiao J. 2023b. Reconstructing 3D human pose and shape from a single image and sparse IMUs. *PeerJ Computer Science* **9(4)**:1–25 DOI 10.7717/peerj-cs.1401.

Lin K, Wang L, Liu Z. 2021a. End-to-end human pose and mesh reconstruction with transformers. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1954–1963 DOI 10.1109/CVPR46437.2021.00199.

Lin K, Wang L, Liu Z. 2021b. Mesh graphormer. In: *Proceedings of the IEEE International Conference on Computer Vision*, 12919–12928 DOI 10.1109/ICCV48922.2021.01270.

Lisanti G, Martinel N, Del BA, Foresti GL. 2017. Group re-identification via unsupervised transfer of sparse features encoding. In: *Proceedings of the IEEE International Conference on Computer Vision 2017-Octob*, 2468–2477 DOI 10.1109/ICCV.2017.268.

Liu H, Member S, Liu T, Zhang Z. 2022. ARHPE: asymmetric relation-aware representation learning for head pose estimation in industrial human-computer interaction. *IEEE Transactions on Industrial Informatics* **18**:7107–7117 DOI 10.1109/TII.2022.3143605.

Liu H, Wu J, He R. 2022. Centre point to pose: multiple views 3D human pose estimation for multi-person. *PLOS ONE* **17(9)**:1–15 DOI 10.1371/journal.pone.0274450.

Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. 2015. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics* **34(6)**:1–16 DOI 10.1145/2816795.2818013.

Ma Q, Saito S, Yang J, Tang S, Black MJ. 2021. Scale: modeling clothed humans with a surface codec of articulated local elements. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 16077–16088 DOI 10.1109/CVPR46437.2021.01582.

Ma X, Su J, Wang C, Zhu W, Wang Y. 2023. 3D human mesh estimation from virtual markers. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2023-June*, 534–543 DOI 10.1109/CVPR52729.2023.00059.

Madhu P, Villar-Corrales A, Kosti R, Bendschus T, Reinhardt C, Bell P, Maier A, Christlein V. 2022. Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning. *Journal on Computing and Cultural Heritage* **16(1)**:1–17 DOI 10.1145/3569089.

Manesco JRR, Berretti S, Marana AN. 2023. DUA: a domain-unified approach for cross-dataset 3D human pose estimation. *Sensors* **7312(17)**:1–19 DOI 10.3390/s23177312.

Martini E, Boldo M, Aldegheri S, Valè N, Filippetti M, Smania N, Bertucco M, Picelli A, Bombieri N. 2022. Enabling gait analysis in the telemedicine practice through portable and accurate 3D human pose estimation. *Computer Methods and Programs in Biomedicine* **225(10)**:107016 DOI 10.1016/j.cmpb.2022.107016.

Mehraban S, Adeli V, Taati B. 2023. MotionAGFormer: enhancing 3D human pose estimation with a transformer-GCNFormer network. ArXiv preprint DOI 10.48550/arXiv.2310.16288.

Mehrizi R, Peng X, Xu X, Zhang S, Li K. 2019. A deep neural network-based method for estimation of 3D lifting motions. *Journal of Biomechanics* **84(1–2)**:87–93 DOI 10.1016/j.jbiomech.2018.12.022.

Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W, Theobalt C. 2018. Monocular 3D human pose estimation in the wild using improved CNN supervision. In: *Proceedings—2017 International Conference on 3D Vision, 3DV 2017*, 506–516 DOI 10.1109/3DV.2017.00064.

Mehta D, Sridhar S, Sotnychenko O, Rhodin H, Shafiei M, Seidel HP, Xu W, Casas D, Theobalt C. 2017. VNect: real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics* **36(4)**:1–13 DOI 10.1145/3072959.3073596.

Moher D, Liberati A, Tetzlaff J, Altman DG. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ (Online)* **339**:332–336 DOI 10.1136/bmj.b2535.

Mondal A, Ghosh S, Ghosh A. 2013. Efficient silhouette-based contour tracking. In: *Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2013*, 1781–1786 DOI 10.1109/ICACCI.2013.6637451.

Mori G, Malik J. 2002. Estimating human body configurations using shape context matching_contexts_unknown_Mori, Malik.pdf. In: *European Conference on Computer: Vision*, 1–8.

Munea TL, Jembre YZ, Weldegebriel HT, Chen L, Huang C, Yang C. 2020. The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access* **8**:133330–133348 DOI 10.1109/ACCESS.2020.3010248.

Nam H, Jung DS, Oh Y, Lee KM. 2023. Cyclic test-time adaptation on monocular video for 3D human mesh reconstruction. ArXiv preprint DOI 10.48550/arXiv.2308.06554.

Niu Z, Lu K, Xue J, Wang J. 2024. Skeleton cluster tracking for robust multi-view multi-person 3D human pose estimation. *Computer Vision and Image Understanding* **246(10)**:104059 DOI 10.1016/j.cviu.2024.104059.

Oreshkin BN. 2023. 3D human pose and shape estimation via HybrIK-transformer. 2–3 DOI 10.48550/arXiv.2302.04774.

Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AA, Di T, Black MJ. 2019. Expressive body capture: 3D hands, face, and body from a single image. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June*, 10967–10977 DOI 10.1109/CVPR.2019.01123.

Pavlakos G, Zhou X, Daniilidis K. 2018. Ordinal depth supervision for 3D human pose estimation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7307–7316 DOI 10.1109/CVPR.2018.00763.

Pavlakos G, Zhou X, Derpanis KG, Daniilidis K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In: *Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua*, 1263–1272 DOI 10.1109/CVPR.2017.139.

Qiu Z, Qiu K, Fu J, Fu D. 2023. Weakly supervised pre-training for 3D human pose estimation via perspective knowledge. *Pattern Recognition* **139(6)**:109497 DOI 10.1016/j.patcog.2023.109497.

Rajasegaran J, Pavlakos G, Kanazawa A, Malik J. 2021. Tracking people by predicting 3D appearance, location and pose. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA, 23703–23713 DOI 10.1109/CVPR52688.2022.00276.

Ran H, Ning X, Li W, Hao M, Tiwari P. 2023. 3D human pose and shape estimation via de-occlusion multi-task learning. *Neurocomputing* **548**:126284 DOI 10.1016/j.neucom.2023.126284.

Rapczyński M, Werner P, Handrich S, Al-Hamadi A. 2021. A baseline for cross-database 3d human pose estimation. *Sensors* **21(11)**:3769 DOI 10.3390/s21113769.

Reddy ND, Guigues L, Pishchulin L, Eledath J, Narasimhan SG. 2021. TesseTrack: end-to-end learnable multi-person articulated 3D pose tracking. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 15185–15195 DOI 10.1109/CVPR46437.2021.01494.

Retsinas G, Efthymiou N, Maragos P. 2023. Mushroom segmentation and 3D pose estimation from point clouds using fully convolutional geometric features and implicit pose encoding. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2023-June*, 6264–6271 DOI 10.1109/CVPRW59228.2023.00666.

Saini N, Bonetto E, Price E, Ahmad A, Black MJ. 2022. AirPose: multi-view fusion network for aerial 3D human pose and shape estimation. *IEEE Robotics and Automation Letters* **7(2)**:4805–4812 DOI 10.1109/LRA.2022.3145494.

Saito S, Huang Z, Natsume R, Morishima S, Li H, Kanazawa A. 2019. PIFu: pixel-aligned implicit function for high-resolution clothed human digitization. In: *Proceedings of the IEEE International Conference on Computer Vision 2019-Octob*, 2304–2314 DOI 10.1109/ICCV.2019.00239.

Šajina R, Ivašić-Kos M. 2022. 3D pose estimation and tracking in handball actions using a monocular camera. *Journal of Imaging* **8(11)**:308 DOI 10.3390/jimaging8110308.

Salisu S, Mohamed SAA, Jaafar HM, Pauzi SBA, Younis HA. 2023. A survey on deep learning-based 2D human pose estimation models. *Computers, Materials & Continua* **76(2)**:2385–2400 DOI 10.32604/cmc.2023.035904.

Shan W, Liu Z, Zhang X, Wang Z, Han K, Wang S, Ma S, Gao W. 2023. Diffusion-based 3D human pose estimation with multi-hypothesis aggregation. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France, 14715–14725 DOI 10.1109/ICCV51070.2023.01356.

Shao R, Zhang H, Zhang H, Chen M, Cao YP, Yu T, Liu Y. 2022. DoubleField: bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2022-June*, 15851–15861 DOI 10.1109/CVPR52688.2022.01541.

Sharma S, Varigonda PT, Bindal P, Sharma A, Jain A. 2019. Monocular 3D human pose estimation by generation and ordinal ranking. In: *Proceedings of the IEEE International Conference on Computer Vision 2019-Octob*, 2325–2334 DOI 10.1109/ICCV.2019.00241.

Silva D, Lima J, Thomas D, Uchiyama H, Teichrieb V. 2023. UMVpose++: unsupervised multi-view multi-person 3D pose estimation using ground point matching. In: *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISIGRAPP-4*. Lisbon, Portugal: ScitePress, 607–614 DOI 10.5220/0011668800003417.

Sun Y, Bao Q, Liu W, Mei T, Black MJ. 2023. TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada, 8856–8866 DOI 10.1109/CVPR52729.2023.00855.

Sun J, Wang M, Zhao X, Zhang D. 2020. Multi-view pose generator based on deep learning for monocular 3D human pose estimation. *Symmetry* **12(7)**:1–14 DOI 10.3390/sym12071116.

Terreran M, Barcellona L, Ghidoni S. 2023. A general skeleton-based action and gesture recognition framework for human-robot collaboration. *Robotics and Autonomous Systems* **170(4)**:104523 DOI 10.1016/j.robot.2023.104523.

Tian Y, Zhang H, Liu Y, Wang L. 2023. Recovering 3D human mesh from monocular images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45(12)**:15406–15425 DOI 10.1109/TPAMI.2023.3298850.

Toshpulatov M, Lee W, Lee S, Haghighian Roudsari A. 2022. Human pose, hand and mesh estimation using deep learning: a survey. *Journal of Supercomputing* **78(6)**:7616–7654 DOI 10.1007/s11227-021-04184-7.

Tralic D, Zupancic I, Grgic S, Grgic M. 2013. CoMoFoD—new database for copy-move forgery detection. In: *Proceedings Elmar—International Symposium Electronics in Marine*, 49–54.

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

Ulku I, Akagündüz E. 2022. A survey on deep learning-based architectures for semantic segmentation on 2D images. *Applied Artificial Intelligence* **36(1)**:2032924 DOI 10.1080/08839514.2022.2032924.

Vukicevic AM, Macuzic I, Mijailovic N, Peulic A, Radovic M. 2021. Assessment of the handcart pushing and pulling safety by using deep learning 3D pose estimation and IoT force sensors. *Expert Systems with Applications* **183(1)**:115371 DOI 10.1016/j.eswa.2021.115371.

Wandt B, Rudolph M, Zell P, Rhodin H, Rosenhahn B. 2021. Canonpose: self-supervised monocular 3D human pose estimation in the wild. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 13289–13299 DOI 10.1109/CVPR46437.2021.01309.

Wang M, Chen Q, Fu Z. 2022. LSNet: learned sampling network for 3D object detection from point clouds. *Remote Sensing* **14(7)**:1–22 DOI 10.3390/rs14071539.

Wang X, Prévost S, Boukhayma A, Desjardin E, Loscos C, Morisset B, Multon F. 2023. Evaluation of hybrid deep learning and optimization method for 3D human pose and shape reconstruction in simulated depth images. *Computers and Graphics (Pergamon)* **115(6)**:158–166 DOI 10.1016/j.cag.2023.07.005.

Wang J, Qiu K, Peng H, Fu J, Zhu J. 2019. AI coach: deep human pose estimation and analysis for personalized athletic training assistance. In: *Proceedings MM'19 of the 27th ACM International Conference on Multimedia. Association for Computing Machinery, Nice, France,* 374–382 DOI 10.1145/3343031.3350910.

Wang J, Tan S, Zhen X, Xu S, Zheng F, He Z, Shao L. 2021. Deep 3D human pose estimation: a review. *Computer Vision and Image Understanding* **210(2)**:103225 DOI 10.1016/j.cviu.2021.103225.

Wang K, Xie J, Zhang G, Liu L, Yang J. 2020. Sequential 3D human pose and shape estimation from point clouds. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7273–7282 DOI 10.1109/CVPR42600.2020.00730.

Wang T, Zhang J, Cai Y, Yan S, Feng J. 2021b. Direct multi-view multi-person 3D pose estimation. *Advances in Neural Information Processing Systems* **16**:13153–13164 DOI 10.48550/arXiv.2111.04076.

Weiming Chen ZJ, Guo H, Ni X. 2020. Fall detection based on key points of human-skeleton using OpenPose weiming. *Symmetry* **12(5)**:744 DOI 10.3390/sym12050744.

Weng CY, Curless B, Kemelmacher-Shlizerman I. 2019. Photo wake-up: 3D character animation from a single photo. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June*, 5901–5910 DOI 10.1109/CVPR.2019.00606.

Wu X, Wang Y, Chen L, Zhang L, Wang L. 2022. Motion parameters measurement of user-defined key points using 3D pose estimation. *Engineering Applications of Artificial Intelligence* **110(1)**:104667 DOI 10.1016/j.engappai.2022.104667.

Xi X, Zhang C, Jia W, Jiang R. 2024. Enhancing human pose estimation in sports training: Integrating spatiotemporal transformer for improved accuracy and real-time performance. *Alexandria Engineering Journal* **109(4)**:144–156 DOI 10.1016/j.aej.2024.08.072.

Xiao Y, Wang X, He M, Jin L, Song M, Zhao J. 2023. A compact and powerful single-stage network for multi-person pose estimation †. *Electronics (Switzerland)* **12(4)**:1–20 DOI 10.3390/electronics12040857.

Xie C, Zhang D, Wu Z, Yu C, Hu Y, Chen Y. 2023. RPM 2.0: RF-based pose machines for multi-person 3D pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology* **34(1)**:490–503 DOI 10.1109/TCSVT.2023.3287329.

Xie C, Zhang D, Wu Z, Yu C, Hu Y, Sun Q, Chen Y. 2022. Accurate human pose estimation using RF signals. In: *2022 IEEE 24th International Workshop on Multimedia Signal Processing, MMSP 2022*, 1–6 DOI 10.1109/MMSP55362.2022.9948797.

Xing Y. 2023. HDG-ODE: a hierarchical continuous-time model for human pose forecasting. *Iccv* **33**:14700–14712 DOI 10.1109/ICCV51070.2023.01351.

Xiong F, Zhang B, Xiao Y, Cao Z, Yu T, Zhou JT, Yuan J. 2019. A2J: anchor-to-joint regression network for 3D articulated pose estimation from a single depth image. In: *Proceedings of the IEEE International Conference on Computer Vision 2019-Octob*, 793–802 DOI 10.1109/ICCV.2019.00088.

Xiu Y, Yang J, Tzionas D, Black MJ. 2022. ICON: implicit clothed humans obtained from normals. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2022-June*, 13286–13296 DOI 10.1109/CVPR52688.2022.01294.

Xu H, Bazavan EG, Zanfir A, Freeman WT, Sukthankar R, Sminchisescu C. 2020. GHUM GHUML: generative 3D human shape and articulated pose models. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6183–6192 DOI 10.1109/CVPR42600.2020.00622.

Xu X, Chen H, Moreno-Noguer F, Jeni LA, De La Torre F. 2022. 3D human pose, shape and texture from low-resolution images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**:4490–4504 DOI 10.1109/TPAMI.2021.3070002.

Xu Y, Kitani K. 2022. Multi-view multi-person 3D pose estimation with uncalibrated camera networks. 1–15 DOI 10.48550/arXiv.2312.01561.

Xul C, Kang Y, Tsujino K, Lu C. 2013. A head posture estimation method based on 3-D image measurement for intuitive human-system interaction. In: *2013 International Joint Conference on Awareness Science and Technology and Ubi-Media Computing: Can We Realize Awareness via Ubi-Media?, iCAST 2013 and UMEDIA 2013*, 377–382 DOI 10.1109/ICAwST.2013.6765469.

Yan G, Yan H, Yao Z, Lin Z, Wang G, Liu C, Yang X. 2024. Monocular 3D multi-person pose estimation for on-site joint flexion assessment: a case of extreme knee flexion detection. *Sensors* **24(19)**:6187 DOI 10.3390/s24196187.

Yang J, Ma Y, Zuo X, Wang S, Gong M, Cheng L. 2022. 3D pose estimation and future motion prediction from 2D images. *Pattern Recognition* **124(6)**:108439 DOI 10.1016/j.patcog.2021.108439.

Yin H, Lv C, Shao Y. 2023. 3D human pose estimation based on transformer. *Journal of Physics: Conference Series* **2562(1)**:1–7 DOI 10.1088/1742-6596/2562/1/012067.

Ying J, Zhao X. 2021. Rgb-D fusion for point-cloud-based 3D human pose estimation. In: *Proceedings—International Conference on Image Processing, ICIP 2021-Septe*, 3109–3112 DOI 10.1109/ICIP42928.2021.9506588.

Yu BXB, Zhang Z, Liu Y, Zhong S, Liu Y, Chen CW. 2023. GLA-GCN: global-local adaptive graph convolutional network for 3D human pose estimation from monocular video. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France, 8784–8795 DOI 10.1109/ICCV51070.2023.00810.

Yu T, Zheng Z, Guo K, Zhao J, Dai Q, Li H, Pons-Moll G, Liu Y. 2018. DoubleFusion: real-time capture of human performances with inner body shapes from a single depth sensor. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7287–7296 DOI 10.1109/CVPR.2018.00761.

Yu T, Zheng Z, Zhong Y, Zhao J, Dai Q, Pons-Moll G, Liu Y. 2019. Simulcap: single-view human performance capture with cloth simulation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June*, 5499–5509 DOI 10.1109/CVPR.2019.00565.

Zecha D, Einfalt M, Eggert C, Lienhart R. 2018. Kinematic pose rectification for performance analysis and retrieval in sports. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2018-June*, 1872–1880 DOI 10.1109/CVPRW.2018.00232.

Zhang J, Gong K, Wang X, Feng J. 2023a. Learning to augment poses for 3D human pose estimation in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45(8)**:10012–10026 DOI 10.1109/TPAMI.2023.3243400.

Zhang J, Shi Y, Ma Y, Xu L, Yu J, Wang J. 2023b. IKOL: inverse kinematics optimization layer for 3D human pose and shape estimation via gauss-newton differentiation. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI* **37(3)**:3454–3462 DOI 10.1609/aaai.v37i3.25454.

Zhang Z, Wang C, Qin W, Zeng W. 2020. Fusing wearable IMUs with multi-view images for human pose estimation: a geometric approach. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2197–2206 DOI 10.1109/CVPR42600.2020.00227.

Zhang D, Wu Y, Guo M, Chen Y. 2021. Deep learning methods for 3D human pose estimation under different supervision paradigms: a survey. *Electronics (Switzerland)* **10(18)**:1–25 DOI 10.3390/electronics10182267.

Zhang X, Zhou Z, Han Y, Meng H, Yang M, Rajasegarar S. 2023c. Deep learning-based real-time 3D human pose estimation. *Engineering Applications of Artificial Intelligence* **119(1)**:105813 DOI 10.1016/j.engappai.2022.105813.

Zhao M, Liu Y, Raghu A, Zhao H, Li T, Torralba A, Di K. 2019. Through-wall human mesh recovery using radio signals. In: *Proceedings of the IEEE International Conference on Computer Vision 2019-Octob*, 10112–10121 DOI 10.1109/ICCV.2019.01021.

Zhao M, Tian Y, Zhao H, Alsheikh MA, Li T, Hristov R, Kabelac Z, Katabi D, Torralba A. 2018. RF-based 3D skeletons. In: *SIGCOMM, 2018—Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, 267–281 DOI 10.1145/3230543.3230579.

Zhao C, Uchitomi H, Ogata T, Ming X, Miyake Y. 2023a. Reducing the device complexity for 3D human pose estimation: a deep learning approach using monocular camera and IMUs. *Engineering Applications of Artificial Intelligence* **124(07)**:106639 DOI 10.1016/j.engappai.2023.106639.

Zhao Q, Zheng C, Liu M, Wang P, Chen C. 2023b. PoseFormerV2: exploring frequency domain for efficient and robust 3D human pose estimation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2023-June*, 8877–8886 DOI 10.1109/CVPR52729.2023.00857.

Zhou K, Bhatnagar BL, Pons-Moll G. 2020. Unsupervised shape and pose disentanglement for 3D meshes. In: Vedaldi A, Bischof H, Brox T, Frahm JM, eds. *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science.* Vol. 12367. Cham: Springer, 341–357 DOI 10.1007/978-3-030-58542-6_21.

Zhou Y, Dong H, EI Saddik A. 2020. Learning to estimate 3D human pose from point cloud. *IEEE Sensors Journal* **20(20)**:12334–12342 DOI 10.1109/JSEN.2020.2999849.

Zhu X, Boukhennoufa I, Liew B, Gao C, Yu W, McDonald-Maier KD, Zhai X. 2023. Monocular 3D human pose markerless systems for gait assessment. *Bioengineering* **10(6)**:1–16 DOI 10.3390/bioengineering10060653.

Zhu M, Derpanis KG, Yang Y, Brahmbhatt S, Zhang M, Phillips C, Lecce M, Daniilidis K. 2014. Single image 3D object detection and pose estimation for grasping. In: *Proceedings—IEEE International Conference on Robotics and Automation*, 3936–3943 DOI 10.1109/ICRA.2014.6907430.

Zhu W, Ma X, Liu Z, Liu L, Wu W, Wang Y. 2022. MotionBERT: a unified perspective on learning human motion representations. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France, 15039–15053 DOI 10.1109/ICCV51070.2023.01385.

Zou S, Guo C, Zuo X, Wang S, Wang P, Hu X, Chen S, Gong M, Cheng L. 2021. EventHPE: event-based 3D human pose and shape estimation. In: *Proceedings of the IEEE International Conference on Computer Vision*, 10976–10985 DOI 10.1109/ICCV48922.2021.01081.

Zou S, Xu Y, Li C, Ma L, Cheng L, Vo M. 2023. Snipper: a spatiotemporal transformer for simultaneous multi-person 3D pose estimation tracking and forecasting on a video snippet. *IEEE Transactions on Circuits and Systems for Video Technology* **33(9)**:4921–4933 DOI 10.1109/TCSVT.2023.3244152.