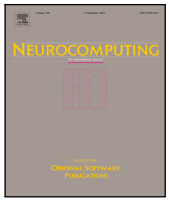
Neurocomputing 596 (2024) 128049

Contents lists available at ScienceDirect

Neurocomputing

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



Survey Paper

Deep learning for 3D human pose estimation and mesh recovery: A survey

Yang Liu, Changzhen Qiu, Zhiyong Zhang *

School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, Guangdong, China

ARTICLE INFO

Communicated by J. Liu

Keywords:
Human pose estimation
3D human pose
Human mesh recovery
Human reconstruction
Deep learning
Literature survey

ABSTRACT

3D human pose estimation and mesh recovery have attracted widespread research interest in many areas, such as computer vision, autonomous driving, and robotics. Deep learning on 3D human pose estimation and mesh recovery has recently thrived, with numerous methods proposed to address different problems in this area. In this paper, to stimulate future research, we present a comprehensive review of recent progress over the past five years in deep learning methods for this area by delving into over 200 references. To the best of our knowledge, this survey is arguably the first to comprehensively cover deep learning methods for 3D human pose estimation, including both single-person and multi-person approaches, as well as human mesh recovery, encompassing methods based on explicit models and implicit representations. We also present comparative results on several publicly available datasets, together with insightful observations and inspiring future research directions. A regularly updated project page can be found at https://github.com/liuyangme/SOTA-3DHPE-HMR.

1. Introduction

1.1. Motivation

Humans are the foremost actors in various social activities, and it is imperative to equip AI with an understanding of humans for contributing to society. Consequently, there are many human-centric tasks positioning the epicenter of research. Among them, 3D **Human Pose Estimation (HPE)** and **Human Mesh Recovery (HMR)** represent crucial tasks in the field of computer vision to interpret the status and behavior of humans in complex real-world environments.

3D human pose estimation can accurately predict human body keypoint coordinates in three-dimensional space. This approach provides more comprehensive and accurate spatial information when compared to its 2D counterpart, thereby facilitating a better understanding of complex human behaviors in higher-level computer vision applications [1–4]. Human mesh recovery reconstructs a three-dimensional digital model of the body, which captures details such as shape [5,6], gestures [7], clothing [8,9], and facial expressions [10], thus offering direct insights into human interactions with the physical world. 3D pose estimation and mesh recovery have a broad range of applications, such as security and surveillance [11], human–computer interaction [12,13], autonomous driving [14,15], and virtual reality [16].

With advanced deep learning technology, 3D human pose estimation and mesh recovery have garnered increasing attention in recent years. 3D pose estimation has evolved from concentrating on single individuals to encompassing multiple persons with more varied data inputs. In human mesh recovery, advancements have been made in terms of data inputs and in capturing more intricate details. The augmentation in explicit model parameters allows for a more nuanced representation of the human body, and the expansion in parameter types facilitates the portrayal of finer surfaces. With the progress of implicit rendering and its incorporation into human mesh recovery, more flexible body representations are achieved. However, both 3D pose estimation and mesh recovery face significant challenges, such as multi-person scenarios, self-occlusion issues, and the detailed reconstruction of bodies. Thus, a systematic and comprehensive review of recent advancements in 3D human pose estimation and mesh recovery is essential.

1.2. Scope of this survey

Over the past five years, numerous reviews have been on 3D human pose estimation and mesh recovery. The reviews [17,18] primarily concentrate on 2D and 3D pose estimation, yet they devote less attention to HMR-related literature. The survey [19] provides a thorough review exclusively dedicated to mesh recovery, focusing on methods based on explicit models. The survey [20] is distinctly centered on mesh recovery using implicit rendering techniques; however, it falls short in offering an exhaustive overview of the latest implicit rendering methods and systems for 3D pose estimation and mesh recovery. These reviews only marginally explore 3D pose estimation and mesh recovery technologies

* Corresponding author.
E-mail address: zhangzhy99@mail.sysu.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.neucom.2024.128049
Received 19 March 2024; Received in revised form 21 May 2024; Accepted 7 June 2024
Available online 10 June 2024

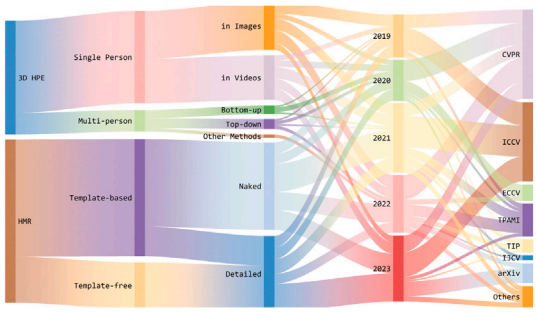


Fig. 1. Recent research of deep learning for 3D HPE and HMR.

but scarcely delve into their applications in other computer vision tasks and future challenges.

This review primarily concentrates on deep learning approaches to 3D human pose estimation and human mesh recovery. 3D pose estimation, both in single-person and multi-person scenarios, is considered. In human mesh recovery, it methodically reviews techniques grounded in both explicit and implicit models. As illustrated in Fig. 1, our survey comprehensively includes the most recent state-of-the-art publications (2019–2023) from mainstream computer vision conferences and journals.

The main contributions of this work compared to the existing literature can be summarized as follows:

(1) To the best of our knowledge, this survey is arguably the first to comprehensively cover deep learning methods for 3D human pose estimation, including both single-person and multi-person approaches, as well as human mesh recovery, encompassing methods based on explicit models and implicit representations.

(2) Unlike existing reviews, we have not overlooked the role of implicit representations in the methods, particularly with the recent rapid advances in implicit rendering. This approach can produce detailed outputs, including clothed human figures with expressions, movements, and other intricacies essential for achieving photorealism.

(3) This paper comprehensively reviews the most recent developments in deep learning for 3D pose estimation and mesh recovery, providing readers with a detailed overview of cutting-edge methodologies. Additionally, it explores how these advancements contribute to various other computer vision tasks and delves into the challenges within this domain.

The structure of this paper is as follows: Section 2 introduces the sensor type and representation for the human body that have been widely used. Section 3 presents the overview of deep learning for 3D human pose estimation and mesh recovery. Section 4 surveys existing single person and multi-person 3D pose estimation methods. Section 5 reviews the methods for human mesh recovery, including template-based and template-free methods. Section 6 introduces the evaluation metrics and datasets for the respective tasks. Moreover, Section 7 discusses the applications, along with their impact on other computer vision tasks. Finally, Section 8 concludes the paper. Fig. 2 shows the taxonomy of deep learning methods for 3D HPE and HMR. Additionally, we host a regularly updated project page, which can be accessed at: https://github.com/liuyangme/SOTA-3DHPE-HMR.

2. Background

2.1. Sensors used for 3D HPE and HMR

There are a variety of sensors that can be used for 3D human pose estimation and mesh recovery, which are mainly categorized as active sensors and passive sensors.

2.1.1. Active sensors

Active sensors operate by emitting a set of signals and measuring by detecting their reflections, such as Motion Capture (MoCap) systems with reflective markers [21], tactile sensing [22], Time of Flight (ToF) cameras [23,24], and Radio Frequency (RF) technologies [25,26]. MoCap and tactile devices are only suitable for cooperative targets. Active cameras generally cannot be used outdoors, and using multiple active devices simultaneously may lead to mutual interference. Thus, these devices have limited application scenarios.

2.1.2. Passive sensors

Passive sensors do not actively emit any signals during the measurement; instead, they rely on signals from the objects or natural sources, including Inertial Measurement Units (IMUs) [27,28] and image sensors [29,30]. Among these, RGB image sensors are particularly notable for being simple, user-friendly, adaptable to various environments, and capable of capturing high-resolution color images. Additionally, multiple RGB image sensors can be combined to form multi-view systems [31–33].

In this survey, considering the widespread applicability of RGB image sensors and the length limitations of the article, we focus on 3D human pose estimation and mesh recovery using RGB image sensors.

2.2. Representation for human body

3D pose estimators and mesh reconstructors can generate corresponding outcomes from the sensor input described previously. The 3D human pose estimation output comprises 3D coordinates detailing positions and joint orientations of the human body, representing the spatial location of each joint (e.g., head, neck, shoulders, elbows, knees), and the skeleton maps the interconnections between joints. Typically, these coordinates are expressed in a global coordinate system or relative to the camera's coordinate system. A keypoints tree structure is commonly employed to illustrate the human pose, where the tree nodes represent the joints and the edges denote the connections between joints.

On the other hand, the skeleton-based human pose representation method does not provide detailed information about the body's surface. The output of human mesh recovery is generally a 3D body model, which encompasses a comprehensive depiction of the body's shape and surface details and offers a more enriched representation. Statistical-based models are widely used in human mesh representations, such as SCAPE [34] and SMPL [35]. SMPL is a learnable skin-vertex model representing the human body as a 3D mesh with a topological structure. The pose and shape of the body are described by pose parameters θ and shape parameters β. The pose parameters control the joint angles and global posture, and the shape parameters determine the body's shape. There are several models based on SMPL to expand the representational capabilities, such as the MANO model (SMPL+H) [36] with hand representation and the FLAME model [37] with facial representation. SMPL-X [38] is a comprehensive model that simultaneously captures the human body, face, and hands by incorporating the FLAME head model [37] and the MANO hand model [36]. H4D [39] can represent the dynamic human shape and pose by building upon the prior knowledge from the SMPL model and extending its capabilities by incorporating a temporal dimension.

In addition to the explicit model representations mentioned above, recent years have seen the development of methods based on implicit models, which offer a more flexible representation of the human body through non-parametric models. The voxel-based model is suitable for volume rendering and physical simulation. However, its resolution is limited by the size of the voxels, which may result in larger storage requirements. Varol et al. [40] proposed an end-to-end network architecture, BodyNet, capable of predicting the volumetric shape of the human body from a single image. Implicit reconstruction transforms inputs into implicit function representations, facilitating the reconstruction of high-quality three-dimensional mesh models from irregular and

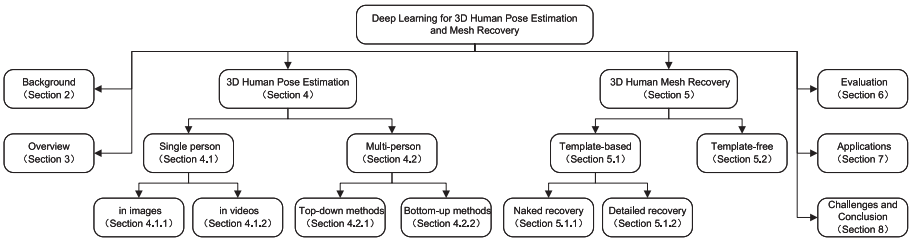


Fig. 2. A taxonomy of deep learning methods for 3D HPE and HMR.

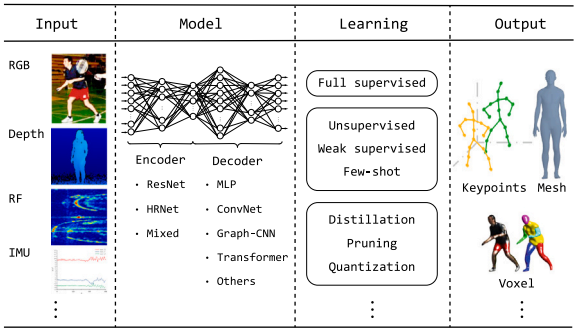


Fig. 3. A basic framework of deep learning for 3D HPE and HMR.

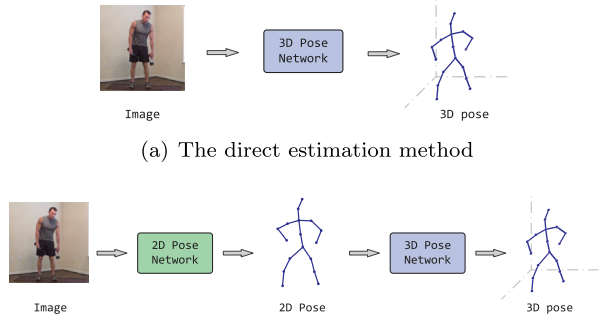


Fig. 4. Typical single person 3D human pose estimation. (a) The direct estimation method; (b) The 2D to 3D lifting method.

noisy data. To address the ill-posed problem of reconstructing a 3D mesh from images, Onizuka et al. proposed the TetraTSDF (Tetrahedral Truncated Signed Distance Function) model [41]. TetraTSDF is a tetrahedral representation model capable of accurately recovering intricate human body shapes, even when the subject is wearing loose clothing. Zheng et al. [42] designed the Parametric Model-Conditioned Implicit Representation (PaMIR), which utilizes the 2D image feature map and corresponding SMPL feature volume to generate an implicit surface representation.

3. Overview of deep learning for 3D HPE and HMR

This section provides an overview of the deep learning framework for 3D human pose estimation and mesh recovery. As shown in Fig. 3, there are four components in deep learning-based systems. First, data are collected from various sensors, including RGB images, depth images, RF signals, IMUs, and so on. Second, the input data are processed through a deep learning model, which typically consists of an encoder and a decoder. The encoder extracts representational features from the input data, such as those obtained using architectures like ResNet [43] or HRNet [44]. The decoder, which could be based on frameworks like MLP (Multi-Layer Perceptron) [45] or Transformer [46], then outputs the body's 3D pose and reconstruction model. This structure enables the model to efficiently process and translate complex input data into detailed and accurate 3D representations of human pose and mesh. Thirdly, various learning methods can be selected during the model learning process. In addition to fully supervised learning, to alleviate data dependency approaches such as weakly supervised learning [47,48], unsupervised learning [49,50], and few-shot learning [51,52] are employed. To reduce the model size, techniques such as knowledge distillation [53,54], model pruning [55], and parameter quantization [56] can be applied. Furthermore, methodologies such as meta-learning [57] and reinforcement learning [58] can also be incorporated, allowing the model to adapt to different scenarios and data constraints. Finally, the deep learning model outputs the results of 3D human pose estimation and mesh recovery. These results can be represented in various forms, including keypoints [59,60], mesh [35,36,61], and voxels [9,41,62]. These representations contribute to a comprehensive understanding of the human body in three dimensions. In the following sections, we will delve into the specifics of 3D human pose estimation and mesh recovery, categorizing and elaborating on each aspect.

4. 3D human pose estimation

3D human pose estimation can provide a more accurate pose by predicting the depth information of body keypoints, but it is much more challenging than 2D pose estimation. 3D pose estimation can be classified into single-person and multi-person estimation, according to the number of targets. Single-person 3D pose estimation has seen rapid progress due to the fast development of deep learning technology, but it still faces many challenges, such as efficiency and the invisibility of certain body parts. Multi-person estimation in crowded scenes is even more challenging because of the interaction and occlusion between bodies and scenes. In this section, we will detail the research progress in this field, categorizing it into two types. Table 1 summarizes all the representative methods.

4.1. Single person 3D pose estimation

As illustrated in Fig. 4, single person 3D pose estimation is mainly classified into the direct estimation method and the 2D to 3D lifting method. The direct method estimates the 3D human pose directly from the input using a predictor, and the 2D to 3D lifting method, which estimates the 3D pose from the results of 2D estimation, involves first detecting the coordinates of human keypoints in 2D space, and then lifting these 2D keypoints onto 3D space coordinates.

4.1.1. Single person 3D pose estimation in images

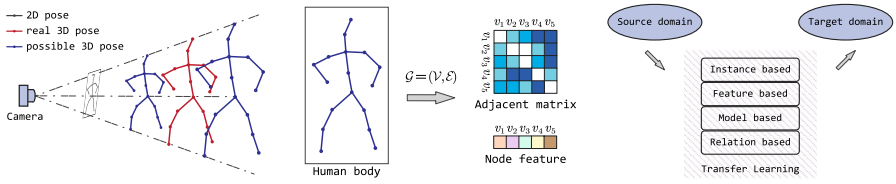
Solving Depth Ambiguity. As shown in Fig. 5(a), different 3D pose coordinates projecting to 2D images may give the same results, leading to an ill-posed problem. This problem can be solved by using the propagation properties of light and camera imaging principle. To address the ill-posed problem, Wei et al. [63] introduced a view-invariant framework to moderate the effects of viewpoint diversity. A

Table 1
Overview of 3D human pose estimation.

Motivations			Methods
Single Person	in Images	Solving Depth Ambiguity	• Optical-aware: VI-HC [63], Ray3D [64] • Appropriate feature representation: HEMlets [65]
		Solving Body Structure Understanding	• Joint aware: JRAN [66] • Limb aware: Wu et al. [67], Deep grammar network [68] • Orientation keypoints: Fisch et al. [69] • Graph-based: Liu et al. [70], LCN [71], Modulated-GCN [72], Skeletal-GNN [73], HopFIR [74], RS-Net [59]
		Solving Occlusion Problems	• Learnable-triangulation [75] • RPSM [76] • Lightweight multi-view [77] • AdaFuse [78] • Bartol et al. [79] • 3D pose consensus [80] • Probabilistic triangulation module [31]
		Solving Data Lacking	• Unsupervised learning: Kudo et al. [81], Chen et al. [82], ElePose [83] • Self-supervised learning: EpipolarPose [84], Wang et al. [85], MRP-Net [86], PoseTriplet [58] • Weakly-supervised learning: Hua et al. [87], CameraPose [47] • Transfer learning: Adaptpose [88]
	in Videos	Solving Single-frame Limitation	• VideoPose3D [89] • PoseFormer [90] • UniPose+ [91] • MHFormer [92] • MixSTE [93] • Honari et al. [94] • HSTFormer [95] • STCFormer [96]
		Solving Real-time Problems	• Temporally sparse sampling: Einfalt et al. [21] • Spatio-temporal sparse sampling: MixSynthFormer [97]
		Solving Body Structure Understanding	• Motion loss: Wang et al. [98] • Human-joint affinity: DG-Net [99] • Anatomy-aware: Chen et al. [100] • Part aware attention: Xue et al. [101]
		Solving Occlusion Problems	• Optical-flow consistency constraint: Cheng et al. [102] • Multi-view: MTF-Transformer [32]
		Solving Data Lacking	• Unsupervised learning: Yu et al. [103] • Weakly-supervised learning: Chen et al. [104] • Semi-supervised learning: MCSS [105] • Self-supervised learning: Kundu et al. [106], P-STMO [107] • Meta-learning: Cho et al. [57] • Data augmentation: PoseAug [108], Zhang et al. [109]
Multi-person	Top-down	Solving Real-time Problems	• Multi-view: Chen et al. [110] • Whole body: AlphaPose [111]
		Solving Representation Limitation	• VoxelTrack [2]
		Solving Occlusion Problems	• Wu et al. [112]
		Solving Data Lacking	• Single-shot: PandaNet [52] • Optical-aware: Moon et al. [113]
	Bottom-up	Solving Real-time Problems	• Fabbri et al. [114]
		Solving Supervisory Limitation	• HMOR [115]
		Solving Data Lacking	• Single-shot: SMAP [116], Benzine et al. [117]
		Solving Occlusion Problems	• Mehta et al. [118] • LCR-Net++ [119]
	Others	Single Stage	• Jin et al. [120]
		Top-down + Bottom-up	• Cheng et al. [121]

View-Invariant Hierarchical Correction (VI-HC) network predicts the 3D pose refinement with view-consistent constraints in the proposed framework. Additionally, a view-invariant discriminative network actualizes high-level constraints after the base network generates an initial estimation. Ray-based 3D (Ray3D) absolute estimation method [64] can convert the input from pixel space to 3D normalized rays. Another helpful approach for this problem is learning a more appropriate feature representation. Part-centric HEatMap triplets (HEMlets) framework [65] utilizes three joint-heatmaps to represent the end-joints relative depth information, which bridges the gap between the 2D location and the 3D human pose.

Solving Body Structure Understanding. Unlike other computer vision tasks, the body's unique structures can provide constraints or prior information to improve pose estimation performance. In the Joint Relationship Aware Network (JRAN) [66], a dual attention module was designed to generate both whole and local feature attention block weights. The limb poses aware network [67] leverages kinematic constraint and trajectory information to prevent errors from accumulating along the human body structure. Pose grammar in [68] learns a 2D-3D mapping function, enabling the input 2D pose to be transformed into 3D space, and integrates three aspects of human structure (kinematics, symmetry, motor coordination) into a Bi-directional Recurrent Neural



(a) Depth ambiguity (b) Graph-based representation (c) Transfer learning

Fig. 5. (a) Depth ambiguity; (b) Graph-based representation for human body; (c) Transfer learning.

Network (RNN). The orientation keypoints-based method [69] uses virtual markers to generate sufficient information for accurately inferring rotations through simple post-processing.

The Graph Neural Network (GNN) is a convolutional network defined on the graph data structure. It is difficult for typical neural networks to handle graph-structured data, such as body structure information. However, this challenge can be more easily addressed if graph-based learning methods are used, as shown in Fig. 5(b). Liu et al. [70] proposed a feature boosting network in which features are learned by the convolutional layers and are boosted with Graphical ConvLSTM to perceive the graphical long short-term dependency among different body parts. The Locally Connected Network (LCN) [71] leverages the allocation of dedicated filters rather than sharing them for different joints. Modulated Graph Convolutional Network (Modulated-GCN) [72] includes weight and affinity modulation to learn modulation vectors for different body joints and to adjust the graph structure, respectively. Zeng et al. [73] designed a skeletal GNN learning framework to address the depth ambiguity and self-occlusion problems in 3D human pose estimation. In this framework, the proposed hop-aware hierarchical channel-squeezing fusion layer is designed to extract relevant information from neighboring nodes. Higher-order Regular Splitting graph Network (RS-Net) [59] captures long-range dependencies between body joints using multi-hop neighborhoods. It learns distinct modulation vectors for different body joints and adds modulation matrices to the corresponding skeletal adjacency matrices. Zhai et al. [74] employed intra-group joint refinement utilizing attention mechanisms to discover potential joint synergies to explore the potential synergies between joints.

Solving Occlusion Problems. Partial occlusion of the human body, including self-occlusion and other occlusions (such as object occlusion and multi-person occlusion), is typical in various scenes. Occlusion may interfere with pose estimation, potentially resulting in the prediction of an error pose. Solving the occlusion problem through a multi-view approach is effective, where an occluded pose not visible in one view is likely to be visible in other views in a multi-view system. Thus, a multi-view based approach can provide more reliable results through cross-view frame inference. In practice, Iskakov et al. [75] combined algebraic and volumetric triangulation for 3D pose estimation from multi-view 2D images. The former method is based on a differentiable algebraic triangulation with confidence weights, while the latter method utilizes volumetric aggregation from intermediate 2D backbone feature maps. Multi-view geometric priors have been incorporated in [76], which combines two steps: first, predicting the 2D poses in multi-view RGB images through cross-view fusion, and second, utilizing the proposed Recursive Pictorial Structure Model (RPSM) to recover the 3D poses from the previously predicted multi-view 2D poses. To improve the real-time performance of multi-view 3D pose estimation, Remelli et al. [77] designed a lightweight framework with a differentiable Direct Linear Transform (DLT) layer. Adaptive multi-view fusion [78] enhances the features in occluded views by determining the correspondence between points in occluded and visible views. To address the problem of generalizability for multi-view estimation, Bartol et al. [79] proposed a stochastic learning framework for pose triangulation. Luvizon et al. [80] effectively integrated 2D annotated data and 3D poses to design a consensus-aware method, which optimizes multi-view poses from uncalibrated images by different coherent estimations up to a scale factor from the intrinsic parameters. Probabilistic triangulation module [31] embedded in a 3D pose estimation framework can extend multi-view methods to uncalibrated scenes. It models camera poses using probability distributions and iteratively updates them from 2D features, replacing the traditional update through camera poses and eliminating the dependency on camera calibration.

Solving Data Lacking. In recent years, the development of diverse and efficient feature representations, along with end-to-end training modes, has significantly enhanced the accuracy of deep learning models in 3D human pose estimation. However, a significant limitation of these models is their reliance on fully supervised training, necessitating vast amounts of expensive and labor-intensive labeled 3D data, predominantly from indoor scenes. Addressing this challenge requires exploring alternative training strategies beyond fully supervised learning, such as unsupervised, semi-supervised, and few-shot learning approaches.

Unsupervised learning typically focuses on learning features rather than specific tasks from unlabeled training data, and it finds relationships between samples by mining the intrinsic features of the data. The first work [81] can predict 3D human pose without any 3D dataset by adversarial learning, which is based on the generative adversarial networks via unsupervised training. The approach [82] utilizes geometric self-supervision and randomly reprojects 2D pose camera viewpoints from the recovered 3D skeleton during training, thereby forming a self-consistency loss in the lift-reproject-lift process. Elepose [83] utilizes random projections, estimating likelihood using normalizing flows on 2D poses in a linear subspace. Furthermore, Elepose also learns the distribution of camera angles to reduce the dependence on camera rotation priors in the training dataset.

Self-supervised learning is a branch of unsupervised learning where the model can learn by itself from unlabeled training data and acquire the representation model on unlabeled data through pretext tasks. During on-the-ground execution, EpipolarPose [84] trains the 3D pose estimator without any 3D ground-truth data or camera extrinsic parameters, predicting 2D poses from multi-view images and obtaining a 3D pose and camera geometry via epipolar geometry. Wang et al. [85] designed a simple yet effective self-supervised correction mechanism for 3D human pose estimation by learning all intrinsic structures of the human pose, which is divided into two parts: the 2D-to-3D pose transformation and 3D-to-2D pose projection. MRP-Net [86] reformulated the self-supervised 3D human pose estimation as an unsupervised domain adaptation problem, which includes model-free joint localization and model-based parametric regression. To reduce dependence on the consistency loss that guides learning, unlike previous works, the method [58] generates 2D-3D pose pairs for augmenting supervision via the proposed self-enhancing dual-loop learning framework.

A weakly-supervised model, characterized by its reliance on only weak labels, often faces the challenge of performing complex tasks with limited or imprecise guidance. Despite the lack of detailed annotations typically required in fully supervised scenarios, these models derive meaningful insights, utilizing vague or less informative labels to infer intricate patterns and relationships within the data. This ability to operate with minimal supervision makes them particularly useful when

acquiring comprehensive labeled data is impractical or costly. In practical applications, Hua et al. [87] proposed a weakly-supervised method that initially lifts the 2D keypoints into coarse 3D poses across two views using triangulation and then refines the 3D pose by employing spatial configurations and cross-view correlations. Yang et al. [47] designed a weakly-supervised framework that utilizes projection relationships to estimate 3D poses solely from 2D pose annotations under the condition of known camera parameters.

Transfer learning enables the transfer of model parameters from the source domain to the target domain, allowing us to share learned model parameters with a new model, as shown in Fig. 5(c). This approach speeds up and optimizes efficiency, eliminating the need to learn from scratch, as is common with most networks. For example, Adaptpose [88] is an end-to-end cross-dataset adaptation 3D human pose estimation predictor that is implemented using transfer learning for limited data applications.

4.1.2. Single person 3D pose estimation in videos

With hardware development, image data are often acquired and processed in videos. The video frame data can provide more continuous information than a single-frame image, predicting human pose estimation in sequence through the spatio-temporal domain. In addition, optical flow and scene flow information can be extracted from videos to predict 3D human pose from data of multimodal [60,122].

Solving Single-frame Limitation. Using continuous image frames from video can capture dynamic changes and temporal information. By analyzing this information, deep learning models can extract motion features and spatio-temporal relationships, thereby effectively improving the effectiveness of 3D human pose estimation. Pavllo et al. [89] proposed a 3D human pose estimation method for video that extracts temporal cues with dilated convolutions over 2D keypoint trajectories, and then they designed a semi-supervised method with back-projection to improve performance. PoseFormer [90] as a spatial–temporal transformer-based framework predicts 3D body pose by analyzing human joint relations within each frame and their temporal correlations across frames. Based on previous work [123], unipose+ [91] leverages multi-scale feature representations to enhance the feature extractors in the framework's backbone. This framework utilizes contextual information across scales and joint localization with Gaussian heatmap modulation to improve decoder estimation accuracy.

Furthermore, Multi-Hypothesis transFormer (MHFormer) [92] learns spatio-temporal representations with multiple plausible pose hypotheses in a three-stage framework. In Mixed Spatio-Temporal Encoder (MixSTE) framework [93], the temporal transformer block learns the temporal motion of each joint and their inter-joint spatial correlation. Honari et al. [94] proposed an unsupervised feature extraction method using Contrastive Self-Supervised (CSS) learning to extract temporal information from videos. Extending this line of research, Hierarchical Spatial-Temporaltrans Formers (HSTFormer) [95] utilizes spatial–temporal correlations of joints at different levels simultaneously, marking a first in studying hierarchical transformer encoders with multi-level fusion. Recently, Spatio-Temporal Criss-cross (STC) attention block [96] decomposes correlation learning into space and time to reduce the computational cost of the joint-to-joint affinity matrix.

Solving Real-time Problems. However, while various methods improve 3D pose estimation performance, they also introduce new problems and challenges in video-based estimation. For example, the continuous video frames add a substantial computational cost, which poses a challenge to the processing efficiency of the system and makes real-time processing more difficult. To reduce the total computational complexity, Einfalt et al. [21] designed a transformer-based scheme that uplifts dense 3D poses from temporally sparse 2D pose sequences by temporal upsampling within Transformer blocks. Sun et al. [97] reduced the complexity of attention calculation in Transformers through spatio-temporal sparse sampling, enabling the estimation of 3D human poses in video sequences on computationally constrained platforms.

Solving Body Structure Understanding. At the same time, the continuous motion information provided by video has led researchers to combine videos with human kinematics. Wang et al. [98] designed a motion loss that computes the difference between the motion patterns of the predicted and ground truth keypoint trajectories, along with motion encoding, a simple yet effective representation of keypoint motion for 3D human pose estimation in videos. Dynamical Graph Network (DG-Net) [99] dynamically determines the human-joint affinity and adaptively predicts human pose via spatial–temporal joint relations in videos. To address the shortcoming of directly regressing the 3D joint locations, Chen et al. [100] proposed an anatomy-aware estimation framework. This human skeleton anatomy-based framework includes a bone direction prediction network and a bone length prediction network, and it effectively utilizes bone features to bridge the gap between 2D keypoints and 3D joints. Xue et al. [101] designed a part-aware temporal attention module capable of extracting each part's temporal dependency separately.

Solving Occlusion Problems. To address occlusion problems, video-based 3D human pose estimation enables the use of a multi-view approach and leverages the continuity of inputs between video frames to predict occluded body parts. Cheng et al. [102] introduced an occlusion-aware model that utilizes estimated 2D confidence heatmaps of keypoints and an optical-flow consistency constraint to generate a more complete 3D pose in occlusion scenes. Additionally, Multi-view and Temporal Fusing Transformer (MTF-Transformer) [32] fuses multi-view sequences in uncalibrated scenes for 3D human pose estimation in videos. To reduce dependency on camera calibration, the framework infers the relationship between pairs of views with a relative attention mechanism.

Solving Data Lacking. Human actions are inherently continuous, making video-based pose estimation critical for enhanced understanding. Compared with 3D annotated still images, high-quality, annotated 3D videos are relatively scarce. Nevertheless, the Internet abounds with a vast repository of unlabeled video data, the utilization of which for learning purposes holds considerable importance. Thus, video-based 3D human pose estimation to alleviate data dependency is essential. In practice, Yu et al. [103] divided the unsupervised 3D pose estimation process into two sub-tasks: a scale estimation module and a pose lifting module. The scale estimation module optimizes the 2D input pose, while the pose lifting module maps the optimized 2D pose to its 3D counterpart. In weakly-supervised methods, Chen et al. [104] proposed a weakly-supervised method, employing a view synthesis framework and a geometry-aware representation. In this framework, the method leverages 2D keypoints for supervision and learns a shared 3D representation between viewpoints by synthesizing the human pose from one viewpoint to another. Semi-supervised learning utilizes only partially labeled data, consisting of a large amount of unlabeled data and a small amount of labeled data. In this context, Mitra et al. [105] present a multi-view consistent semi-supervised method, which can regress 3D human pose from unannotated, uncalibrated, but synchronized multi-view videos. In self-supervised methods, Kundu et al. [106] employed prior knowledge of the body skeleton and disentangles the inherent factors of variation through part-guided human image synthesis. Similarly, Shan et al. [107] randomly mask the body joints in spatial and temporal domains to better capture spatial and temporal dependencies. Meta-learning (a.k.a. "learning to learn") is a process where algorithms optimize their learning strategy based on experience. It involves training a model on various learning tasks, enabling it to learn new tasks more efficiently using fewer data samples. Cho et al. [57] present an optimization-based meta-learning algorithm for 3D human pose estimation. This algorithm can adapt to arbitrary camera distortion by generating synthetic distorted data from undistorted 2D keypoints during model training. Apart from changing the learning supervision methods, data augmentation is also a practical approach to

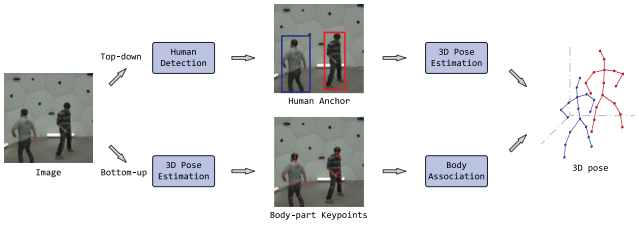


Fig. 6. Typical multi-person 3D pose estimation.

increasing the amount of data and enhancing the model's generalization performance. PoseAug [108] adjusts geometric factors such as posture, body shape, and viewpoint for model learning under the premise of the discriminative module controlling the feasibility of augmented poses. Zhang et al. [109] generates more diverse and challenging poses online.

4.2. Multi-person 3D pose estimation

As depicted in Fig. 6, multi-person 3D pose estimation can be divided into two main categories: the bottom-up method (estimation + association) and the top-down method (detection + estimation). With a two-stage pipeline, the top-down method first detects every person in the input images and then extracts each person's keypoints in the previously detected bounding box. The bottom-up method initially detects all keypoints in one stage and subsequently associates them with their respective persons. For instance, Cheng et al. [121] integrated both top-down and bottom-up approaches in multi-person pose estimation, complementing each other's shortcomings. The top-down method shows robustness against potential erroneous bounding boxes, while the bottom-up network is more robust in handling scale variations. Finally, the 3D poses estimated from both top-down and bottom-up networks are fed into an integrated network to obtain the final 3D pose. Some single-stage methods and approaches combine the two methods as mentioned above. Unlike mainstream two-stage solutions for multi-person 3D human pose estimation, Jin et al. [120] introduced a single-stage method. This method expresses the 2D pose in the image plane and the depth information of a 3D body instance via the proposed decoupled representation and predicts the scale information of instances by extracting 2D pose features and enabling depth regression.

4.2.1. Top-down methods

In top-down methods, AlphaPose [111] predicts whole-body multi-person 3D human pose including face, body, hand, and foot. In the proposed framework, the symmetric integral keypoint regression module achieves fast and fine localization, the parametric pose non-maximum suppression module helps to eliminate redundant human detections, and the pose aware identity embedding module enables joint pose estimation and tracking. To achieve better real-time performance in multi-person 3D pose estimation, Chen et al. [110] proposed a multi-view temporal consistency method in real-time from videos, which matches the 2D inputs with 3D poses directly in three-dimensional space and achieves over 150 FPS (frames per second) on a 12-camera setup and 34 FPS on a 28-camera setup. Additionally, choosing appropriate representations can effectively improve multi-person pose estimation results. For instance, Zhang et al. [2] employed a voxel representation to predict multi-person 3D pose and tracking, where the proposed voxel representation can determine whether each voxel contains a particular body joint. In multi-person scenarios, occlusion issues may be more severe and complex than in single-person scenarios. To address this challenge, Wu et al. [112] utilized graph neural networks to boost information passing efficiency for multi-person, multi-view 3D human pose estimation. The multi-view matching graph module associates coarse cross-view poses within the networks, and the center refinement graph module further refines the results. Single-shot learning trains the model using significantly less data than is typically required for supervised learning, aiming to improve effectiveness. PandaNet [52] as an anchor-based model introduces a pose-aware anchor selection strategy to discard ambiguous anchors. Moon et al. [113] proposed a top-down, camera distance-aware method for multi-person 3D pose estimation without relying on ground-truth information.

4.2.2. Bottom-up methods

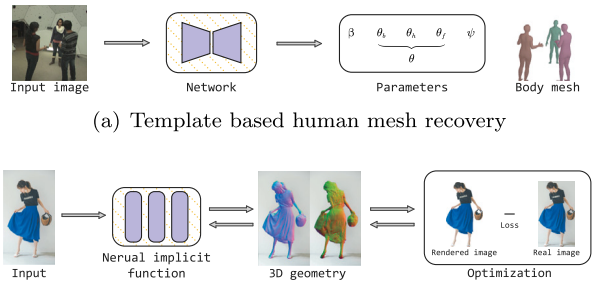
MubyNet [124] is a typical bottom-up, multi-task method for multi-person 3D pose estimation, which allows for training all component parameters. In the research on the lightweight top-down framework, Fabbri et al. [114] utilized high-resolution volumetric heatmaps to improve the estimation performance and designed a volumetric heatmap autoencoder that can compress the size of the representation. Designing better supervisory methods could also be effective, Wang et al. [115] introduced a novel supervisory approach named Hierarchical Multi-person Ordinal Relations (HMOR) using a monocular camera and designed a comprehensive top-level model to learn these ordinal relations, enhancing the accuracy of human depth estimation through a coarse-to-fine architecture. In the top-down estimation on single-shot learning, Mehta et al. [118] inferred whole body pose under strong partial occlusions through occlusion-robust pose-maps in their proposed single-shot framework. Zhen et al. [116] designed a single-shot, bottom-up framework that estimates the absolute positions of multiple people by leveraging depth-related cues across the entire image. After their previous work [52], Benzine et al. [117] proposed a single-shot 3D human pose estimation method that predicts multi-person 3D poses without the need for bounding boxes. This method extends the Stacked Hourglass Network [125] to handle multi-person situations. To address the issue of occlusion, LCR-Net++ [119] integrates adjacent pose hypotheses to predict the multi-person 2D and 3D poses without approximating initial human localization when a person is partially occluded or truncated by the image boundaries.

4.3. Summary of 3D pose estimation

Research in single-person 3D pose estimation predominantly focuses on some of the aforementioned critical issues. While there are commendable studies in this field, the challenges are yet to be fully resolved, necessitating more comprehensive research. In contrast to image-based methods, utilizing video-based methods is unequivocally seen as the trajectory of future advancements. The methodology adopted in multi-person 3D pose estimation must be tailored to the specific context. The current prevalent techniques can be categorized into two main approaches, each with inherent limitations. Top-down methods depend heavily on human detection and are susceptible to detection inaccuracies, often leading to unreliable pose estimations in environments with multiple people. Conversely, bottom-up methods, which operate independently of human detection and thus are immune to errors, face challenges in accurately processing all individuals in a scene concurrently, particularly affecting the detection of smaller-scale figures. In line with other domains in computer vision, the one-stage, end-to-end methods represent this field's future direction.

5. 3D human mesh recovery

Human mesh recovery can be divided into two categories based on their representation models: template-based (parametric) methods and template-free (non-parametric) methods, as shown in Fig. 7. Template-based human mesh recovery reconstructs predefined models (such as SCAPE [34], SMPL [35]) by estimating the model's parameters. In contrast, template-free human mesh recovery predicts the 3D body directly from input data without reliance on predefined models. The



(a) Template based human mesh recovery

(b) Template-free human mesh recovery

Fig. 7. Typical human mesh recovery. (a) Template based human mesh recovery method; (b) Template-free human mesh recovery method.

parametric approaches can be more robust than non-parametric methods with the templates' prior knowledge, but their flexibility and detail are inherently limited even when extended with more parameters like SMPL+H [36] and SMPL+X [61]. Human mesh recovery is a crucial technique for digitizing the human body, posing significant challenges in computer vision and computer graphics due to the complex nature of geometric textures and color variations. Moreover, human mesh recovery faces challenges similar to 3D pose estimation, including environmental interference, multi-person scenarios, and occlusion issues. In this section, we will introduce the dominant methods based on these categories and challenges, with a comprehensive summary provided in Table 2.

5.1. Template-based human mesh recovery

5.1.1. Naked human body recovery

Multimodal Methods. Multimodality in human mesh recovery harnesses the potential by combining various modalities of data, such as RGB images, depth information, and optical flow. Integrating data in different modalities can significantly enhance the robustness and precision of mesh recovery. Rong et al. [126] proposed a hybrid annotation method with a hybrid training strategy for human mesh recovery to reduce annotation costs, which effectively utilizes various types of heterogeneous annotations, including 3D and 2D annotations, body part segmentation, and dense correspondence. Deep Two-Stream Video Inference for Human Body Pose and Shape Estimation (DTS-VIBE) [127] method redefines the task as a multimodal problem, merging RGB data with optical flow to achieve a more reliable estimation and address temporal inconsistencies from RGB videos. LASOR [128] estimates 3D pose and shape by synthesizing occlusion-aware silhouettes and 2D keypoints data in scenes with inter-person occlusion. CLIFF [129] utilizes cropped images and bounding box information from the pre-training phase as inputs to enhance the accuracy of global rotation estimation in the camera coordinate system.

Utilizing Attention Mechanism. Transformer [212] as a self-attention model has demonstrated remarkable success in Natural Language Processing (NLP) [213,214]. Following this, the Vision Transformer (ViT) [46] has mirrored these successes in the field of computer vision [45,215], and numerous Transformer-based methods are now being employed in mesh recovery. The attention mechanism serves to amplify the importance of certain parts in the neural network. To address the partial occlusion issues, Part Attention REgressor (PARE) [130] leverages the relationships between body parts derived from segmentation masks, prompting the network to enhance predictions for occluded body parts. Mesh Graphormer [131] utilizes a GCNN-reinforced transformer for estimating 3D mesh vertices and body joints, while a GCNN is employed to infer interactions among neighboring vertices based on pre-existing mesh topology. The architecture effectively merges graph-based networking with the attention mechanism of transformers, enabling the modeling of local and global interactions. MPS-Net [132] employs motion continuity attention to capture temporal coherence. This approach then leverages a hierarchical attentive feature mechanism to combine features from temporally adjacent representations more effectively. FastMETRO [134] leverages self-attention mechanisms for non-adjacent vertices based on the topology of the body's triangular mesh to minimize memory overhead and accelerate inference speed. Xue et al. [135] proposed a learnable sampling module for human mesh recovery to reduce inherent depth ambiguity. This module aggregates global information by generating joint adaptive tokens, utilizing non-local information within the input image. METRO [136] is a mesh transformer for end-to-end human mesh recovery, in which the encoder captures interactions between vertices and joints, and the decoder outputs 3D joint coordinates and the mesh structure. Qiu et al. [133] proposed an end-to-end Transformer-based method for multi-person human mesh recovery in videos. This approach utilizes a spatio-temporal encoder to extract global features, followed by a spatio-temporal pose and shape decoder to predict human pose and mesh.

Exploiting Temporal Information. With the advancement of video technology, video-based human mesh recovery methods that extract temporal information from adjacent frames have shown increased potential. To learn the 3D dynamics of the human body more accurately and effectively, Kanazawa et al. [137] proposed a framework that produces smooth 3D meshes from videos. This framework also predicts past and future 3D motions by temporally encoding image features. Video Inference for Body pose and shape Estimation (VIBE) [138] estimates kinematically plausible motion sequences through self-attention mechanism-based temporal network and adversarial training without requiring any ground-truth 3D labels. To diminish the dependency on static features in the current frame, addressing a limitation of previous temporal-based methods, Choi et al. [139] developed the Temporally Consistent Mesh Recovery (TCMR) system. The system utilizes temporal information to ensure consistency and effectively recovers smooth 3D human motion by incorporating data from past and future frames. Multi-level Attention Encoder–Decoder (MAED) network [140] captures relationships at multiple levels, including the spatial–temporal and human joint levels, through its multi-level attention mechanism. Wang et al. [141] introduced the Temporally embedded 3D human body Pose and shape estimation (TePose) method, specifically tailored for live stream videos. They designed a motion discriminator for adversarial training, utilizing datasets without any 3D labels, through a multi-scale spatio-temporal graph convolutional network. Additionally, they employed a sequential data loading strategy to accommodate the unique start-to-end data processing requirements of live streaming. To effectively balance the learning of short-term and long-term temporal correlations, Global-to-Local Transformer (GLoT) [142] structurally decouples the modeling of long-term and short-term correlations.

Multi-view Methods. Dong et al. [143] designed a practical multi-view framework that combines 2D observations from multi-view images into a unified 3D representation for individual instances using a confidence-aware majority voting mechanism. Sengupta et al. [144] introduced a probabilistic-based multi-view method without constraints such as specific target poses, viewpoints, or background conditions across image sequences. Huang et al. [33] proposed a dynamic physics-geometry consistency approach for multi-person multi-view mesh recovery. This method integrates motion priors, extrinsic camera parameters, and human mesh data to mitigate the impact of noisy human semantic data. Zhuo et al. [145] proposed a cross-view fusion method that predicts foot posture by achieving a more refined 3D intermediate representation and alleviating inconsistencies across different views.

Boosting Efficiency. Maintaining excellent performance with a lightweight model and low computational cost is essential and challenging, especially in applications such as wearable devices, power-limited systems, and edge computing. SCOPE [146] efficiently computes the Gauss–Newton direction using 2D and 3D keypoints for human mesh

Table 2
Overview of human mesh recovery.

Main ideas			Methods
Template-based	Naked	Multimodal Methods	• Hybrid annotations: Rong et al. [126] • Optical flow: DTS-VIBE [127] • Silhouettes: LASOR [128] • Cropped image and bounding box: CLIFF [129]
		Utilizing Attention Mechanism	• Part-driven attention: PARE [130] • Graph attention: Mesh Graphormer [131] • Spatio-temporal attention: MPS-Net [132], PSVT [133] • Efficient architecture: FastMETRO [134], Xue et al. [135] • End-to-end structure: METRO [136]
		Exploiting Temporal Information	• Temporally encoding features: Kanazawa et al. [137] • Self-attention temporal: VIBE [138] • Temporally consistent: TCMR [139] • Multi-level spatial–temporal attention: MAED [140] • Temporally embedded live stream: TePose [141] • Short-term and long-term temporal correlations: GLoT [142]
		Multi-view Methods	• Confidence-aware majority voting mechanism: Dong et al. [143] • Probabilistic-based multi-view: Sengupta et al. [144] • Dynamic physics-geometry consistency: Huang et al. [33] • Cross-view fusion: Zhuo et al. [145]
		Boosting Efficiency	• Sparse constrained formulation: SCOPE [146] • Single-stage model: BMP [147] • Process heatmap inputs: HeatER [148] • Removing redundant tokens: TORE [149]
		Developing Various Representations	• Texture map: TexturePose [150] • UV map: Zhang et al. [151], DecoMR [152], Zhang et al. [153] • Heat map: Sun et al. [154], 3DCrowdNet [155] • Uniform representation: DSTformer [156]
		Utilizing Structural Information	• Part-based: holopose [157] • Skeleton disentangling: Sun et al. [158] • Hybrid inverse kinematics: HybrIK [159], NIKI [160] • Uncertainty-aware: Lee et al. [161] • Kinematic tree structure: Sengupta et al. [162] • Kinematic chains: SGRE [163]
		Choosing Appropriate Learning Strategies	• Self-improving: SPIN [164], ReFit [165], You et al. [5] • Novel losses: Zanfir et al. [48], Jiang et al. [166] • Unsupervised learning: Madadi et al. [167], Yu et al. [50] • Bilevel online adaptation: Guan et al. [168] • Single-shot: Pose2UV [169] • Contrastive learning: JOTR [170] • Domain adaptation: Nam et al. [171]
	Detailed	With Clothes	• Alldieck et al. [172] • Multi-Garment Network (MGN) [173] • Texture map: Tex2Shape [174] • Layered garment representation: BCNet [175] • Temporal span: H4D [39]
		With Hands	• Linguistic priors: SGNify [176] • Two-hands interaction: [177] • Hand-object interaction: [178]
		Whole Body	• PROX [179] • ExPose [180] • FrankMocap [181] • PIXIE [182] • Moon et al. [183] • PyMAF [184] • OSX [185] • HybrIK-X [186]

(*continued on next page*)

recovery. The method capitalizes on inherent sparsity and employs a sparse constrained formulation, achieving real-time performance at over 30 FPS. To implement a single-stage multi-person human mesh recovery model, Body Meshes as Points (BMP) [147] represents multiple people as points and associates each with a single body mesh to significantly improve efficiency and performance. HeatER [148] processes heatmap inputs directly to reduce memory and computational costs. Dou et al. [149] designed an efficient transformer for human mesh recovery to reduce model complexity and computational cost by removing redundant tokens.

Developing Various Representations. Stable and effective feature representation is crucial for enhancing the capabilities of deep learning algorithms in human mesh recovery, as it enables the efficient extraction of meaningful patterns from complex input data. Based on the assumption that image texture remains constant across frames, TexturePose [150] leverages the appearance constancy of the body across different frames. It measures the texture map for each frame using a texture consistency loss. DenseRaC [216] generates a pixel-to-surface correspondence map to optimize the estimation of parameterized human pose and shape. Zhang et al. [151] established a connection

Table 2 (*continued*).

Main ideas		Methods
Template-free	Regression-based	• FACSIMILE [187], PeeledHuman [188], GTA [189], NSF [190]
	Optimization-based Differentiable	• DiffPhy [191], AG3D [192]
	Implicit Representations	• PIFu [193], PIFuHD [194] • Canonical space: ARCH [195], ARCH++ [196], CAR [197] • Geometric priors: GeoPIFu [198] • Novel representations: Peng et al. [199], 3DNBF [200]
	Neural Radiance Fields	• Volume deformation scheme [201] • ActorsNeRF [51]
	Diffusion Models	• HMDiff [202]
	Implicit + Explicit	• HMD [203], IP-Net [204], PaMIR [42], Zhu et al. [205], ICON [206], ECON [207], DELTA [10], GETAvatar [208]
	Diffusion + Explicit	• DINAR [209]
	NeRF + Explicit	• TransHuman [210]
	Gaussian Splatting + Explicit	• Animatable 3D Gaussian [211]

between 2D pixels and 3D vertices by using dense correspondences of body parts, effectively addressing related issues. Their proposed DaNet model concentrates on learning the 2D-to-3D mapping, while the Part-Drop strategy ensures that the model focuses more on complementary body parts and adjacent positional features. To address the lack of dense correspondence between image features and the 3D mesh, DecoMR [152] recovers the human mesh by establishing a pixel-to-surface dense correspondence map in the UV space. Zhang et al. [153] designed a two-branch network that utilizes a partial UV map to represent the human body when occluded by objects, effectively converting this map into an estimation of the 3D human shape. Sun et al. [154] proposed a body-center-guided representation method that predicts both the body center heatmap and the mesh parameter map. This approach describes the 3D body mesh at the pixel level, enabling one-stage multi-person 3D mesh regression. 3DCrowdNet [155] employs a joint-based regressor to isolate target features from others for recovering multi-person meshes in crowded scenes, which utilizes a crowded scene-robust image feature heatmap instead of the full feature map within a bounding box. Dual-stream Spatio-temporal Transformer (DSTformer) [156] extracts long-range spatio-temporal relationships among skeletal joints to effectively capture unified human motion representations from large-scale and heterogeneous video data for human-centric tasks, such as human mesh recovery.

Utilizing Structural Information. The structural information of the body acts as unique prior knowledge in human mesh recovery, enhancing the understanding of body relations. Furthermore, introducing additional physical constraints, which describe the interrelationships between various body structures, can significantly improve the performance. Holopose [157] employs a part-based multi-task regression network for 2D, 3D, and dense pose to estimate 3D human surfaces. Sun et al. [158] introduced an end-to-end method for human mesh recovery from single images and monocular videos. This method employs skeleton disentangling to reduce the complexity of decoupling and incorporates temporal coherence, efficiently capturing both short and long-term temporal cues. Hybrid Inverse Kinematics (HybrIK) [159] calculates the swing rotation from 3D joints and employs a network to predict the twist rotation through the twist-and-swing decomposition. The method can tackle non-linearity challenges and misalignment between images and models in mesh estimation from 3D poses. In their later work, Li et al. [160] designed NIKI, a model capable of learning from both the forward and inverse processes using invertible networks. To address the non-linear mapping and drifting joint position issues, Lee et al. [161] introduced an uncertainty-aware method for human mesh recovery, leveraging information from 2D poses to address the inherent ambiguities in 2D. Sengupta et al. [162] presented a probabilistic approach to circumvent the ill-posed problem, which integrates the kinematic tree structure of the human body with a Gaussian distribution over SMPL parameters. This method then predicts the hierarchical matrix-fisher distribution of 3D joint rotation matrices. SGRE [163] estimates the global rotation matrix of joints directly to avoid error accumulation along the kinematic chains in human mesh recovery.

Choosing Appropriate Learning Strategies. SPIN [164] incorporates an initial estimate optimization routine into the training loop by the self-improving neural network, which can fit the body mesh estimate to 2D joints. Zanfir et al. [48] developed a method integrating kinematic latent normalizing flow representations and dynamical models with structured, differentiable, semantic body part alignment loss functions aimed at enhancing semi-supervised and self-supervised 3D human pose and shape estimation. Jiang et al. [166] introduced two novel loss functions for multi-person mesh recovery from single images: a distance field-based collision loss penalizing interpenetration between constructed figures and a depth ordering-aware loss addressing occlusions and promoting accurate depth ordering of targets. Madadi et al. [167] presented an unsupervised denoising autoencoder network to recover invisible landmarks using sparse motion capture data effectively. To tackle the challenges of pose failure and shape ambiguity in the unsupervised human mesh recovery task, Yu et al. [50] devised a strategy that decouples the task into unsupervised 3D pose estimation and leverages kinematic prior knowledge. Bilevel Online Adaptation (BOA) [217] employs bilevel optimization to reconcile conflicts between 2D and temporal constraints in out-of-domain streaming videos human mesh recovery. In their later work, Dynamic Bilevel Online Adaptation (DBOA) [168] integrates temporal constraints to compensate for the absence of 3D annotations. Huang et al. [169] developed Pose2UV, a single-shot human mesh recovery method capable of extracting target features under occlusions using a deep UV prior. JOTR [170] fuses 2D and 3D features and incorporates supervision for the 3D feature through a Transformer-based contrastive learning framework. ReFit [165] reprojects keypoints and refines the human model via a feedback-update loop mechanism. You et al. [5] introduced a co-evolution method for human mesh recovery that utilizes 3D pose as an intermediary. This method divides the process into two distinct stages: initially, it estimates the 3D human pose from video, and subsequently, it regresses mesh vertices based on the estimated 3D pose, combined with temporal image features. To bridge the gap between training and test data, CycleAdapt [171] proposed a domain adaptation method including a mesh reconstruction network and a motion denoising network enabling more effective adaptation.

5.1.2. Detailed human body recovery

Model-based human mesh recovery has moderately satisfactory results but still lacks detailed body representations; expansions on the naked parametric model now allow for parameterized depictions of various body parts and details, including clothing [39,175], hands [36], face [37], and the entire body [38], as discussed in Section 2.2.

With Clothes. Alldieck et al. [172] estimated the parameters of the SMPL model, including clothing and hair, from 1 to 8 frames of a monocular video. Bhatnagar et al. [173] developed the Multi-Garment Network (MGN) to reconstruct body shape and clothing layers on top of the SMPL model. Tex2Shape [174] converts the human body mesh regression problem into an image-to-image estimation task. Specifically, it predicts a partial texture map of the visible region and then reconstructs the body shape, adding details to visible and occluded parts. BCNet [175] features a layered garment representation atop the SMPL model and innovatively decouples the skinning weight of the garment from the body mesh. Jiang et al. [39] introduced a method that employs a temporal span, SMPL parameters of shape and initial pose, and latent codes encoding motion and auxiliary information. This approach facilitates the recovery of detailed body shapes, including visible and occluded parts, by utilizing a texture map of the visible region.

With Hands. Forte et al. [176] proposed SGNify, a model that captures hand pose, facial expression, and body movement from sign language videos. It employs linguistic priors and constraints on 3D hand pose to effectively address the ambiguities in isolated signs. Additionally, the relationship between Two-Hands [177], and Hand-Object [178] effectively reconstructs the hand's details.

Whole Body. To address the inconsistency between 3D human mesh and 3D scenes, Hassan et al. [179] introduced a human whole body and scenes recovery method named Proximal Relationships with Object eXclusion (PROX). EXpressive POse and Shape rEgression (ExPose) framework [180] employs a body-driven attention mechanism and adopts a regression approach for holistic expressive body reconstruction to mitigate local optima issues in optimization-based methods. FrankMocap [181] operates by independently running 3D mesh recovery regression for face, hands, and body and subsequently combining the outputs through an integration module. PIXIE [182] integrates independent estimates from the body, face, and hands using the shared shape space of SMPL-X across all body parts. Moon et al. [183] developed an end-to-end framework for whole-body human mesh recovery named Hand4Whole, which employs joint features for 3D joint rotations to enhance the accuracy of 3D hand predictions. Zhang et al. [184] enhanced the PyMAF framework [218], developing PyMAF-X for the detailed reconstruction of full-body models. This advancement aims to resolve the misalignment issues in regression-based, one-stage human mesh recovery methods by employing a feature pyramid approach and refining the mesh-image alignment parameters. OSX [185] employs a simple yet effective component-aware transformer that includes a global body encoder and a local face/hand decoder instead of separate networks for each part. Li et al. [186] extended the HybrIK [159] framework and proposed HybrIK-X, a one-stage model based on a hybrid analytical-neural inverse kinematics framework to recover comprehensive whole-body meshes with details.

5.2. Template-free human body recovery

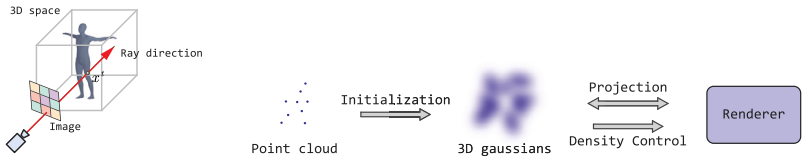
Template-free methods for human mesh recovery, such as neural network regression models, optimization models based on differentiable rendering, implicit representation models, Neural Radiance Fields (NeRF), and Gaussian Splatting, demonstrate enhanced flexibility over template-based approaches, enabling the depiction of richer details.

Regression-based Methods. Regression-based human mesh recovery bypasses the limitations and biases inherent in template-based models and directly outputs the template-free 3D models of the human body. This approach allows for a more dynamic and flexible generation of models, which can capture a wider variety of human body shapes and postures. FACSIMILE (FAX) utilizes an image-translation network to recover geometry at the original image resolution directly, bypassing the need for indirect output through representations. Jinka et al. [188] proposed a robust shape representation specifically for scenes with self-occlusions, where the method encodes the human body using peeled RGB and depth maps, significantly enhancing accuracy and efficiency during both training and inference. Neural Surface Fields (NSF) [190] models a continuous and flexible displacement field on the base surface for 3D clothed human mesh recovery from monocular depth. This approach adapts to base surfaces with varying resolutions and topologies without retraining during inference. Zhang et al. [189] introduced the Global-correlated 3D-decoupling Transformer for clothed Avatar reconstruction (GTA), a model that employs an encoder to capture globally-correlated image features and a decoder to decouple tri-plane features using cross-attention.

Optimization-based Methods. Optimization-based differentiable rendering integrates the rendering process into the optimization method by minimizing the rendering error. AG3D [192] captures the body's and loose clothing's shape and deformation by adopting a holistic 3D generator and integrating geometric cues in the form of predicted 2D normal maps. Gärtner et al. [191] designed DiffPhy, a differentiable physics-based model that incorporates a physically plausible body representation with anatomical joint limits, significantly enhancing the performance and robustness of human body recovery.

Implicit Representation Methods. Implicit representations do not directly represent an object's geometric data, such as vertices or meshes, but rather define whether a point in space belongs to the object through a function. The advantage of implicit representations lies in their ability to compactly and flexibly represent complex shapes, including those with intricate topologies or discontinuities. Thus, they have achieved impressive outcomes in human body reconstruction. Saito et al. [193] proposed the Pixel-aligned Implicit Function (PIFu), an implicit representation that aligns pixels of 2D images with a 3D object, and designed an end-to-end deep learning framework for inferring both 3D surface and texture from a single image. PIFu was the first to apply implicit representation in human mesh recovery, enabling the reconstruction of geometric details of the human body in clothing. Subsequently, PIFuHD [194] extended the work of PIFu to 4 K resolution images, further enhancing detail. Huang et al. [195] designed the Animatable Reconstruction of Clothed Humans (ARCH) method by creating a semantic space and a semantic deformation field. He et al. [196] introduced ARCH++, a co-supervising framework with cross-space consistency, which jointly estimates occupancy in both the posed and canonical spaces. They transform the human mesh recovery problem with implicit representation from pose space to canonical space for processing. However, this approach's limitation is that it heavily relies on accurate pose estimation, and the clothing expression based on skinning weights still lacks sufficient naturalness in detail. GeoPIFu [198] learns latent voxel features using a structure-aware 3D U-Net incorporating geometric priors. To tackle the ill-posed problem in representation learning when views are incredibly sparse, Peng et al. [199] developed a novel human body representation. This representation assumes that the neural representations learned at different frames share latent codes anchored to a deformable mesh. Clothed Avatar Reconstruction (CAR) method [197] employs a learning-based implicit model to initially form the general shape in canonical space, followed by refining surface details through predicting non-rigid optimization deformation in the posed space. 3D-aware Neural Body Fitting (3DNBF) [200] addresses the challenges of occlusion and 2D-3D ambiguity by a volumetric human representation using Gaussian ellipsoidal kernels.

Neural Radiance Fields. As illustrated in Fig. 8(a), Neural Radiance Fields [219] is based on the principle of implicit representation, using neural networks to learn the continuous volumetric density and color distribution of a scene, allowing for generating a high-quality 3D model from arbitrary viewpoints. Gao et al. [201] utilized a specialized representation that combines a canonical Neural Radiance Field (NeRF) with a volume deformation scheme, enabling the recovery of novel views and poses for a person not seen during training. Mu et al. [51]



(a) NeRF (b) 3D Guassian Splatting

Fig. 8. (a) NeRF; (b) 3D Guassian Splatting.

introduce ActorsNeRF, a NeRF-based human representation for human mesh recovery from a few monocular images by encoding the category-level prior through parameter sharing with a 2-level canonical space.

Diffusion models. Diffusion models are based on a series of diffusion processes, transforming original data by adding random noise and then gradually removing this noise to generate new data through a reverse process. Human Mesh Diffusion (HMDiff) [202] treats mesh recovery as a reverse diffusion process, incorporates input-specific distribution information into the diffusion process, and introduces prior knowledge to simplify the task.

Implicit Representation with Explicit Parameter. Methods based on implicit representations can recover free-form geometric shapes, but they may have excessive dependence on accurate poses or generate unsatisfactory shapes for novel poses or clothing. To increase robustness in these scenarios, existing work employs explicit body models to constrain mesh reconstruction in implicit methods. Researchers desire an approach that combines implicit and explicit methods based on a coarse body mesh's geometry to refine and produce detailed human models meticulously. Hierarchical Mesh Deformation (HMD) [203] uses constraints from body joints, silhouettes, and per-pixel shading information to combine the robustness of a parametric model with the flexibility of free-form 3D deformation. Bhatnagar et al. [204] introduced an Implicit Part Network (IP-Net) to predict the detailed human surface, including the 3D surface of the dressed person, the inner body surface, and the parametric body model. Parametric Model-Conditioned Implicit Representation (PaMIR) [42] combines a parametric body model with a free-form deep implicit function to enhance generalization through regularization. Zhu et al. [205] utilized a 'project-predict-deform' strategy that refines the SMPL model generated by human recovery methods using supervision from joints, silhouettes, and shading information. ICON [206] generates a stable, coarse human mesh using the SMPL model, then renders the front and back body normals, which provide rich texture details and combine with the original image. Finally, the body's normal clothed normal, and Signed Distance Function (SDF) information are fed into an implicit MLP to obtain the final result. Subsequently, Xiu et al. proposed ECON [207], which integrates normal estimation, normal integration, and shape completion by using SMPL-X depth as a soft geometric constraint within the optimization equation of Normal Integration. It endeavors to maintain coherence with nearby surfaces during the integration of normals. Feng et al. [10] presented Disentangled Avatars (DELTA), a model representing humans with hybrid explicit–implicit 3D representations. GETAvatar [208] directly produces explicit textured 3D human meshes. GETAvatar creates an articulated 3D human representation with explicit surface modeling, enriches it with realistic surface details derived from 2D normal maps of 3D scan data, and utilizes a rasterization-based renderer for surface rendering. Diffusion Inpainting of Neural Avatars (DINAR) [209] combines neural textures with the SMPL-X body model and employs a latent diffusion model to recover textures of both seen and unseen regions, subsequently integrating these onto the base SMPL-X mesh. TransHuman [210] employs transformers to project the SMPL model into a canonical space and associates each output token with a deformable radiance field. This field encodes the query point in the observation space and is further utilized to integrate fine-grained information from reference images. Gaussian Splatting [220] is a technique for processing and rendering point clouds, using a Gaussian function to create an influence range for each point, resulting in a smoother and more natural representation in two-dimensional images, as shown in Fig. 8(b). It has achieved good results in SLAM (Simultaneous Localization and Mapping) [221], generative human modeling [222], dynamic scene reconstruction [223], and multimodal generation [224]. Reconstruction based on Gaussian Splatting requires an initial point cloud input, which poses a challenge for human mesh recovery from image inputs. However, explicit models can provide initial points, serving as a stepping stone to make Gaussian splatting human mesh recovery from RGB inputs feasible. Animatable 3D Gaussian [211] learns human avatars from input images and poses by extending 3D Gaussian to dynamic human scenes. In the proposed framework, they model a set of skinned 3D Gaussian and a corresponding skeleton in canonical space and deform 3D Gaussian to posed space according to the input poses.

5.3. Summary of human mesh recovery

This chapter provides a comprehensive review of human mesh recovery utilizing both explicit and implicit models. Explicit models excel in robustly reconstructing human mesh but often fall short in capturing intricate details, prompting a range of extensions for greater detail accuracy. In contrast, implicit-based human mesh recovery tends to lack stability, which is known for its flexibility and adaptability. Consequently, several studies have explored integrating implicit models with explicit counterparts to synergize their respective strengths. The trade-off between flexibility and robustness in human mesh recovery represents a pivotal and enduring area of research.

6. Evaluation

6.1. Evaluation metrics

There are many evaluation metrics that can be used fairly to measure the performance of deep models in human pose estimation, and some of the key evaluation metrics are provided below.

Mean Per Joint Position Error (MPJPE) is widely used to evaluate the accuracy performance of 3D human pose estimation by calculating the L2 distance between the predicted joint coordinates and their ground truth counterparts. Denote the estimated coordinates of the jth joint as p_j^* and the ground truth coordinates as p_j. The MPJPE of jth joint in the skeleton can be computed as:

$$MPJPE = \frac{1}{N}\sum_{j=1}^{N}\left\|p_j - p_j^*\right\|_2 \tag{1}$$

where the skeleton comprises N joints, and unlike previously in 2D, where the error is quantified in pixels, the joint coordinates in 3D are measured and reported in millimeters (mm).

Mean Per Joint Angle Error (MPJAE) quantifies the angular discrepancy between the estimated and groundtruth joints. The function r_j^* returns the estimated angle of jth joint, and function r_j returns the

Table 3
The overview of the mainstream datasets.

Dataset	Type	Data	Total frames	Feature	Download link
Human3.6M [225]	3D/Mesh	Video	3.6M	multi-view	Website
3DPW [227]	3D/Mesh	Video	51K	multi-person	Website
MPI-INF-3DHP [228]	2D/3D	Video	2K	in-wild	Website
HumanEva [229]	3D	Video	40K	multi-view	Website
CMU-Panoptic [230]	3D	Video	1.5M	multi-view/multi-person	Website
MuCo-3DHP [118]	3D	Image	8K	multi-person/occluded scence	Website
SURREAL [231]	2D/3D/Mesh	Video	6.0M	synthetic model	Website
3DOH50K [153]	2D/3D/Mesh	Image	51K	object-occluded	Website
3DCP [232]	Mesh	Mesh	190	contact	Website
AMASS [233]	Mesh	Motion	11K	soft-tissue dynamics	Website
DensePose [234]	Mesh	Image	50K	multi-person	Website
UP-3D [235]	3D/Mesh	Image	8K	sport scence	Website
THuman2.0 [236]	Mesh	Image	7K	textured surface	Website

ground truth angle. The MPJAE is computed in three dimensions as follows:

$$MPJAE = \frac{1}{3N}\sum_{j=1}^{3N}\left|\left(r_j - r_j^*\right) \bmod \pm 180\right| \tag{2}$$

Mean Per Joint Localization Error (MPJLE) is a more perceptive and robust evaluation metric than MPJPE and MPJAE [225]. It allows for an adjustable tolerance level through a perceptual threshold t :

$$MPJLE = \frac{1}{N}\sum_{j=1}^{N}\mathbb{1}_{\left\|l_j - l_j^*\right\|_2 \geq t} \tag{3}$$

where $l(\cdot)$ indicates the joint localization.

There are various modifications of MPJPE, MPJAE, and MPJLE, including **Procrustes-aligned (PA-)** MPJPE, MPJAE, MPJLE and **Normalized (N-)** MPJPE, MPJAE, MPJLE. The former refers to procrustes-aligned metrics, while the latter denotes normalized metrics. Additionally, some 2D metrics are adaptable to 3D, such as the **3D Percentage of Correct Keypoints (3D PCK)** and the **3D Area Under the Curve (3D AUC)**. The PCK [226] measures the distance between predicted and groundtruth keypoints, considering keypoints correct if this distance is less than a predefined threshold. The AUC signifies the aggregate area beneath the PCK threshold curve as the threshold varies.

Mean Per Vertex Position Error (MPVPE) is a metric used to evaluate the human mesh reconstruction by computing the L2 distance between the predicted and ground truth mesh points. In some published articles, the MPVPE is also called Vertex-to-Vertex (V2V) error and Per-Vertex Error (PVE).

6.2. Datasets

Datasets are essential in developing deep learning-based human pose estimation and mesh recovery. Researchers have developed many datasets to train models and facilitate fair comparisons among different methods. In this section, we introduce the details of related datasets from recent years. Summary of these datasets, including information on 3D pose and 3D mesh, are presented in Table 3.

Human3.6M dataset [225] is the most widely used dataset in the evaluation of 3D human pose estimation. It encompasses a vast collection of 3.6 million poses captured using RGB and ToF cameras from diverse viewpoints within a real-world setting. Additionally, this dataset incorporates high-resolution 3D scanner data of body meshes, playing a pivotal role in the progression of human sensing systems. Table 4 exhibits the performance of state-of-the-art 3D human pose estimation methods on the Human3.6M dataset.

MPI-INF-3DHP dataset [228] offers over 2 K videos featuring joint annotations of 13 keypoints in outdoor scenes, apt for 2D and 3D human pose estimation. The ground truth was obtained using a multi-camera arrangement and a marker-less MoCap system, representing a shift from traditional marker-based MoCap systems that involve real individuals. Table 5 showcases the performance of state-of-the-art methods on the 3DHP dataset.

3DPW dataset [227] captures 51,000 sequences of single-view video, complemented by IMUs data. These videos were recorded using a handheld camera, with the IMU data facilitating the association of 2D poses with their 3D counterparts. 3DPW stands out as one of the most formidable datasets, establishing itself as a benchmark for 3D pose estimation in multi-person, in-wild scenarios of recent times. Table 6 shows the performance of state-of-the-art human mesh recovery methods on the Human3.6M and 3DPW datasets.

HumanEva dataset [229] is a multi-view 3D human pose estimation dataset comprising two versions: HumanEva-I and HumanEva-II. In HumanEva-I, the dataset includes around 40,000 multi-view video frames captured from seven cameras positioned at the front, left, and right (RGB) and four corners (Mono). Additionally, HumanEva-II features approximately 2460 frames, recorded with four cameras at each corner.

CMU-Panoptic dataset [230,251] includes 65 frame sequences, approximately 5.5 h of footage, and features 1.5 million 3D annotated poses. Recorded via a massively multi-view system equipped with 511 calibrated cameras and 10 RGB-D sensors featuring hardware-based synchronization, this dataset is crucial for developing weakly supervised methods through multi-view geometry. These methods address the occlusion problems commonly encountered in traditional computer vision techniques.

Multiperson Composited 3D Human Pose (MuCo-3DHP) dataset [118] serves as a large-scale, multi-person occluded training set for 3D human pose estimation. Frames in the MuCo-3DHP are generated from the MPI-INF-3DHP dataset through a compositing and augmentation scheme.

SURREAL dataset [231] is a large synthetic human body dataset containing 6 million RGB video frames. It provides a range of accurate annotations, including depth, body parts, optical flow, 2D/3D poses, and surfaces. In the SURREAL dataset, images exhibit variations in texture, view, and pose, and the body models are based on the SMPL parameters, a widely-recognized mesh representation standard.

3DOH50K dataset [153] offers a collection of 51,600 images obtained from six distinct viewpoints in real-world settings, predominantly featuring object occlusions. Each image is annotated with ground truth 2D and 3D poses, SMPL parameters, and a segmentation mask. Utilized for training human estimation and reconstruction models, the 3DOH50K dataset facilitates exceptional performance in occlusion scenarios.

3DCP dataset [232] represents a 3D human mesh dataset, derived from AMASS [233]. It includes 190 self-contact meshes spanning six human subjects (three males and three females), each modeled with an SMPL-X parameterized template.

AMASS dataset [233] constitutes a comprehensive and diverse human motion dataset, encompassing over 11,000 motions from 300 subjects, totaling more than 40 h. The motion data, accompanied by SMPL parameters for skeleton and mesh representation, is derived from a marker-based MoCap system utilizing 15 optical markers.

Table 4
Comparisons of 3D pose estimation methods on Human3.6M [225].

Method	Year	Publication	Highlight	MPJPE↓	PMPJPE↓	Code
Graformer [237]	2022	CVPR'22	graph-based transformer	35.2	–	Code
GLA-GCN [238]	2023	ICCV'23	adaptive GCN	34.4	37.8	Code
PoseDA [49]	2023	arXiv'23	domain adaptation	49.4	34.2	Code
GFPose [239]	2023	CVPR'23	gradient fields	35.6	30.5	Code
TP-LSTMs [240]	2022	TPAMI'22	pose similarity metric	40.5	31.8	–
FTCM [122]	2023	TCSVT'23	frequency-temporal collaborative	28.1	-	Code
VideoPose3D [89]	2019	CVPR'19	semi-supervised	46.8	36.5	Code
PoseFormer [90]	2021	ICCV'21	spatio-temporal transformer	44.3	34.6	Code
STCFormer [96]	2023	CVPR'23	spatio-temporal transformer	40.5	31.8	Code
3Dpose_ssl [85]	2020	TPAMI'20	self-supervised	63.6	63.7	Code
MTF-Transformer [32]	2022	TPAMI'22	multi-view temporal fusion	26.2	–	Code
AdaptPose [88]	2022	CVPR'22	cross datasets	42.5	34.0	Code
3D-HPE-PAA [101]	2022	TIP'22	part aware attention	43.1	33.7	Code
DeciWatch [241]	2022	ECCV'22	efficient framework	52.8	-	Code
Diffpose [242]	2023	CVPR'23	pose refine	36.9	28.7	Code
Elepose [83]	2022	CVPR'22	unsupervised	-	36.7	Code
Uplift and Upsample [21]	2023	CVPR'23	efficient transformers	48.1	37.6	Code
RS-Net [59]	2023	TIP'23	regular splitting graph network	48.6	38.9	Code
HSTFormer [95]	2023	arXiv'23	spatial–temporal transformers	42.7	33.7	Code
PoseFormerV2 [60]	2023	CVPR'23	frequency domain	45.2	35.6	Code
DiffPose [243]	2023	ICCV'23	diffusion models	42.9	30.8	Code

Table 5
Comparisons of 3D pose estimation methods on MPI-INF-3DHP [228].

Method	Year	Publication	Highlight	MPJPE↓	PCK↑	AUC↑	Code
HSTFormer [95]	2023	arXiv'23	spatial–temporal transformers	28.3	98.0	78.6	Code
PoseFormerV2 [60]	2023	CVPR'23	frequency domain	27.8	97.9	78.8	Code
Uplift and Upsample [21]	2023	CVPR'23	efficient transformers	46.9	95.4	67.6	Code
RS-Net [59]	2023	TIP'23	regular splitting graph network	-	85.6	53.2	Code
Diffpose [242]	2023	CVPR'23	pose refine	29.1	98.0	75.9	Code
FTCM [122]	2023	TCSVT'23	frequency-temporal collaborative	31.2	97.9	79.8	Code
STCFormer [96]	2023	CVPR'23	spatio-temporal transformer	23.1	98.7	83.9	Code
PoseDA [49]	2023	arXiv'23	domain adaptation	61.3	92.0	62.5	Code
TP-LSTMs [240]	2022	TPAMI'22	pose similarity metric	48.8	82.6	81.3	–
AdaptPose [88]	2022	CVPR'22	cross datasets	77.2	88.4	54.2	Code
3D-HPE-PAA [101]	2022	TIP'22	part aware attention	69.4	90.3	57.8	Code
Elepose [83]	2022	CVPR'22	unsupervised	54.0	86.0	50.1	Code

Table 6
Comparisons of human mesh recovery methods on Human3.6M [225] and 3DPW [227].

Method	Publication	Highlight	Human3.6M		3DPW			Code
			MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	PVE↓	
VirtualMarker [244]	CVPR'23	novel intermediate representation	47.3	32.0	67.5	41.3	77.9	Code
NIKI [160]	CVPR'23	inverse kinematics	-	-	71.3	40.6	86.6	Code
TORE [149]	ICCV'23	efficient transformer	59.6	36.4	72.3	44.4	88.2	Code
JOTR [170]	ICCV'23	contrastive learning	-	-	76.4	48.7	92.6	Code
HMDiff [202]	ICCV'23	reverse diffusion processing	49.3	32.4	72.7	44.5	82.4	Code
ReFit [165]	ICCV'23	recurrent fitting network	48.4	32.2	65.8	41.0	-	Code
PyMAF-X [184]	TPAMI'23	regression-based one-stage whole body	–	–	74.2	45.3	87.0	Code
PointHMR [245]	CVPR'23	vertex-relevant feature extraction	48.3	32.9	73.9	44.9	85.5	-
PLIKS [246]	CVPR'23	inverse kinematics	47.0	34.5	60.5	38.5	73.3	Code
ProPose [247]	CVPR'23	learning analytical posterior probability	45.7	29.1	68.3	40.6	79.4	Code
POTTER [248]	CVPR'23	pooling attention transformer	56.5	35.1	75.0	44.8	87.4	Code
PoseExaminer [249]	ICCV'23	automated testing of out-of-distribution	-	-	74.5	46.5	88.6	Code
MotionBERT [156]	ICCV'23	pretrained human representations	43.1	27.8	68.8	40.6	79.4	Code
3DNBF [200]	ICCV'23	analysis-by-synthesis approach	-	-	88.8	53.3	-	Code
FastMETRO [134]	ECCV'22	efficient architecture	52.2	33.7	73.5	44.6	84.1	Code
CLIFF [129]	ECCV'22	multi-modality inputs	47.1	32.7	69.0	43.0	81.2	Code
PARE [130]	ICCV'21	part-driven attention	–	–	74.5	46.5	88.6	Code
Graphormer [131]	ICCV'21	GCNN-reinforced transformer	51.2	34.5	74.7	45.6	87.7	Code
PSVT [133]	CVPR'23	spatio-temporal encoder	–	–	73.1	43.5	84.0	–
GLoT [142]	CVPR'23	short-term and long-term temporal correlations	67.0	46.3	80.7	50.6	96.3	Code
MPS-Net [132]	CVPR'23	temporally adjacent representations	69.4	47.4	91.6	54.0	109.6	Code
MAED [140]	ICCV'21	multi-level attention	56.4	38.7	79.1	45.7	92.6	Code
Lee et al. [161]	ICCV'21	uncertainty-aware	58.4	38.4	92.8	52.2	106.1	–
TCMR [139]	CVPR'21	temporal consistency	62.3	41.1	95.0	55.8	111.3	-
VIBE [138]	CVPR'20	self-attention temporal network	65.6	41.4	82.9	51.9	99.1	Code
ImpHMR [250]	CVPR'23	implicitly imagine person in 3D space	-	-	74.3	45.4	87.1	-
SGRE [163]	ICCV'23	sequentially global rotation estimation	–	–	78.4	49.6	93.3	Code
PMCE [5]	ICCV'23	pose and mesh co-evolution network	53.5	37.7	69.5	46.7	84.8	Code

DensePose dataset [234] features 50,000 manually annotated real images, comprising 5 million image-to-surface correspondence pairs extracted from the COCO [252] dataset. This dataset proves instrumental for training in dense human pose estimation, as well as in detection and segmentation tasks.

UP-3D dataset [235] is a dedicated 3D human pose and shape estimation dataset featuring extensive annotations in sports scenarios. The UP-3D comprises approximately 8000 images from the LSP and MPII datasets. Additionally, each image in UP-3D is accompanied by a metadata file indicating the quality (medium or high) of the 3D fit.

THuman dataset [236] constitutes a 3D real-world human mesh dataset. It includes 7000 RGBD images, each featuring a textured surface mesh obtained using a Kinect camera. Including surface mesh with detailed texture and the aligned SMPL model is anticipated to significantly enhance and stimulate future research in human mesh reconstruction.

7. Applications

In this section, we review related works about human pose estimation and mesh recovery for a few popular applications.

Motion Retargeting. Human motion retargeting can transfer human actions onto actors. Using pose estimation eliminates the need for motion capture systems and achieves image-to-image translation. Therefore, 3D human pose estimation is crucial for retargeting. Recently, there has been a significant amount of retargeting work based on 3D human pose estimation [253–255]. End-to-end methods and related datasets are also designed [256]. Additionally, unsupervised methods in In-the-Wild scenarios [257] have been developed, achieved through canonicalization operations and derived regularizations. Beyond bodily retargeting, facial retargeting [258,259] has also gained prominence, wherein the intricacies of facial expressions offer a refined portrayal of the actor's emotional and psychological states.

Human Avatars. Human avatars are human-like virtual entities created through digital technology, which can interact and express themselves on various digital media platforms. Human pose estimation and mesh recovery enable digital avatars to simulate human movements and behaviors more realistically and vividly, allowing developers to create virtual characters with behaviors highly similar to real individuals [209,260]. For instance, Luo et al. [261] proposed a real-time multi-person avatar controller using noisy poses from video-based pose estimators and language-based motion generators.

Action Recognition. Action recognition employs algorithms to identify and analyze human movements from images or videos. The results derived from 3D human pose estimation and reconstruction play a pivotal role in deciphering the dynamics of human motion within a three-dimensional context, thereby transforming these movements into actionable behavioral insights [1,262–264]. Moreover, these advancements can significantly augment the efficiency of action recognition [265]. It is also possible to address pose estimation and action recognition within the same framework through multi-task learning and sharing features [266].

Security Monitoring. In video surveillance systems at public places or critical facilities, pedestrian tracking and re-identification are key tasks. Combined with human pose estimation, Human tracking is utilized for the surveillance and analysis of pedestrian flow, tracking specific targets, and tracking and analyzing human behavior in space. This approach is highly beneficial for tracking humans in complex scenarios [267–270]. Considering the constrained viewpoints of individual cameras in such systems, Re-identification enables the recognition and tracking of the same person as they move across different camera views. Human pose estimation significantly contributes to the enhancement of this re-identification [11].

SLAM. SLAM precisely estimates its location by gathering sensor data from its environment, employing technologies such as cameras and LiDAR. It simultaneously constructs or refines an environmental map, facilitating self-localization and map creation in unfamiliar territories. Diverging from the conventional SLAM systems that predominantly concentrate on objects, recent advancements have shifted focus towards incorporating humans within the environmental context. Notably, Dai et al. [271] have illustrated the significance of 3D human trajectory reconstruction in indoor settings, enhancing indoor navigation capabilities. Furthermore, Kocabas et al. [272] have innovatively integrated human motion priors into SLAM, effectively merging human pose estimation with scene analysis.

Autonomous Driving. In robotic navigation and autonomous vehicle applications, estimating human pose enables these systems to comprehend human behavior and intentions better. This understanding facilitates more intelligent decision-making and interaction. Zheng et al. [14] propose a multi-modal approach employing 2D labels on RGB images as weak supervision for 3D human pose estimation in the context of autonomous vehicles. Wang et al. [15] have developed a comprehensive framework to learn physically plausible human dynamics from real driving scenarios, effectively bridging the gap between actual and simulated human behavior in safety-critical applications.

Human–Computer Interaction. Human–computer interaction entails the bidirectional communication between humans and computers, underscored by the critical need for computers to interpret human poses accurately. Liu et al. [12] advanced this field by proposing asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction, which utilizes an effective Lorentz distribution learning scheme. In Augmented Reality (AR) systems, the accuracy of human pose estimation significantly elevates the interaction quality between virtual objects and the real environment, fostering more natural interactions between humans and virtual entities. In this context, Weng et al. [16] developed a novel method and application for animating a human subject from a single photo within AR.

8. Challenges and conclusion

In this survey, we have presented a contemporary overview of recent deep learning-based 3D human pose estimation and mesh recovery methods. A comprehensive taxonomy and performance comparison of these methods has been covered. We further point out a few promising research directions, hoping to promote advances in this field.

Speed. Speed is an essential aspect to consider in the practical deployment of algorithms. While most current research papers report achieving real-time performance on GPUs, a wide array of applications necessitates real-time and efficient processing on edge computing platforms, notably on ARM processors within smartphones. The disparity in performance between ARM processors and GPUs is pronounced, underscoring the immense value of optimizing for speed. Although some video-based visual processing tasks [273–275] have already acknowledged the importance of processing speed for real-time application scenarios, further research is still needed in the context of 3D cases [77].

Crowding and Occlusion Challenges. In open-world scenarios, the phenomena of crowding and occlusion are prevalent, and they represent long-standing challenges in the field of object detection. Currently, top-down methods depend on object detection, rendering these issues inescapable. Although bottom-up strategies may circumvent object detection, they confront formidable challenges regarding key point assembly. Therefore, some researchers [121] attempt to integrate both methods together. Moreover, incorporating temporal consistency constraints (such as optical flow [102]) during mining can also alleviate occlusion issues.

Large Models. The efficacy of large models in language [276,277] and foundational computer vision tasks, such as segmentation [278] and tracking [279], has been universally acknowledged and is quite remarkable. Additionally, while there is burgeoning research in human-centric computer vision tasks [280], the area of 3D tasks still demands

further investigation. Furthermore, the development of large models not only constitutes a significant area of research but also the exploration of their effective utilization presents substantial challenges and practical importance. Exploratory efforts to fuse pose estimation with large language models [281] and the combination of large visual models with pose estimation are also meaningful.

More Detailed Reconstruction. At present, explicit model-based methodologies such as SMPL [35] and SMPL-X [61], fall short of meeting the demands for detail that people expect. On the other hand, implicit representation approaches [193,195,198], as well as rendering techniques such as NeRF [51,201] and Gaussian Splatting [211], are capable of capturing fine details but lack sufficient robustness in pose estimation. Bridging the gap between robust pose estimation and surface details remains a formidable challenge that necessitates collaborative efforts from computer vision and computer graphics researchers.

Reconstructing with the Environment. Humans engage with various objects and environments within the real world in the daily life. The ability to reconstruct these interactions in 3D is paramount for comprehending and simulating their dynamics. This aspect holds particular significance in the realm of artificial intelligence applied to human avatars. In the reconstruction of object interactions, challenges such as pose initialization and sparse observation of objects in Structure from Motion are encountered [7]. Moreover, Yi et al. [282] indicate a reciprocal enhancement between environmental and human reconstruction, resulting in enhanced effectiveness across both domains.

Controllability and Animatability. Controllability refers to the precise control over the postures, movements, and expressions of human avatars, enabling them to exhibit various desired behaviors in virtual environments. Animation, on the other hand, denotes the ability of avatars to vividly and smoothly demonstrate various movements and expressions, making them appear more realistic and natural. Jiang et al. [283] employed an explicit deformation field to deform the neural implicit field, thus enabling the animation of human characters. They mapped the target human body mesh to the template human body mesh, represented by a parameterized human body model, thereby simplifying the animation and reshaping of the generated avatars through the control of pose and shape parameters. By enhancing the controllability and animatability of human body mesh reconstruction, developers can create digital characters that are more realistic, vivid, and expressive. This, in turn, enhances their interactivity and appeal in the virtual world.

CRediT authorship contribution statement

Yang Liu: Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Changzhen Qiu:** Writing – review & editing. **Zhiyong Zhang:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2969–2978.

[2] Y. Zhang, C. Wang, X. Wang, W. Liu, W. Zeng, Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2022) 2613–2626.

[3] Y. Zhu, H. Shuai, G. Liu, Q. Liu, Multilevel spatial–temporal excited graph network for skeleton-based action recognition, IEEE Trans. Image Process. 32 (2022) 496–508.

[4] J. Yang, Z. Zhang, S. Xiao, S. Ma, Y. Li, W. Lu, X. Gao, Efficient data-driven behavior identification based on vision transformers for human activity understanding, Neurocomputing 530 (2023) 104–115.

[5] Y. You, H. Liu, T. Wang, W. Li, R. Ding, X. Li, Co-evolution of pose and mesh for 3D human body estimation from video, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14963–14973.

[6] S. Tripathi, L. Müller, C.-H.P. Huang, O. Taheri, M.J. Black, D. Tzionas, 3D human pose estimation via intuitive physics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4713–4725.

[7] Z. Fan, M. Parelli, M.E. Kadoglou, M. Kocabas, X. Chen, M.J. Black, O. Hilliges, HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video, 2023, arXiv preprint arXiv:2311.18448.

[8] L. Dai, L. Ma, S. Qian, H. Liu, Z. Liu, H. Xiong, Cloth2Body: Generating 3D human body mesh from 2D clothing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15007–15017.

[9] S. Tang, G. Wang, Q. Ran, L. Li, L. Shen, P. Tan, High-resolution volumetric reconstruction for clothed humans, ACM Trans. Graph. 42 (5) (2023) 1–15.

[10] Y. Feng, W. Liu, T. Bolkart, J. Yang, M. Pollefeys, M.J. Black, Learning disentangled avatars with hybrid 3d representations, 2023, arXiv preprint arXiv:2309.06441.

[11] P. Wang, Z. Zhao, F. Su, X. Zu, N.V. Boulgouris, HOReID: Deep high-order mapping enhances pose alignment for person re-identification, IEEE Trans. Image Process. 30 (2021) 2908–2922.

[12] H. Liu, T. Liu, Z. Zhang, A.K. Sangaiah, B. Yang, Y. Li, Arhpe: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction, IEEE Trans. Ind. Inform. 18 (10) (2022) 7107–7117.

[13] J. Zou, Simplified neural architecture for efficient human motion prediction in human-robot interaction, Neurocomputing 588 (2024) 127683.

[14] J. Zheng, X. Shi, A. Gorban, J. Mao, Y. Song, C.R. Qi, T. Liu, V. Chari, A. Cornman, Y. Zhou, et al., Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4478–4487.

[15] J. Wang, Y. Yuan, Z. Luo, K. Xie, D. Lin, U. Iqbal, S. Fidler, S. Khamis, Learning human dynamics in autonomous driving scenarios, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 20796–20806.

[16] C.-Y. Weng, B. Curless, I. Kemelmacher-Shlizerman, Photo wake-up: 3d character animation from a single photo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5908–5917.

[17] W. Liu, Q. Bao, Y. Sun, T. Mei, Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective, ACM Comput. Surv. 55 (4) (2022) 1–41.

[18] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, M. Shah, Deep learning-based human pose estimation: A survey, ACM Comput. Surv. 56 (1) (2023) 1–37.

[19] Y. Tian, H. Zhang, Y. Liu, L. Wang, Recovering 3d human mesh from monocular images: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[20] L. Chen, S. Peng, X. Zhou, Towards efficient and photorealistic 3d human reconstruction: A brief survey, Vis. Inform. 5 (4) (2021) 11–19.

[21] M. Einfalt, K. Ludwig, R. Lienhart, Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2903–2913.

[22] Y. Luo, Y. Li, M. Foshey, W. Shou, P. Sharma, T. Palacios, A. Torralba, W. Matusik, Intelligent carpet: Inferring 3d human pose from tactile signals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11255–11265.

[23] A. Ruget, M. Tyler, G. Mora Martín, S. Scholes, F. Zhu, I. Gyongy, B. Hearn, S. McLaughlin, A. Halimi, J. Leach, Pixels2Pose: Super-resolution time-of-flight imaging for 3D pose estimation, Sci. Adv. 8 (48) (2022) eade0123.

[24] R. Pandey, A. Tkach, S. Yang, P. Pidlypenskyi, J. Taylor, R. Martin-Brualla, A. Tagliasacchi, G. Papandreou, P. Davidson, C. Keskin, et al., Volumetric capture of humans with a single rgbd camera via semi-parametric learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9709–9718.

[25] Y. Ren, Z. Wang, Y. Wang, S. Tan, Y. Chen, J. Yang, GoPose: 3D human pose estimation using WiFi, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6 (2) (2022) 1–25.

[26] T. Li, L. Fan, Y. Yuan, D. Katabi, Unsupervised learning for human sensing using radio signals, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3288–3297.

[27] J.L. Ponton, H. Yun, A. Aristidou, C. Andujar, N. Pelechano, SparsePoser: Real-time full-body motion reconstruction from sparse data, ACM Trans. Graph. 43 (1) (2023) 1–14.

[28] F. Huang, A. Zeng, M. Liu, Q. Lai, Q. Xu, Deepfuse: An imu-aware network for real-time 3d human pose estimation from multi-view image, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 429–438.

[29] S. Zou, X. Zuo, S. Wang, Y. Qian, C. Guo, L. Cheng, Human pose and shape estimation from single polarization images, IEEE Trans. Multimed. (2022).

[30] L. Xu, W. Xu, V. Golyanik, M. Habermann, L. Fang, C. Theobalt, Eventcap: Monocular 3d capture of high-speed human motions using an event camera, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4968–4978.

[31] B. Jiang, L. Hu, S. Xia, Probabilistic triangulation for uncalibrated multi-view 3D human pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14850–14860.

[32] H. Shuai, L. Wu, Q. Liu, Adaptive multi-view and temporal fusing transformer for 3d human pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2022) 4122–4135.

[33] B. Huang, Y. Shu, T. Zhang, Y. Wang, Dynamic multi-person mesh recovery from uncalibrated multi-view cameras, in: 2021 International Conference on 3D Vision, 3DV, IEEE, 2021, pp. 710–720.

[34] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, J. Davis, Scape: Shape completion and animation of people, in: ACM SIGGRAPH 2005 Papers, 2005, pp. 408–416.

[35] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: A skinned multi-person linear model, ACM Trans. Grap. (TOG) 34 (6) (2015) 1–16.

[36] J. Romero, D. Tzionas, M.J. Black, Embodied hands: Modeling and capturing hands and bodies together, 2022, arXiv preprint arXiv:2201.02610.

[37] T. Li, T. Bolkart, M.J. Black, H. Li, J. Romero, Learning a model of facial shape and expression from 4D scans, ACM Trans. Graph. 36 (6) (2017) 1–194.

[38] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A.A. Osman, D. Tzionas, M.J. Black, Expressive body capture: 3d hands, face, and body from a single image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10975–10985.

[39] B. Jiang, Y. Zhang, X. Wei, X. Xue, Y. Fu, H4d: Human 4d modeling by learning neural compositional representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19355–19365.

[40] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, C. Schmid, Bodynet: Volumetric inference of 3d human body shapes, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 20–36.

[41] H. Onizuka, Z. Hayirci, D. Thomas, A. Sugimoto, H. Uchiyama, R.-i. Taniguchi, TetraTSDF: 3D human reconstruction from a single image with a tetrahedral outer shell, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6011–6020.

[42] Z. Zheng, T. Yu, Y. Liu, Q. Dai, Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, IEEE Trans. Pattern Anal. Mach. Intell. 44 (6) (2021) 3170–3184.

[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[44] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3349–3364.

[45] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., Mlp-mixer: An all-mlp architecture for vision, in: Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 24261–24272.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[47] C.-Y. Yang, J. Luo, L. Xia, Y. Sun, N. Qiao, K. Zhang, Z. Jiang, J.-N. Hwang, C.-H. Kuo, CameraPose: Weakly-supervised monocular 3D human pose estimation by leveraging in-the-wild 2D annotations, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2924–2933.

[48] A. Zanfir, E.G. Bazavan, H. Xu, W.T. Freeman, R. Sukthankar, C. Sminchisescu, Weakly supervised 3d human pose and shape reconstruction with normalizing flows, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, Springer, 2020, pp. 465–481.

[49] W. Chai, Z. Jiang, J.-N. Hwang, G. Wang, Global adaptation meets local generalization: Unsupervised domain adaptation for 3D human pose estimation, 2023, arXiv preprint arXiv:2303.16456.

[50] Z. Yu, J. Wang, J. Xu, B. Ni, C. Zhao, M. Wang, W. Zhang, Skeleton2mesh: Kinematics prior injected unsupervised human mesh recovery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8619–8629.

[51] J. Mu, S. Sang, N. Vasconcelos, X. Wang, ActorsNeRF: Animatable few-shot human rendering with generalizable NeRFs, 2023, arXiv preprint arXiv:2304.14401.

[52] A. Benzine, F. Chabot, B. Luvison, Q.C. Pham, C. Achard, Pandanet: Anchor-based single-shot multi-person 3d pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6856–6865.

[53] Z. Yang, A. Zeng, C. Yuan, Y. Li, Effective whole-body pose estimation with two-stages distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4210–4220.

[54] S. Tripathi, S. Ranade, A. Tyagi, A. Agrawal, Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation, in: 2020 International Conference on 3D Vision, 3DV, IEEE, 2020, pp. 311–321.

[55] H. Liu, C. Ren, An effective 3D human pose estimation method based on dilated convolutions for videos, in: 2019 IEEE International Conference on Robotics and Biomimetics, ROBIO, IEEE, 2019, pp. 2327–2331.

[56] S. Choi, S. Choi, C. Kim, MobileHumanPose: Toward real-time 3D human pose estimation in mobile devices, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2328–2338.

[57] H. Cho, Y. Cho, J. Yu, J. Kim, Camera distortion-aware 3d human pose estimation in video with optimization-based meta-learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11169–11178.

[58] K. Gong, B. Li, J. Zhang, T. Wang, J. Huang, M.B. Mi, J. Feng, X. Wang, PoseTriplet: Co-evolving 3D human pose estimation, imitation, and hallucination under self-supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11017–11027.

[59] M.T. Hassan, A.B. Hamza, Regular splitting graph network for 3d human pose estimation, IEEE Trans. Image Process. (2023).

[60] Q. Zhao, C. Zheng, M. Liu, P. Wang, C. Chen, PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8877–8886.

[61] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, Y. Wang, H.E. Pang, H. Mei, M. Zhang, L. Zhang, et al., Smpler-x: Scaling up expressive human pose and shape estimation, 2023, arXiv preprint arXiv:2309.17448.

[62] R. Li, Y. Xiu, S. Saito, Z. Huang, K. Olszewski, H. Li, Monocular real-time volumetric performance capture, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16, Springer, 2020, pp. 49–67.

[63] G. Wei, C. Lan, W. Zeng, Z. Chen, View invariant 3D human pose estimation, IEEE Trans. Circuits Syst. Video Technol. 30 (12) (2019) 4601–4610.

[64] Y. Zhan, F. Li, R. Weng, W. Choi, Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13116–13125.

[65] K. Zhou, X. Han, N. Jiang, K. Jia, J. Lu, HEMlets posh: Learning part-centric heatmap triplets for 3D human pose and shape estimation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (6) (2021) 3000–3014.

[66] X. Zheng, X. Chen, X. Lu, A joint relationship aware neural network for single-image 3D human pose estimation, IEEE Trans. Image Process. 29 (2020) 4747–4758.

[67] L. Wu, Z. Yu, Y. Liu, Q. Liu, Limb pose aware networks for monocular 3d pose estimation, IEEE Trans. Image Process. 31 (2021) 906–917.

[68] Y. Xu, W. Wang, T. Liu, X. Liu, J. Xie, S.-C. Zhu, Monocular 3d pose estimation via pose grammar and data augmentation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2021) 6327–6344.

[69] M. Fisch, R. Clark, Orientation keypoints for 6D human pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2021) 10145–10158.

[70] J. Liu, H. Ding, A. Shahroudy, L.-Y. Duan, X. Jiang, G. Wang, A.C. Kot, Feature boosting network for 3D pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2019) 494–501.

[71] H. Ci, X. Ma, C. Wang, Y. Wang, Locally connected network for monocular 3D human pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (3) (2020) 1429–1442.

[72] Z. Zou, W. Tang, Modulated graph convolutional network for 3D human pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11477–11487.

[73] A. Zeng, X. Sun, L. Yang, N. Zhao, M. Liu, Q. Xu, Learning skeletal graph neural networks for hard 3d pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11436–11445.

[74] K. Zhai, Q. Nie, B. Ouyang, X. Li, S. Yang, HopFIR: Hop-wise GraphFormer with intragroup joint refinement for 3D human pose estimation, 2023, arXiv preprint arXiv:2302.14581.

[75] K. Iskakov, E. Burkov, V. Lempitsky, Y. Malkov, Learnable triangulation of human pose, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7718–7727.

[76] H. Qiu, C. Wang, J. Wang, N. Wang, W. Zeng, Cross view fusion for 3d human pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4342–4351.

[77] E. Remelli, S. Han, S. Honari, P. Fua, R. Wang, Lightweight multi-view 3d pose estimation through camera-disentangled representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6040–6049.

[78] Z. Zhang, C. Wang, W. Qiu, W. Qin, W. Zeng, Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild, Int. J. Comput. Vis. 129 (2021) 703–718.

[79] K. Bartol, D. Bojanić, T. Petković, T. Pribanić, Generalizable human pose triangulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11028–11037.

[80] D.C. Luvizon, D. Picard, H. Tabia, Consensus-based optimization for 3D human pose estimation in camera coordinates, Int. J. Comput. Vis. 130 (3) (2022) 869–882.
[81] Y. Kudo, K. Ogaki, Y. Matsui, Y. Odagiri, Unsupervised adversarial learning of 3d human pose from 2d joint locations, 2018, arXiv preprint arXiv:1803.08244.
[82] C.-H. Chen, A. Tyagi, A. Agrawal, D. Drover, R. Mv, S. Stojanov, J.M. Rehg, Unsupervised 3d pose estimation with geometric self-supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5714–5724.
[83] B. Wandt, J.J. Little, H. Rhodin, ElePose: Unsupervised 3D human pose estimation by predicting camera elevation and learning normalizing flows on 2D poses, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6635–6645.
[84] M. Kocabas, S. Karagoz, E. Akbas, Self-supervised learning of 3d human pose using multi-view geometry, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1077–1086.
[85] K. Wang, L. Lin, C. Jiang, C. Qian, P. Wei, 3D human pose machines with self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 42 (5) (2019) 1069–1082.
[86] J.N. Kundu, S. Seth, P. YM, V. Jampani, A. Chakraborty, R.V. Babu, Uncertainty-aware adaptation for self-supervised 3d human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20448–20459.
[87] G. Hua, H. Liu, W. Li, Q. Zhang, R. Ding, X. Xu, Weakly-supervised 3D human pose estimation with cross-view U-shaped graph convolutional network, IEEE Trans. Multimed. (2022).
[88] M. Gholami, B. Wandt, H. Rhodin, R. Ward, Z.J. Wang, AdaptPose: Cross-dataset adaptation for 3D human pose estimation by learnable motion generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13075–13085.
[89] D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli, 3D human pose estimation in video with temporal convolutions and semi-supervised training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7753–7762.
[90] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, Z. Ding, 3D human pose estimation with spatial and temporal transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11656–11665.
[91] B. Artacho, A. Savakis, Unipose+: A unified framework for 2d and 3d human pose estimation in images and videos, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2021) 9641–9653.
[92] W. Li, H. Liu, H. Tang, P. Wang, L. Van Gool, Mhformer: Multi-hypothesis transformer for 3d human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13147–13156.
[93] J. Zhang, Z. Tu, J. Yang, Y. Chen, J. Yuan, Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13232–13242.
[94] S. Honari, V. Constantin, H. Rhodin, M. Salzmann, P. Fua, Temporal representation learning on monocular videos for 3D human pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[95] X. Qian, Y. Tang, N. Zhang, M. Han, J. Xiao, M.-C. Huang, R.-S. Lin, HSTFormer: Hierarchical spatial-temporal transformers for 3D human pose estimation, 2023, arXiv preprint arXiv:2301.07322.
[96] Z. Tang, Z. Qiu, Y. Hao, R. Hong, T. Yao, 3D human pose estimation with spatio-temporal criss-cross attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4790–4799.
[97] Y. Sun, A.W. Dougherty, Z. Zhang, Y.K. Choi, C. Wu, MixSynthFormer: A transformer encoder-like structure with mixed synthetic self-attention for efficient human pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14884–14893.
[98] J. Wang, S. Yan, Y. Xiong, D. Lin, Motion guided 3d pose estimation from videos, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, Springer, 2020, pp. 764–780.
[99] J. Zhang, Y. Wang, Z. Zhou, T. Luan, Z. Wang, Y. Qiao, Learning dynamical human-joint affinity for 3d pose estimation in videos, IEEE Trans. Image Process. 30 (2021) 7914–7925.
[100] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, J. Luo, Anatomy-aware 3d human pose estimation with bone-based pose decomposition, IEEE Trans. Circuits Syst. Video Technol. 32 (1) (2021) 198–209.
[101] Y. Xue, J. Chen, X. Gu, H. Ma, H. Ma, Boosting monocular 3D human pose estimation with part aware attention, IEEE Trans. Image Process. 31 (2022) 4278–4291.
[102] Y. Cheng, B. Yang, B. Wang, W. Yan, R.T. Tan, Occlusion-aware networks for 3d human pose estimation in video, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 723–732.
[103] Z. Yu, B. Ni, J. Xu, J. Wang, C. Zhao, W. Zhang, Towards alleviating the modeling ambiguity of unsupervised monocular 3d human pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8651–8660.
[104] X. Chen, K.-Y. Lin, W. Liu, C. Qian, L. Lin, Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10895–10904.
[105] R. Mitra, N.B. Gundavarapu, A. Sharma, A. Jain, Multiview-consistent semi-supervised learning for 3d human pose estimation, in: Proceedings of the Ieee/Cvf Conference on Computer Vision and Pattern Recognition, 2020, pp. 6907–6916.
[106] J.N. Kundu, S. Seth, V. Jampani, M. Rakesh, R.V. Babu, A. Chakraborty, Self-supervised 3d human pose estimation via part guided novel image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6152–6162.
[107] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, W. Gao, P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V, Springer, 2022, pp. 461–478.
[108] K. Gong, J. Zhang, J. Feng, Poseaug: A differentiable pose augmentation framework for 3d human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8575–8584.
[109] J. Zhang, K. Gong, X. Wang, J. Feng, Learning to augment poses for 3D human pose estimation in images and videos, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[110] L. Chen, H. Ai, R. Chen, Z. Zhuang, S. Liu, Cross-view tracking for multi-human 3d pose estimation at over 100 fps, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3279–3288.
[111] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, C. Lu, Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[112] S. Wu, S. Jin, W. Liu, L. Bai, C. Qian, D. Liu, W. Ouyang, Graph-based 3d multi-person pose estimation using multi-view images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11148–11157.
[113] G. Moon, J.Y. Chang, K.M. Lee, Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10133–10142.
[114] M. Fabbri, F. Lanzi, S. Calderara, S. Alletto, R. Cucchiara, Compressed volumetric heatmaps for multi-person 3d pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7204–7213.
[115] C. Wang, J. Li, W. Liu, C. Qian, C. Lu, Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 242–259.
[116] J. Zhen, Q. Fang, J. Sun, W. Liu, W. Jiang, H. Bao, X. Zhou, Smap: Single-shot multi-person absolute 3d pose estimation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16, Springer, 2020, pp. 550–566.
[117] A. Benzine, B. Luvison, Q.C. Pham, C. Achard, Single-shot 3D multi-person pose estimation in complex images, Pattern Recognit. 112 (2021) 107534.
[118] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, C. Theobalt, Single-shot multi-person 3d pose estimation from monocular rgb, in: 2018 International Conference on 3D Vision, 3DV, IEEE, 2018, pp. 120–130.
[119] G. Rogez, P. Weinzaepfel, C. Schmid, Lcr-net++: Multi-person 2d and 3d pose detection in natural images, IEEE Trans. Pattern Anal. Mach. Intell. 42 (5) (2019) 1146–1161.
[120] L. Jin, C. Xu, X. Wang, Y. Xiao, Y. Guo, X. Nie, J. Zhao, Single-stage is enough: Multi-person absolute 3D pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13086–13095.
[121] Y. Cheng, B. Wang, R.T. Tan, Dual networks based 3d multi-person pose estimation from monocular video, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2022) 1636–1651.
[122] Z. Tang, Y. Hao, J. Li, R. Hong, FTCM: Frequency-temporal collaborative module for efficient 3D human pose estimation in video, IEEE Trans. Circuits Syst. Video Technol. (2023).
[123] B. Artacho, A. Savakis, Unipose: Unified human pose estimation in single images and videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7035–7044.
[124] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, C. Sminchisescu, Deep network for the integrated 3d sensing of multiple people in natural images, in: Advances in Neural Information Processing Systems, vol. 31, 2018.
[125] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, Springer, 2016, pp. 483–499.
[126] Y. Rong, Z. Liu, C. Li, K. Cao, C.C. Loy, Delving deep into hybrid annotations for 3d human recovery in the wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5340–5348.

[127] Z. Li, B. Xu, H. Huang, C. Lu, Y. Guo, Deep two-stream video inference for human body pose and shape estimation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 430–439.

[128] K. Yang, R. Gu, M. Wang, M. Toyoura, G. Xu, LASOR: Learning accurate 3D human pose and shape via synthetic occlusion-aware data and neural mesh rendering, IEEE Trans. Image Process. 31 (2022) 1938–1948.

[129] Z. Li, J. Liu, Z. Zhang, S. Xu, Y. Yan, Cliff: Carrying location information in full frames into human pose and shape estimation, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V, Springer, 2022, pp. 590–606.

[130] M. Kocabas, C.-H.P. Huang, O. Hilliges, M.J. Black, PARE: Part attention regressor for 3D human body estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11127–11137.

[131] K. Lin, L. Wang, Z. Liu, Mesh graphormer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12939–12948.

[132] W.-L. Wei, J.-C. Lin, T.-L. Liu, H.-Y.M. Liao, Capturing humans in motion: temporal-attentive 3D human pose and shape estimation from monocular video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13211–13220.

[133] Z. Qiu, Q. Yang, J. Wang, H. Feng, J. Han, E. Ding, C. Xu, D. Fu, J. Wang, PSVT: End-to-end multi-person 3D pose and shape estimation with progressive video transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21254–21263.

[134] J. Cho, K. Youwang, T.-H. Oh, Cross-attention of disentangled modalities for 3D human mesh recovery with transformers, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I, Springer, 2022, pp. 342–359.

[135] Y. Xue, J. Chen, Y. Zhang, C. Yu, H. Ma, H. Ma, 3D human mesh reconstruction by learning to sample joint adaptive tokens for transformers, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6765–6773.

[136] K. Lin, L. Wang, Z. Liu, End-to-end human pose and mesh reconstruction with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1954–1963.

[137] A. Kanazawa, J.Y. Zhang, P. Felsen, J. Malik, Learning 3d human dynamics from video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5614–5623.

[138] M. Kocabas, N. Athanasiou, M.J. Black, Vibe: Video inference for human body pose and shape estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5253–5263.

[139] H. Choi, G. Moon, J.Y. Chang, K.M. Lee, Beyond static features for temporally consistent 3d human pose and shape from a video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1964–1973.

[140] Z. Wan, Z. Li, M. Tian, J. Liu, S. Yi, H. Li, Encoder-decoder with multi-level attention for 3D human shape and pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13033–13042.

[141] Z. Wang, S. Ostadabbas, Live stream temporally embedded 3D human body pose and shape estimation, 2022, arXiv preprint arXiv:2207.12537.

[142] X. Shen, Z. Yang, X. Wang, J. Ma, C. Zhou, Y. Yang, Global-to-local modeling for video-based 3D human pose and shape estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8887–8896.

[143] Z. Dong, J. Song, X. Chen, C. Guo, O. Hilliges, Shape-aware multi-person pose estimation from multi-view images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11158–11168.

[144] A. Sengupta, I. Budvytis, R. Cipolla, Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16094–16104.

[145] L. Zhuo, J. Cao, Q. Wang, B. Zhang, L. Bo, Towards stable human pose estimation via cross-view fusion and foot stabilization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 650–659.

[146] T. Fan, K.V. Alwala, D. Xiang, W. Xu, T. Murphey, M. Mukadam, Revitalizing optimization for 3d human pose and shape estimation: A sparse constrained formulation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11457–11466.

[147] J. Zhang, D. Yu, J.H. Liew, X. Nie, J. Feng, Body meshes as points, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 546–556.

[148] C. Zheng, M. Mendieta, T. Yang, C. Chen, HeatER: An efficient and unified network for human reconstruction via heatmap-based transformer, 2022, arXiv preprint arXiv:2205.15448.

[149] Z. Dou, Q. Wu, C. Lin, Z. Cao, Q. Wu, W. Wan, T. Komura, W. Wang, Tore: Token reduction for efficient human mesh recovery with transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15143–15155.

[150] G. Pavlakos, N. Kolotouros, K. Daniilidis, Texturepose: Supervising human mesh estimation with texture consistency, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 803–812.

[151] H. Zhang, J. Cao, G. Lu, W. Ouyang, Z. Sun, Learning 3d human shape and pose from dense body parts, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2020) 2610–2627.

[152] W. Zeng, W. Ouyang, P. Luo, W. Liu, X. Wang, 3D human mesh regression with dense correspondence, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7054–7063.

[153] T. Zhang, B. Huang, Y. Wang, Object-occluded human shape and pose estimation from a single color image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7376–7385.

[154] Y. Sun, Q. Bao, W. Liu, Y. Fu, M.J. Black, T. Mei, Monocular, one-stage, regression of multiple 3d people, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11179–11188.

[155] H. Choi, G. Moon, J. Park, K.M. Lee, Learning to estimate robust 3D human mesh from in-the-wild crowded scenes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1475–1484.

[156] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, Y. Wang, Motionbert: A unified perspective on learning human motion representations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15085–15099.

[157] R.A. Guler, I. Kokkinos, Holopose: Holistic 3d human reconstruction in-the-wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10884–10894.

[158] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, T. Mei, Human mesh recovery from monocular images via a skeleton-disentangled representation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5349–5358.

[159] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, C. Lu, Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3383–3393.

[160] J. Li, S. Bian, Q. Liu, J. Tang, F. Wang, C. Lu, NIKI: Neural inverse kinematics with invertible neural networks for 3D human pose and shape estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12933–12942.

[161] G.-H. Lee, S.-W. Lee, Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12375–12384.

[162] A. Sengupta, I. Budvytis, R. Cipolla, Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11219–11229.

[163] D. Wang, S. Zhang, 3D human mesh recovery with sequentially global rotation estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14953–14962.

[164] N. Kolotouros, G. Pavlakos, M.J. Black, K. Daniilidis, Learning to reconstruct 3D human pose and shape via model-fitting in the loop, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2252–2261.

[165] Y. Wang, K. Daniilidis, Refit: Recurrent fitting network for 3d human recovery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14644–14654.

[166] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, K. Daniilidis, Coherent reconstruction of multiple humans from a single image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5579–5588.

[167] M. Madadi, H. Bertiche, S. Escalera, Deep unsupervised 3D human body reconstruction from a sparse set of landmarks, Int. J. Comput. Vis. 129 (8) (2021) 2499–2512.

[168] S. Guan, J. Xu, M.Z. He, Y. Wang, B. Ni, X. Yang, Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2022) 5070–5086.

[169] B. Huang, T. Zhang, Y. Wang, Pose2UV: Single-shot multiperson mesh recovery with deep UV prior, IEEE Trans. Image Process. 31 (2022) 4679–4692.

[170] J. Li, Z. Yang, X. Wang, J. Ma, C. Zhou, Y. Yang, JOTR: 3D joint contrastive learning with transformers for occluded human mesh recovery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 9110–9121.

[171] H. Nam, D.S. Jung, Y. Oh, K.M. Lee, Cyclic test-time adaptation on monocular video for 3D human mesh reconstruction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14829–14839.

[172] T. Alldieck, M. Magnor, B.L. Bhatnagar, C. Theobalt, G. Pons-Moll, Learning to reconstruct people in clothing from a single RGB camera, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1175–1186.

[173] B.L. Bhatnagar, G. Tiwari, C. Theobalt, G. Pons-Moll, Multi-garment net: Learning to dress 3d people from images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5420–5430.

[174] T. Alldieck, G. Pons-Moll, C. Theobalt, M. Magnor, Tex2shape: Detailed full human body geometry from a single image, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2293–2303.

[175] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, H. Bao, Bcnet: Learning body and cloth shape from a single image, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, Springer, 2020, pp. 18–35.

[176] M.-P. Forte, P. Kulits, C.-H.P. Huang, V. Choutas, D. Tzionas, K.J. Kuchenbecker, M.J. Black, Reconstructing signing avatars from video using linguistic priors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12791–12801.

[177] B. Zhang, Y. Wang, X. Deng, Y. Zhang, P. Tan, C. Ma, H. Wang, Interacting two-hand 3d pose and shape reconstruction from single color image, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11354–11363.

[178] Y. Chen, Z. Tu, D. Kang, R. Chen, L. Bao, Z. Zhang, J. Yuan, Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion, IEEE Trans. Image Process. 30 (2021) 4008–4021.

[179] M. Hassan, V. Choutas, D. Tzionas, M.J. Black, Resolving 3D human pose ambiguities with 3D scene constraints, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2282–2292.

[180] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, M.J. Black, Monocular expressive body regression through body-driven attention, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16, Springer, 2020, pp. 20–40.

[181] Y. Rong, T. Shiratori, H. Joo, Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1749–1759.

[182] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, M.J. Black, Collaborative regression of expressive bodies using moderation, in: 2021 International Conference on 3D Vision, 3DV, IEEE, 2021, pp. 792–804.

[183] G. Moon, H. Choi, K.M. Lee, Accurate 3D hand pose estimation for whole-body 3D human mesh estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2308–2317.

[184] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, Y. Liu, Pymaf-x: Towards well-aligned full-body model regression from monocular images, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[185] J. Lin, A. Zeng, H. Wang, L. Zhang, Y. Li, One-stage 3D whole-body mesh recovery with component aware transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21159–21168.

[186] J. Li, S. Bian, C. Xu, Z. Chen, L. Yang, C. Lu, HybrIK-X: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery, 2023, arXiv preprint arXiv:2304.05690.

[187] D. Smith, M. Loper, X. Hu, P. Mavroidis, J. Romero, Facsimile: Fast and accurate scans from an image in less than a second, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5330–5339.

[188] S.S. Jinka, R. Chacko, A. Sharma, P. Narayanan, Peeledhuman: Robust shape representation for textured 3d human body reconstruction, in: 2020 International Conference on 3D Vision, 3DV, IEEE, 2020, pp. 879–888.

[189] Z. Zhang, L. Sun, Z. Yang, L. Chen, Y. Yang, Global-correlated 3d-decoupling transformer for clothed avatar reconstruction, 2023, arXiv preprint arXiv:2309.13524.

[190] Y. Xue, B.L. Bhatnagar, R. Marin, N. Sarafianos, Y. Xu, G. Pons-Moll, T. Tung, Nsf: Neural surface fields for human modeling from monocular depth, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15049–15060.

[191] E. Gärtner, M. Andriluka, E. Coumans, C. Sminchisescu, Differentiable dynamics for articulated 3d human motion reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13190–13200.

[192] Z. Dong, X. Chen, J. Yang, M.J. Black, O. Hilliges, A. Geiger, AG3D: Learning to generate 3D Avatars from 2D image collections, 2023, arXiv preprint arXiv:2305.02312.

[193] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, H. Li, Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2304–2314.

[194] S. Saito, T. Simon, J. Saragih, H. Joo, Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 84–93.

[195] Z. Huang, Y. Xu, C. Lassner, H. Li, T. Tung, Arch: Animatable reconstruction of clothed humans, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3093–3102.

[196] T. He, Y. Xu, S. Saito, S. Soatto, T. Tung, Arch++: Animation-ready clothed human reconstruction revisited, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11046–11056.

[197] T. Liao, X. Zhang, Y. Xiu, H. Yi, X. Liu, G.-J. Qi, Y. Zhang, X. Wang, X. Zhu, Z. Lei, High-fidelity clothed Avatar reconstruction from a single image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8662–8672.

[198] T. He, J. Collomosse, H. Jin, S. Soatto, Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction, Adv. Neural Inf. Process. Syst. 33 (2020) 9276–9287.

[199] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, X. Zhou, Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9054–9063.

[200] Y. Zhang, P. Ji, A. Wang, J. Mei, A. Kortylewski, A. Yuille, 3D-aware neural body fitting for occlusion robust 3D human pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 9399–9410.

[201] X. Gao, J. Yang, J. Kim, S. Peng, Z. Liu, X. Tong, Mps-nerf: Generalizable 3d human rendering from multiview images, IEEE Trans. Pattern Anal. Mach. Intell. (2022).

[202] L.G. Foo, J. Gong, H. Rahmani, J. Liu, Distribution-aligned diffusion for human mesh recovery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 9221–9232.

[203] H. Zhu, X. Zuo, S. Wang, X. Cao, R. Yang, Detailed human shape estimation from a single image by hierarchical mesh deformation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4491–4500.

[204] B.L. Bhatnagar, C. Sminchisescu, C. Theobalt, G. Pons-Moll, Combining implicit function learning and parametric models for 3d human reconstruction, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer, 2020, pp. 311–329.

[205] H. Zhu, X. Zuo, H. Yang, S. Wang, X. Cao, R. Yang, Detailed avatar recovery from single image, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2021) 7363–7379.

[206] Y. Xiu, J. Yang, D. Tzionas, M.J. Black, Icon: Implicit clothed humans obtained from normals, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2022, pp. 13286–13296.

[207] Y. Xiu, J. Yang, X. Cao, D. Tzionas, M.J. Black, ECON: Explicit clothed humans optimized via normal integration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 512–523.

[208] X. Zhang, J. Zhang, R. Chacko, H. Xu, G. Song, Y. Yang, J. Feng, Getavatar: Generative textured meshes for animatable human avatars, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2273–2282.

[209] D. Svitov, D. Gudkov, R. Bashirov, V. Lempitsky, Dinar: Diffusion inpainting of neural textures for one-shot human avatars, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7062–7072.

[210] X. Pan, Z. Yang, J. Ma, C. Zhou, Y. Yang, TransHuman: A transformer-based human representation for generalizable neural human rendering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3544–3555.

[211] Y. Liu, X. Huang, M. Qin, Q. Lin, H. Wang, Animatable 3D Gaussian: Fast and high-quality reconstruction of multiple human avatars, 2023, arXiv preprint arXiv:2311.16482.

[212] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

[213] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[214] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877–1901.

[215] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[216] Y. Xu, S.-C. Zhu, T. Tung, Denserac: Joint 3d pose and shape estimation by dense render-and-compare, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7760–7770.

[217] S. Guan, J. Xu, Y. Wang, B. Ni, X. Yang, Bilevel online adaptation for out-of-domain human mesh reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10472–10481.

[218] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, Z. Sun, Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11446–11456.

[219] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis, Commun. ACM 65 (1) (2021) 99–106.

[220] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, 3D Gaussian splatting for real-time radiance field rendering, ACM Trans. Graph. 42 (4) (2023).

[221] C. Yan, D. Qu, D. Wang, D. Xu, Z. Wang, B. Zhao, X. Li, GS-SLAM: Dense visual SLAM with 3D Gaussian splatting, 2023, arXiv preprint arXiv:2311.11700.

[222] X. Liu, X. Zhan, J. Tang, Y. Shan, G. Zeng, D. Lin, X. Liu, Z. Liu, HumanGaussian: Text-driven 3D human generation with Gaussian splatting, 2023, arXiv preprint arXiv:2311.17061.

[223] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, X. Wang, 4D Gaussian splatting for real-time dynamic scene rendering, 2023, arXiv preprint arXiv:2310.08528.

[224] Z. Chen, F. Wang, H. Liu, Text-to-3d using Gaussian splatting, 2023, arXiv preprint arXiv:2309.16585.

[225] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1325–1339, http://dx.doi.org/10.1109/TPAMI.2013.248.

[226] Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2878–2890, http://dx.doi.org/10.1109/TPAMI.2012.261.

[227] T. Von Marcard, R. Henschel, M.J. Black, B. Rosenhahn, G. Pons-Moll, Recovering accurate 3d human pose in the wild using imus and a moving camera, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 601–617.

[228] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3d human pose estimation in the wild using improved cnn supervision, in: 2017 International Conference on 3D Vision, 3DV, IEEE, 2017, pp. 506–516.

[229] L. Sigal, A.O. Balan, M.J. Black, Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, Int. J. Comput. Vis. 87 (1–2) (2010) 4.

[230] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, Y. Sheikh, Panoptic studio: A massively multiview system for social motion capture, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2015.

[231] G. Varol, J. Romero, X. Martin, N. Mahmood, M.J. Black, I. Laptev, C. Schmid, Learning from synthetic humans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 109–117.

[232] L. Muller, A.A. Osman, S. Tang, C.-H.P. Huang, M.J. Black, On self-contact and human pose, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9990–9999.

[233] N. Mahmood, N. Ghorbani, N.F. Troje, G. Pons-Moll, M.J. Black, AMASS: Archive of motion capture as surface shapes, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5442–5451.

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

[235] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M.J. Black, P.V. Gehler, Unite the people: Closing the loop between 3d and 2d human representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6050–6059.

[236] Z. Zheng, T. Yu, Y. Wei, Q. Dai, Y. Liu, Deephuman: 3d human reconstruction from a single image, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7739–7749.

[237] W. Zhao, W. Wang, Y. Tian, Graformer: Graph-oriented transformer for 3d pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20438–20447.

[238] B.X. Yu, Z. Zhang, Y. Liu, S.-h. Zhong, Y. Liu, C.W. Chen, Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8818–8829.

[239] H. Ci, M. Wu, W. Zhu, X. Ma, H. Dong, F. Zhong, Y. Wang, Gfpose: Learning 3d human pose prior with gradient fields, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4800–4810.

[240] K. Lee, W. Kim, S. Lee, From human pose similarity metric to 3D human pose estimator: Temporal propagating LSTM networks, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2022) 1781–1797.

[241] A. Zeng, X. Ju, L. Yang, R. Gao, X. Zhu, B. Dai, Q. Xu, Deciwatch: A simple baseline for 10× efficient 2d and 3d pose estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 607–624.

[242] J. Gong, L.G. Foo, Z. Fan, Q. Ke, H. Rahmani, J. Liu, Diffpose: Toward more reliable 3d pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13041–13051.

[243] K. Holmquist, B. Wandt, Diffpose: Multi-hypothesis human pose estimation using diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15977–15987.

[244] X. Ma, J. Su, C. Wang, W. Zhu, Y. Wang, 3D human mesh estimation from virtual markers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 534–543.

[245] J. Kim, M.-G. Gwon, H. Park, H. Kwon, G.-M. Um, W. Kim, Sampling is matter: Point-guided 3D human mesh reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12880–12889.

[246] K. Shetty, A. Birkhold, S. Jaganathan, N. Strobel, M. Kowarschik, A. Maier, B. Egger, PLIKS: A pseudo-linear inverse kinematic solver for 3D human body estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 574–584.

[247] Q. Fang, K. Chen, Y. Fan, Q. Shuai, J. Li, W. Zhang, Learning analytical posterior probability for human mesh recovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8781–8791.

[248] C. Zheng, X. Liu, G.-J. Qi, C. Chen, POTTER: Pooling attention transformer for efficient human mesh recovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1611–1620.

[249] Q. Liu, A. Kortylewski, A.L. Yuille, PoseExaminer: Automated testing of out-of-distribution robustness in human pose and shape estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 672–681.

[250] H. Cho, Y. Cho, J. Ahn, J. Kim, Implicit 3D human mesh recovery using consistency with pose and shape from unseen-view, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21148–21158.

[251] T. Simon, H. Joo, I. Matthews, Y. Sheikh, Hand keypoint detection in single images using multiview bootstrapping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1145–1153.

[252] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[253] K. Aberman, P. Li, D. Lischinski, O. Sorkine-Hornung, D. Cohen-Or, B. Chen, Skeleton-aware networks for deep motion retargeting, ACM Trans. Graph. 39 (4) (2020) 1–62.

[254] Z. Yang, W. Zhu, W. Wu, C. Qian, Q. Zhou, B. Zhou, C.C. Loy, Transmomo: Invariance-driven unsupervised video motion retargeting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5306–5315.

[255] W.-Y. Yu, L.-M. Po, R.C. Cheung, Y. Zhao, Y. Xue, K. Li, Bidirectionally deformable motion modulation for video-based human pose transfer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7502–7512.

[256] T.L. Gomes, R. Martins, J. Ferreira, R. Azevedo, G. Torres, E.R. Nascimento, A shape-aware retargeting approach to transfer human motion and appearance in monocular videos, Int. J. Comput. Vis. 129 (7) (2021) 2057–2075.

[257] W. Zhu, Z. Yang, Z. Di, W. Wu, Y. Wang, C.C. Loy, Mocanet: Motion retargeting in-the-wild via canonicalization networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, (no. 3) 2022, pp. 3617–3625.

[258] L. Mo, H. Li, C. Zou, Y. Zhang, M. Yang, Y. Yang, M. Tan, Towards accurate facial motion retargeting with identity-consistent and expression-exclusive constraints, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, (no. 2) 2022, pp. 1981–1989.

[259] X. Chen, M. Mihajlovic, S. Wang, S. Prokudin, S. Tang, Morphable diffusion: 3D-consistent diffusion for single-image avatar creation, 2024, arXiv preprint arXiv:2401.04728.

[260] Z. Su, L. Hu, S. Lin, H. Zhang, S. Zhang, J. Thies, Y. Liu, Caphy: Capturing physical properties for animatable human avatars, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14150–14160.

[261] Z. Luo, J. Cao, K. Kitani, W. Xu, et al., Perpetual humanoid control for real-time simulated avatars, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10895–10904.

[262] H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li, S.J. Maybank, Feedback graph convolutional network for skeleton-based action recognition, IEEE Trans. Image Process. 31 (2021) 164–175.

[263] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, M. Chiaberge, Action transformer: A self-attention model for short-time pose-based human action recognition, Pattern Recognit. 124 (2022) 108487.

[264] Z. Lu, H. Wang, Z. Chang, G. Yang, H.P. Shum, Hard no-box adversarial attack on skeleton-based human action recognition with skeleton-motion-informed gradient, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4597–4606.

[265] C. Bian, W. Feng, L. Wan, S. Wang, Structural knowledge distillation for efficient skeleton-based action recognition, IEEE Trans. Image Process. 30 (2021) 2963–2976.

[266] D.C. Luvizon, D. Picard, H. Tabia, Multi-task deep learning for real-time 3D human pose estimation and action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (8) (2020) 2752–2764.

[267] Q. Bao, W. Liu, Y. Cheng, B. Zhou, T. Mei, Pose-guided tracking-by-detection: Robust multi-person pose tracking, IEEE Trans. Multimed. 23 (2020) 161–175.

[268] N.D. Reddy, L. Guigues, L. Pishchulin, J. Eledath, S.G. Narasimhan, Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15190–15200.

[269] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, J. Malik, Humans in 4D: Reconstructing and tracking humans with transformers, 2023, arXiv preprint arXiv:2305.20091.

[270] Y. Sun, Q. Bao, W. Liu, T. Mei, M.J. Black, TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8856–8866.
[271] Y. Dai, C. Wen, H. Wu, Y. Guo, L. Chen, C. Wang, Indoor 3D human trajectory reconstruction using surveillance camera videos and point clouds, IEEE Trans. Circuits Syst. Video Technol. 32 (4) (2021) 2482–2495.
[272] M. Kocabas, Y. Yuan, P. Molchanov, Y. Guo, M.J. Black, O. Hilliges, J. Kautz, U. Iqbal, PACE: Human and camera motion estimation from in-the-wild videos, 2023, arXiv preprint arXiv:2310.13768.
[273] A. Habibian, D. Abati, T.S. Cohen, B.E. Bejnordi, Skip-convolutions for efficient video processing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2695–2704.
[274] Y. Tay, M. Dehghani, D. Bahri, D. Metzler, Efficient transformers: A survey, ACM Comput. Surv. 55 (6) (2022) 1–28.
[275] L.G. Foo, J. Gong, Z. Fan, J. Liu, System-status-aware adaptive network for online streaming video understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10514–10523.
[276] R. Anil, A.M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al., Palm 2 technical report, 2023, arXiv preprint arXiv:2305.10403.
[277] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, 2023, arXiv preprint arXiv:2303.08774.
[278] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, 2023, arXiv preprint arXiv:2304.02643.
[279] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, F. Zheng, Track anything: Segment anything meets videos, 2023, arXiv preprint arXiv:2304.11968.
[280] Y. Ci, Y. Wang, M. Chen, S. Tang, L. Bai, F. Zhu, R. Zhao, F. Yu, D. Qi, W. Ouyang, UniHCP: A unified model for human-centric perceptions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17840–17852.
[281] Y. Feng, J. Lin, S.K. Dwivedi, Y. Sun, P. Patel, M.J. Black, PoseGPT: Chatting about 3D human pose, 2023, arXiv preprint arXiv:2311.18836.
[282] H. Yi, C.-H.P. Huang, D. Tzionas, M. Kocabas, M. Hassan, S. Tang, J. Thies, M.J. Black, Human-aware object placement for visual environment reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3959–3970.
[283] R. Jiang, C. Wang, J. Zhang, M. Chai, M. He, D. Chen, J. Liao, Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14371–14382.

Yang Liu is currently pursuing his Ph.D. degree in the School of Electronics and Communication Engineering, Sun Yat-sen University, China. His research interests include human-centric computer vision and real-time image processing.

Changzhen Qiu received a Ph.D. from the National University of Defense Technology (NUDT), China, in 2015. He works in the School of Electronics and Communication Engineering at Sun Yat-sen University. His research interests include image processing, target detection and target tracking.

Zhiyong Zhang received a Ph.D. from the National University of Defense Technology (NUDT), China, in 2006. He currently works as a professor at the School of Electronics and Communication Engineering, Sun Yat-sen University. His research interests include pattern recognition, machine/deep learning, and computer vision.