



A survey on deep 3D human pose estimation

Rama Bastola Neupane¹ · Kan Li¹ · Tesfaye Fenta Boka¹

Accepted: 6 November 2024 / Published online: 25 November 2024
© The Author(s) 2024

Abstract

3D Human Pose Estimation (3D-HPE) is a highly active and evolving research area in computer vision with numerous applications such as extended reality, action recognition, and video surveillance. The field has significantly advanced with deep learning, public datasets, and enhanced computational power, addressing challenges like depth ambiguity, occlusion, and data scarcity. Researchers confront scenario-specific issues such as ill-posed problems in monocular setups, cross-view aggregation with camera synchronizations in multi-view systems, and inter-person occlusion in multi-person scenarios. This survey comprehensively reviews contemporary strategies covering a technological spectrum including Convolutional Neural Networks, Graph Convolutional Networks, Transformers, and their combinations employed to address these challenges. It includes scenarios such as monocular and multi-view setups, single and multi-person cases, as well as image and video inputs. The survey explores various solution paradigms, including single-stage vs 2D-to-3D lifting, absolute vs relative keypoints, pixel vs voxel vs Neural Radiance Field spaces, and deterministic, probabilistic, or diffusion-based strategies, along with top-down vs bottom-up approaches. It examines advanced learning techniques beyond supervised methods and data augmentation for diverse pose datasets. It analyzes the performance of recent methods on benchmark datasets for different scenarios. Challenges are categorized into common and scenario-specific issues, and future research directions are proposed to foster further advancements in the field. Additionally, key sections are summarized in tables or visual formats for quick understanding. This survey is a valuable resource and a solid reference for researchers in the dynamic landscape of 3D human pose estimation.

Keywords 3D human pose estimation · Single-person · Multi-person · Data augmentation · Learning techniques · Scenario-specific challenges

✉ Kan Li
likan@bit.edu.cn
Rama Bastola Neupane
rama@bit.edu.cn
Tesfaye Fenta Boka
tfenta@bit.edu.cn

¹ School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

1 Introduction

3D Human Pose Estimation (3D-HPE) is one of the most active and dynamic fields in computer vision aiming to predict 3D coordinates of articulated human joints from images or video frames (Zhang et al. 2023a; Shuai et al. 2023; Zhang et al. 2023b). When a 3D scene with a person is projected onto a 2D image, both the positional and depth information of all joints are lost. The ultimate aim of 3D pose estimation is to accurately determine their 2D positions in the images and their depth within the camera's coordinate system (Ci et al. 2022). Applications of 3D-HPE are diverse, spanning human-computer interaction (Huo et al. 2023), action recognition (Rajasegaran et al. 2023), extended reality (XR) (Zhang 2012), autonomous vehicles (Lamas et al. 2022), video surveillance (Matsukawa et al. 2020), gaming (Tuyls et al. 2021), and medical fields (Liu et al. 2023). The performance of these applications relies heavily on the precision of 3D-HPE, making it a vital area of study in computer vision. The advancements in deep learning techniques, availability of large-scale 3D pose datasets such as Human3.6M (Ionescu et al. 2014), and computational resources have accelerated research in this area, leading to significant contributions from academia and industry.

The 3D-HPE problem depends on the number of sensors capturing the images and individuals in the scene, as both affect estimation performance. Based on the number of cameras, input data can be either monocular (single camera) or multi-view (multiple cameras), hence the study of 3D-HPE can be categorized into four scenarios: Monocular images, Monocular videos, Multi-view images, and Multi-view videos. These scenarios can further involve single-person or multi-person pose estimation, leading to eight distinct scenarios. Consequently, each scenario presents unique challenges alongside common issues such as depth ambiguity, occlusion, inadequate data, changes in appearance (such as clothing and lighting), and complexity of the articulated human body. Capturing an RGB image is much easier and more affordable in uncontrolled environments, however, depth ambiguity and self-occlusion are major issues in monocular images (Ci et al. 2022; Xu et al. 2022). Monocular video requires methods to extract and utilize temporal information (Zheng et al. 2021; Honari et al. 2023) from frame-to-frame relationships for better pose estimation. Multi-view Images are better at handling self-occlusion and depth ambiguity through triangulation or similar techniques (Bartol et al. 2022; Wan et al. 2023), however, they require precise camera calibration and synchronization (Shuai et al. 2023).

Moreover, we cannot directly extend the methods designed for single-person to multi-persons due to new challenges in multi-person scenarios such as intra-person occlusion and requirements of camera-centric coordinates (Cheng et al. 2023). Researchers have developed various methods to address these challenges, often leveraging deep learning architectures such as convolutional neural networks (CNNs) (Li et al. 2020b), graph convolution networks (GCN) (Kipf and Welling 2017), transformers (Vaswani et al. 2017), and their combinations. Moreover, online/offline data augmentation techniques and learning techniques beyond supervision have been utilized to handle the data scarcity problem making the model more adaptive to natural scenarios. Continuous advancements in this field will likely lead to more accurate and efficient models, benefiting various technological and scientific domains. This survey provides a comprehensive overview of the current state of 3D-HPE techniques, highlighting common and scenario-specific challenges, and future research directions.

1.1 Previous survey

The early review work (Sarafianos et al. 2016) offers a taxonomy of methods depending on input types such as images or image sequences and single or multiple views scenarios. It covers both traditional and a few deep learning-based approaches. The study also introduces a novel synthetic dataset, SynPose300, and evaluates several state-of-the-art techniques on this dataset. The review work (Gamra and Akhloufi 2021) categorizes 2D single-person pipelines into regression-based and detection-based methods and 3D approaches into one-stage and two-stage methodologies. It also examines multi-person pipelines through top-down and bottom-up approaches. The survey (Desmarais et al. 2021) reviews 3D-HPE techniques based on the network backbone architecture and input types. The performance of those methods is compared based on their accuracy, speed, and robustness. The work (Wang et al. 2021) reviews existing deep learning approaches for 3D-HPE across a spectrum of input types, number of individuals, and direct 3D estimation versus lifting from 2D to 3D pose. The review work (Liu et al. 2022) focuses on monocular 2D and 3D HPE. The work (Yan et al. 2022) reviews the 3D HPE from classical methods to deep learning methods and includes learning methods. The survey (Tian et al. 2023) centers on categorizing monocular 3D human mesh recovery techniques according to their design paradigms, reconstruction detail, and application contexts. The review (Zheng et al. 2023) examines contemporary deep learning approaches for 2D and 3D pose estimation and compares methods based on their input data and inference procedures.

The field of 3D-HPE is advancing quickly, requiring frequent reviews of new methods and trends. Survey papers are essential for understanding the progress in this domain, as they offer comprehensive overviews of techniques available at the time of their publication. This work aims to deliver a thorough and well-structured review of the latest developments in 3D-HPE, building on previous surveys and providing valuable insights into recent technological advancements.

1.2 Scope of the study

Most survey papers highlight common challenges and suggest future directions in the field. While data scarcity and generalization issues are frequently mentioned among these challenges, there is often little focus on the strategies used by existing techniques to address these problems. Our review addresses this gap by exploring and organizing solutions such as advanced learning algorithms beyond supervised learning and online/offline pose data augmentation techniques. This structured presentation aims to provide readers with a clear and practical understanding of how these strategies can mitigate the issues of data scarcity and generalization.

With the advancement of machine learning algorithms, problem-solving approaches have shifted from CNN-based methods to hybrid algorithms such as graph-transformers, solution spaces from pixel-based to voxel/ NeRF (Neural Radiance Fields), deterministic to probabilistic multi-hypotheses and recently diffusion-based approaches. We focus on these latest trends and perspectives, thoroughly reviewing recent papers from top journals and conferences in computer vision, covering the period from 2020 to March 2024. A few relevant studies beyond this period are also included. Our review includes comprehensive summaries of key sections, presented with tabular or visual illustrations for clarity.

Additionally, we have extensively studied existing problems, categorizing them into common challenges and scenario-specific issues, and providing future directions with actionable recommendations for research. We believe these insights will significantly benefit readers and help them make rapid progress in the outlined areas. Figure 1 illustrates the focus area of this survey, which centers on estimating 3D human pose from RGB images. The main contributions of this survey are summarized as follows:

1. Unlike previous works, we begin by examining various recent problem-solving strategies (Section 2) and then review cutting-edge deep learning approaches for different 3D-HPE scenarios (Sections 3 and 4). This approach provides a comprehensive understanding of the rapidly evolving field of 3D-HPE.
2. We specifically explore learning strategies beyond traditional supervision, including semi-, weakly-, unsupervised, and self-supervised approaches. Additionally, we review pose data augmentation techniques into offline and online categories. These approaches are highly regarded by researchers for addressing the common challenges of data inadequacy and diversity in 3D-HPE.
3. We present the performance of state-of-the-art methods across benchmark datasets, highlighting comparisons of different deep learning algorithms across various scenarios. We uniquely categorize the challenges into common and scenario-specific issues, offering future directions with actionable recommendations. The rest of this survey is organized as follows: Sect. 2 categorizes 3D-HPE strategies based on problem-solving stages, joint keypoint coordinates, multi-person paradigms, probabilistic/diffusion-based approaches, and solution spaces. Sections 3 and 4 summarize state-of-the-art techniques for single-person and multi-person scenarios, respectively. Section 5 describes learning techniques beyond supervised learning, while Sect. 6 presents benchmark datasets, online/offline pose data augmentation techniques and evaluation metrics. Section 7 discusses the performance of various categories of 3D HPE models on benchmark datasets. Section 8 categorizes challenges in the field of 3D-HPE into common and scenario-specific challenges and suggests potential future directions. Finally, Sect. 9 concludes the study.

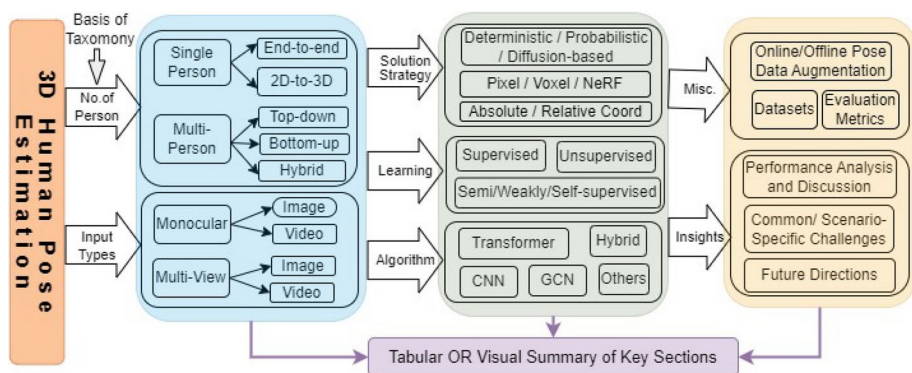


Fig. 1 Taxonomic landscape of 3D human pose estimation survey

2 3D-HPE problem-solving strategies

This section examines the problem-solving strategies utilized by current methods for 3D-HPE. These methods are analyzed based on several factors: the number of stages in the solution process, which includes end-to-end single-stage approaches or two-stage methods involving 2D-to-3D lifting; the coordinate system for keypoint estimation, whether absolute or relative; multi-person paradigms that address scenarios with multiple individuals; deterministic or probabilistic and diffusion-based approaches; the problem-solving space, which can involve pixel space, voxel space, or NeRF; and emerging trends and needs. Before delving into the technical specifics, Table 1 provides brief definitions of key terms to better understand the concept.

2.1 Problem-solving stage

Based on problem-solving stages, the existing 3D-HPE methods are categorized into single-stage or end-to-end and two-stage or 2D to 3D lifting paradigms. End-to-end learning approaches aim to directly estimate 3D poses from images using methods such as coordinate regression, nearest neighbor matching between images and poses, or classification over a set of pose classes. In the two-stage approach, 2D pose-aware features or directly the 2D human pose are utilized as intermediate inputs to estimate 3D coordinates. The field of 2D-HPE has made significant advancements due to the availability of large-scale datasets and advancements in deep learning techniques. Building on these developments, researchers have applied state-of-the-art 2D-HPE networks, such as Stack Hourglass (Newell et al. 2016), Cascaded Pyramid Network (CPN) (Chen et al. 2018), and High-Resolution Network (HRNet) (Sun et al. 2019), to recover a 3D human pose. The two-stage strategy mitigates the risk of overfitting to small 3D datasets and enables the utilization of existing 2D pose datasets such as MPII (Andriluka et al. 2014) and COCO (Lin et al. 2014). Additional image information such as texture and semantic details can aid in accurately predicting 3D poses from 2D poses (Zhou et al. 2024a).

2.1.1 Single-stage or End-to-end

Umar et al. (2020) propose to utilize multi-view data to ensure spatial consistency without costly 3D pose annotations for 3D-HPE from monocular images. LCR-Net++ (Rogez et al. 2020) and UniPose+ (Artacho and Savakis 2022) present end-to-end methods for 2D and 3D pose estimation. These methods incorporate depth regression directly into the pose estimation network. Wang et al. (2021b) introduce a Multi-view Pose Transformer model for multi-person 3D pose estimation. This model uses a direct regression method with a hierarchical joint query embedding scheme and projective attention mechanism. Despite its superior performance and speed, it struggles with data scarcity and varying camera setups. Reddy et al. (2021) address this issue with TesseTrack, which simultaneously addresses 3D joint reconstruction and person association in space and time using a spatio-temporal formulation in a common voxelized feature space. Honari et al. (2023) employ contrastive self-supervised learning for single-stage human pose estimation from RGB video. The research work (Luvizon et al. 2023) addresses scalability issues that employ a scalable sequential

Table 1 Glossary of key terminologies

Key terms	Brief concepts
Convolutional Neural Networks (CNNs)	Mostly used to extract features from 2D images or video frames to predict the 2D location of body keypoints or estimate depth maps, which are then used to infer 3D poses
Data Augmentation	Involves artificially increasing the size and diversity of the dataset by applying transformations to images, such as rotation, flipping, scaling, or altering keypoint positions. This technique enhances the ability of the model to generalize, enabling pose estimation models to effectively manage different human poses, body shapes, and camera perspectives
Diffusion Models	These are generative models that synthesize human poses by progressively refining random noise into a plausible human skeleton. These models aim to predict missing pose data or generate diverse pose samples
Graph Convolutional Networks (GCNs)	Useful for modeling the skeleton as a graph where joints are nodes, and bones are edges, learning how the spatial configuration of these joints evolves. GCNs are particularly effective in inferring the 3D pose from 2D pose keypoints by learning joint relationships
Hybrid Approach	In 3D-HPE, a hybrid approach combines different methods (e.g., CNNs, GCNs, or transformers) to better capture both local and global relationships between body joints. Top-down and bottom-up approaches are also combined to improve the performance of multi-person pose estimation
Neural Radiance Field (NeRF)	NeRF is a neural rendering technique to model and synthesize realistic 3D scenes from 2D images. It works by learning a volumetric scene representation where each point in 3D space is associated with color and density. In 3D-HPE, NeRF can be used to represent a human body's volumetric scene in 3D by learning the appearance and geometry from a set of 2D images
Transformer	In 3D-HPE, transformers can model long-range dependencies between body joints. They use self-attention mechanisms to focus on how joints relate to each other spatially and temporally across video frames
Voxel	In 3D-HPE, voxels represent the 3D volume of the human body or parts of the body on a grid of 3D units called volumetric pixels (voxels). Each voxel contains information about whether it is occupied by part of the human body and may carry appearance information such as color or texture

pyramid network for regressing the pose at multiple scales using a sequential coarse-to-fine approach.

Pros and cons of single-stage paradigm: 2D pose estimates tend to be quite noisy in real-world applications, which can significantly and irreversibly affect the subsequent 3D pose estimation step (Zhang et al. 2023c). Single-stage methods, which directly estimate 3D poses from 2D images without relying on intermediate 2D estimators, avoid this issue. However, the focus of convolutional operations on local features and the lack of intermediate processing make these methods computationally intensive (Chen et al. 2024) and hence unsuitable for real-time processing. Additionally, single-stage methods require 3D annotations for training so the scarcity of large-scale labeled datasets in natural environments can lead to overfitting and poor generalization of the pose estimator. The absence of intermediate supervision also makes these methods susceptible to variations in background and lighting, and the complexity of features makes the learning process for a single model extremely difficult (Zhang et al. 2023b). Moreover, directly regressing from an image to a 3D human pose is a highly nonlinear problem, leading to a large estimation model, a vast solution search space, and a high likelihood of sub-optimal solutions. Therefore, lifting the predicted 2D human pose to a 3D pose has emerged as a viable direction within the two-stage methods in this field (Nie et al. 2023).

2.1.2 2D-to-3D lifting

Depending on the type of intermediate input used for 3D coordinate estimation, two-stage methods for 3D human pose estimation can be classified into 2D Pose-Aware Feature-Based and 2D Pose-Based lifting methods. 2D Pose-Aware Feature-Based Lifting methods utilize 2D pose-aware features derived from the detected 2D human pose, such as heatmaps, part affinity fields, or other keypoints descriptors. These features capture the spatial and contextual information from the 2D pose, which is then used to estimate the 3D coordinates (Pavlakos et al. 2017; Wang et al. 2020; Liu et al. 2020a; Kundu et al. 2020a; Remelli et al. 2020; Benzine et al. 2020; Wu and Xiao 2020; Tu et al. 2020). The work (Liu et al. 2021) uses 2D poses as input and employs an attention-based temporal convolutional neural network to estimate 3D poses. Zhan et al. (2022) transform the input from 2D-pixel space to 3D rays in a normalized coordinate system, mitigating variations due to changes in camera intrinsic parameters and pitch angle.

Some research works (Wandt et al. 2021; Cheng et al. 2021a; Wehrbein et al. 2021; Kundu et al. 2022; Usman et al. 2022) incorporate 2D pose confidence scores and heatmaps to estimate 3D poses. Such methods use high-confidence joints to correct the positions of low-confidence, uncertain joints, enhancing overall prediction accuracy. Kim et al. (2024) utilize the confidence levels of joints in 2D poses to create pseudo-labels. The model generates 3D pseudo ground-truths by calculating a weighted average for each joint based on the confidence of the 2D poses, which are then used for self-supervised learning.

2D Pose-Based Lifting methods (Zhou et al. 2017; Pavlo et al. 2019; Liu et al. 2020b; Shuai et al. 2023) directly use the 2D human pose, consisting of the detected keypoints or joints in 2D space, as the intermediate input. The spatial configuration of these 2D joints is leveraged to infer the corresponding 3D positions, aiming to reconstruct the 3D pose from the 2D joint locations. The research work (Xu et al. 2020) refines 2D poses using perspective projection and estimates 3D poses by applying kinematic constraints. Zhang et

al. (2020) leverage pose geometry, including poses and viewpoints, to lift 2D poses into 3D poses. AdaFuse (Zhang et al. 2021) utilizes a 2D pose method to generate heatmaps for each view and fuse them using epipolar geometry prior to estimating the 3D pose.

Pros and cons of 2D-to-3D lifting: The task of estimating 2D human poses from images is a well-solved problem (Nie et al. 2023), providing a clear visualization of joint positions and aligning well spatially with the 3D pose, making it robust against image distortions such as varying backgrounds, clothing, and colors, given sufficient ground-truth (GT) data. This approach can significantly reduce uncertainty and semantic loss (Lee et al. 2023). Additionally, there are many 2D labels available, especially in-the-wild, because labeling 2D poses is easier and cheaper than capturing 3D ground truth (Nie et al. 2023). The widespread availability of 2D human pose detectors and the lightweight nature of 2D skeleton representations have made lifting-based methods prevalent in 3D pose estimation (Zhao et al. 2023b). However, the 2D-to-3D pipeline involves several variable factors, such as the number of views, video sequence length, and whether camera calibration is used (Shuai et al. 2023). Moreover, the 2D-to-3D approach inherits errors from the 2D model and increases computational complexity by sequentially calculating depth in a second step, rather than estimating depth concurrently.

2.2 Joint keypoint coordinates

Existing 3D-HPE methods predict 3D poses either relative to the root body joints or the camera coordinates. Based on the reference coordinate system, these methods are classified into person-centric and camera-centric. Some models estimate keypoints in both person-centric and camera-centric coordinates to compare their performance with state-of-the-art technologies and the datasets used for training. For instance, TesseTrack (Reddy et al. 2021) calculates the root-centered MPJPE metric on the Human3.6M dataset and the non-root-centered MPJPE on the CMU Panoptic dataset. It is relevant to estimate 3D pose in relative body coordinates for a single person, however, it requires absolute coordinate for multi-person from monocular images and global coordinate is meaningful for multi-person pose estimation from multiple camera views.

2.2.1 Person-centric

Person-centric, also known as relative coordinates or root-centric methods predict 3D coordinates for keypoints relative to a central body point, such as the pelvis, with this root joint centered at the origin and other joints estimated in relation to it. This approach is primarily useful for single-person pose estimation using monocular images. Various methods for estimating pose in root-centric coordinates for single-person 3D-HPE with monocular images are described in works (Zhang et al. 2023a; Ci et al. 2022; Zheng et al. 2021; Bartol et al. 2022; Cheng et al. 2021a; Wehrbein et al. 2021; Usman et al. 2022). Additionally, some studies (Honari et al. 2023; Liu et al. 2021; Lee et al. 2023; Choi et al. 2021a; Kundu et al. 2021) focus on video input to estimate 3D pose in person-centric coordinates. However, person-centric methods are less suitable for scenarios involving multiple people or multiple views because they lose the location of persons in the scene and do not know where to place them.

2.2.2 Camera-centric

Camera-centric or absolute coordinates methods predict 3D keypoints relative to the camera as a reference point. The distance between the camera and the person defines the person's size and keypoints, which is crucial in multi-person 3D HPE involving multi-view image or video input. For example, a person farther from the camera appears shorter than a person closer to the camera, even if the distant person is taller. Person-centric coordinates limit generalization for multi-view setups because relative poses predicted from one camera cannot be easily projected into a different view, complicating occlusion handling in multi-view scenarios. In contrast, when estimations are made relative to a static reference, predictions can be easily projected from one view to another (Luvizon et al. 2022). For multiple camera scenarios, it is relevant to estimate 3D pose in a global or world coordinate system. Estimating absolute 3D poses is more advantageous than root-relative 3D poses in real-world applications. For instance, in an unmanned store, it is essential to detect the merchandise picked up by a customer, which relies on accurate hand localization in the world coordinate system (Zhan et al. 2022). Similarly, applications such as augmented reality require camera-centric human body keypoints for re-localizing a person in the world coordinate system. The research works (Cheng et al. 2023; Wu and Xiao 2020; Cheng et al. 2021a; Luvizon et al. 2022; Wei et al. 2022; Cheng et al. 2021b; Zou and Tang 2021; Jin et al. 2022; Wang et al. 2022) use a monocular image or video input to find human keypoints in camera coordinates. Most multi-view research works (Zhang et al. 2021; Dong et al. 2022; Gholami et al. 2022) estimate human keypoints with respect to world coordinates, enabling articulation across multiple views.

2.3 Multi-person paradigms

The problem-solving paradigms used in 3D multi-person pose estimation (3D-MPPE) can be divided into top-down, bottom-up, and hybrid approaches as shown in Fig. 2. The top-down paradigm involves first identifying each person using a bounding box and then detecting their keypoints. In contrast, the bottom-up paradigm generates heatmaps for all individuals at once and subsequently assigns keypoints to each person. Generally, top-down methods tend to be more accurate, while bottom-up methods are usually faster. The hybrid approach integrates aspects of both top-down and bottom-up methods to leverage their respective advantages. Additionally, some research tackles the multi-person 3D-HPE problem using a single-stage approach, differing from the typical two-stage solutions of top-down and bottom-up paradigms.

2.3.1 Top-down paradigm

Top-down methods first identify bounding boxes likely to contain a person and then perform single-person HPE for each detected individual. Existing methods (Reddy et al. 2021; Cheng et al. 2021a; Dong et al. 2022; Han et al. 2022; Guo et al. 2021; Lin and Lee 2021; Chen et al. 2020; Wu et al. 2021) utilize this top-down approach to estimate 3D poses in multi-person scenarios. Generally, top-down approaches offer superior accuracy in pose estimation and are ideal for applications requiring high precision (Wu and Xiao 2020). However, these methods heavily depend on the accuracy of the people detector (Han et al.

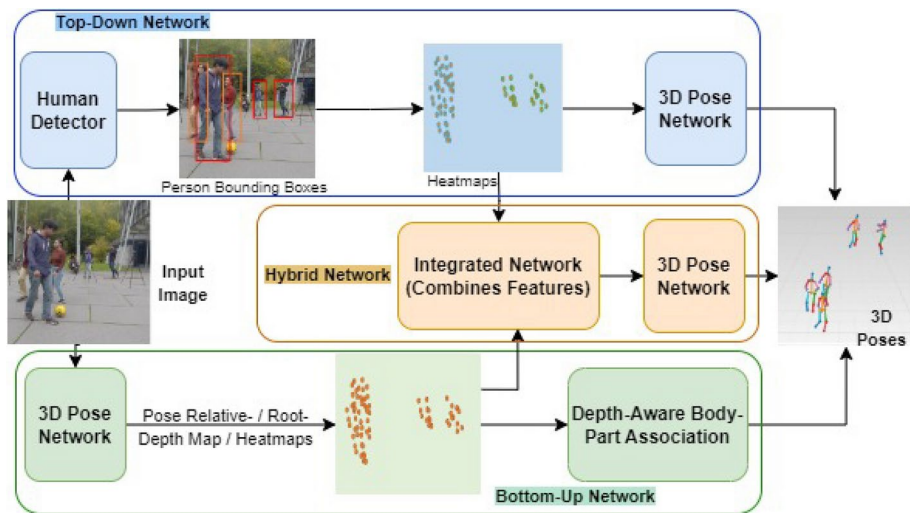


Fig. 2 Multi-person paradigms: top-down, bottom-up, and hybrid approaches

2022) and often fail to detect individuals who are mostly occluded. Additionally, they do not scale well in crowded scenes, as their computational complexity increases with the number of people, resulting in slower performance and not suitable for real-time processing. The research work (Benzine et al. 2020) employs bounding box detection without depending on the number of persons in the images.

2.3.2 Bottom-up paradigm

Bottom-up approaches do not require human detection and can estimate the poses of multiple people simultaneously. Typically, these methods involve generating heatmaps followed by post-processing steps to assemble joint detections into complete human skeletons. Zhou et al. (2021, 2023) present a bottom-up method for instant-aware human body parsing, where semantic and geometry-rich human joints are regressed as pixel embeddings, enabling efficient grouping of fine-grained semantics through joint association. Existing methods (Wang et al. 2021b; Zhang et al. 2023c; Tu et al. 2020; Benzine et al. 2021; Fabri et al. 2020) adhere to the bottom-up paradigm for multi-person 3D-HPE and are more efficient than top-down methods. Unlike top-down approaches, bottom-up methods produce multi-person joint locations in a single pass, allowing for 3D pose inference even under significant occlusion (Benzine et al. 2020). However, most current bottom-up methods process all individuals at a uniform scale, making them sensitive to variations in scale among multiple people and less capable of detecting joints of smaller individuals.

2.3.3 Hybrid/single-stage

Top-down methods are generally effective in considering the scale of people within a scene but are susceptible to detection errors due to occlusion within a person and require more resources as the number of people increases in crowded scenes. On the other hand, bottom-

up methods are usually efficient and do not rely on the number of people in the scene but tend to make errors with small-scale individuals. Cheng et al. (2021b, 2023) propose combining top-down and bottom-up networks to take advantage of both approaches. In their integrated network, the top-down network estimates the joints of all individuals in an image patch, while the bottom-up network uses human detection based on normalized heatmaps, enhancing the model's robustness to scale variations.

The authors of the papers (Jin et al. 2022; Wang et al. 2022) argue that current multi-person methods using either two-stage top-down or bottom-up approaches face issues with redundant computations and high computational costs, resulting in inadequate efficiency for real-time processing. Jin et al. (2022) propose a single-stage solution with a decoupled regression model for multi-person 3D human pose estimation using regression maps. Wang et al. (2022) introduce a single-stage distribution-aware method that enhances body joint estimation recursively.

2.4 Deterministic vs. probabilistic vs. diffusion-based

Existing research on 3D-HPE can be categorized into deterministic and probabilistic approaches based on the number of solutions they provide. Deterministic approaches generate a single, definitive 3D pose for each image, making them practical for real-world applications. Probabilistic approaches model the 2D-to-3D conversion as a probability distribution and produce a set of possible solutions for each image, accommodating uncertainty and ambiguity in the lifting process. To make these probabilistic approaches practical, aggregation techniques combine multiple hypotheses into a single, higher-quality 3D pose. Recently, diffusion models have been popular in image generation tasks and have also been explored for 3D pose estimation. These models are based on the concept of gradually transforming simple data distributions such as Gaussian noise into more complex distributions that represent the final 3D pose. The generic flow of these techniques is illustrated in Fig. 3.

2.4.1 Deterministic approach

Most of the existing 3D-HPE methods (Honari et al. 2023; Artacho and Savakis 2022; Nie et al. 2023; Liu et al. 2020a; Wang et al. 2020; Zhao et al. 2023b; Luvizon et al. 2022; Tang et al. 2023) follow a deterministic approach, providing a definite, single 3D pose for each frame. Deterministic methods are favored because they produce consistent and straightforward results, making them practical for various applications. However, accurately reconstructing the correct 3D pose from 2D joint detections is challenging due to depth ambiguities and occluded body parts, leading to multiple feasible solutions. Approaches based on determinism often overlook these ambiguities by assuming a single solution exists, which can result in less satisfactory results.

2.4.2 Probabilistic approach

Probabilistic methods aim to model the uncertainty in pose estimation mostly from 2D detectors. Bertoni et al. (2019) tackles the ambiguity in the localization by estimating confidence intervals using a loss function derived from the Laplace distribution. Rogez et al. (2020); Wehrbein et al. (2021) propose a probabilistic model to generate multiple 3D pose

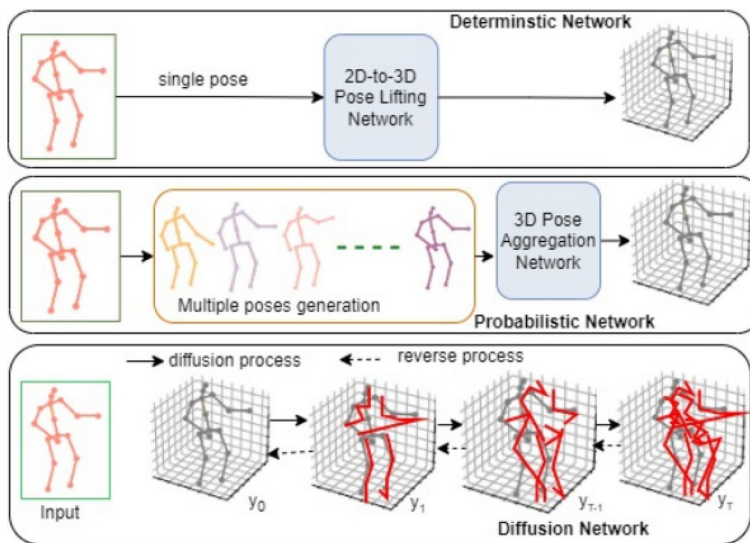


Fig. 3 Generic flow of deterministic, probabilistic, and diffusion-based approaches

hypotheses to create a series of pose suggestions for the localization process in pose detection. Han et al. (2022) proposed uncertainty learning to tackle uncertainties stemming from noisy labeling, imprecise person detection, and occlusion by modeling uncertainties in joint locations using probability. Wandt et al. (2022) employ normalizing flows to establish a prior distribution of poses, facilitating the identification of the most probable up-to-scale 3D pose from random projections. Research in (Li et al. 2023b, 2022) proposes a methodology involving initial one-to-many mapping followed by a many-to-one mapping through multiple intermediate hypotheses. Bartol et al. (2022) propose a method for triangulating human poses that generates multiple 3D pose hypotheses by triangulating random subsets of views for each joint. Lee et al. (2023) utilize a posture entropy-based probabilistic model, Jiang et al. (2023) utilize probability distributions to represent camera poses whereas MHCanonNet (Kim et al. 2024) generates multiple hypotheses for pose and rotation within the spatial domain while capturing temporal correlations of features across multiple frames. A probabilistic 3D-HPE framework can be enhanced by integrating dynamic parameters and heuristic functions, such as joint or motion constraints. A recent study by (Sun et al. 2024) exemplifies this approach by incorporating proprioceptive feedback into their probabilistic model.

2.4.3 Diffusion-based approach

Diffusion models such as (Gong et al. 2023) start from an initial noisy pose estimate and iteratively refine it over time through a learned diffusion process. These methods often involve a learned reverse process where noise is removed step-by-step to recover the true pose. Shan et al. (2023) propose a diffusion-based method that is compatible with existing 3D-HPE methods as a backbone and allows for a customizable number of hypotheses during inference. Holmquist and Wandt (2023) introduce a conditional diffusion model to predict multiple hypotheses as the gradual transformation of a vector representing 3D joint

coordinates into a Gaussian distribution. The denoising phase relies on joint-wise heatmaps generated by a 2D joint detector that uses an embedding transformer. This embedding transformer non-linearly embeds all samples for each joint into a single vector, preserving their multi-modality. Cai et al. (2024a) utilize a disentanglement approach using hierarchical information to explicitly model pose priors in the forward diffusion process. In the reverse process, they use a denoiser consisting of a spatial transformer related to the hierarchy and a temporal transformer.

2.5 Problem solving space

Existing methods are examined based on the problem-solving space where the actual 3D pose estimation task is performed. Many methods for 3D-HPE operate in pixel space, while a few research efforts have proposed models that function in 3D space using discretized voxels. Recently, some studies have leveraged NeRF for various tasks related to 3D human modeling.

2.5.1 Pixel-based approach

Most existing methods (Honari et al. 2023; Cheng et al. 2023; Kim et al. 2024; Chen et al. 2024; Nie et al. 2023; Lee et al. 2023; Zhao et al. 2023b; Chen et al. 2022; Tang et al. 2023; Gong et al. 2023; Yu et al. 2023) perform the task 3D-HPE within the pixel space. These approaches typically begin by extracting image features related to human pose and representing them in a 2D space. The subsequent pose estimation task is then performed on these 2D feature maps. These pixel-based methods are computationally efficient, but they face significant challenges due to the highly non-linear process of regressing 3D coordinates from 2D feature space.

2.5.2 Voxel-based approach

Volumetric supervision in 3D pose estimation offers detailed ground truth data for each voxel in 3D space, providing richer information. This approach directly regresses 3D coordinates by transforming them into predicting confidence estimates for each voxel (Pavlakos et al. 2017). Additionally, directly regressing joint coordinates becomes impractical in multi-person environments when the number of people is unknown beforehand, making volumetric heatmaps a natural choice for bottom-up multi-person 3D-HPE (Fabbri et al. 2020). Tu et al. (2020); Reddy et al. (2021); Zhang et al. (2023c) propose voxel-based approach for estimating the 3D poses of multiple people from multiple camera views, operating directly in 3D voxel space by integrating all camera views. In this technique, the integrated 3D representation effectively eliminates the need for associating 2D poses across different views, which is beneficial for handling occlusions. However, the accuracy is highly dependent on the resolution of the volumetric heatmap, with precision linked to the size of each voxel, and high-resolution heatmaps require significant memory, leading to scalability problems. Consequently, these methods are either accurate but slow or fast but inaccurate, and a new model must be trained for each specific speed-accuracy trade-off (Luvizon et al. 2023).

2.5.3 NeRF-based approach

Recently, Neural Radiance Field (NeRF) (Mildenhall et al. 2020) has gained prominence for generating novel views by synthesizing scenes from multiple images. Researchers propose to leverage the NeRF for 3D human modeling in various tasks such as animation, avatar, and portrait synthesis. Some recent works, such as A-NeRF (Su et al. 2021) and Neural Body (Peng et al. 2021, 2023) have specifically addressed the representation of the articulated human body using NeRF. A-NeRF integrates articulated skeleton models and implicit functions to create 3D representations from 2D images. It initializes articulated body poses using a standard 3D pose estimator and utilizes spatial latent variables. Neural Body focuses on generating new views of moving humans from sparse multi-view videos by using latent codes to encode local features and appearance. These codes are associated with vertices from the SMPL (Loper et al. 2015) model, which captures human figure dynamics. Recent research enhancements (Peng et al. 2023) integrate an implicit surface model to improve geometry extraction and constrain learned geometry during training. Gholami et al. (2024) propose a NeRF-based data augmentation method incorporating A-NeRF (Su et al. 2021) to generate diverse 3D human pose data. Numerous methods employ NeRF for human modeling, typically relying on off-the-shelf 3D pose estimators or heavy parametric models such as SMPL, which implicitly include 3D pose information. At present, there are no widely recognized methods that specifically target 3D human pose estimation using NeRF.

2.6 Emerging techniques and needs

In addition to the strategies outlined in the previous subsections, several new techniques have recently been proposed. With the success of large language models (LLMs) in natural language processing, some recent studies (Wang et al. 2024; Feng et al. 2024) have begun exploring their application to 3D-HPE. LLMs hold significant promise for improving 3D-HPE by integrating semantic understanding from language models with geometric reasoning, potentially leading to more robust and accurate pose estimation systems.

Lightweight models for real-time 3D-HPE are becoming increasingly important with the growing use of mobile devices, tablets, and embedded systems. Some methods (Choi et al. 2021b; Ye et al. 2022; Liu et al. 2020b) have been proposed to achieve real-time 3D-HPE, however, the key challenge is designing models that are both efficient and accurate, able to run on devices with limited computational resources, such as lower memory and processing power. These advancements would enable real-time applications such as AR/VR experiences, fitness tracking, and human-computer interaction on portable, resource-constrained devices, making 3D-HPE more accessible and practical for everyday use.

Many existing methods for estimating 3D poses focus on scenarios with uniform backgrounds. However, exploring the interaction between humans and dynamic 3D scenes for real-world applications is crucial. This involves understanding how humans interact with their environment, which is essential for developing more robust and applicable pose estimation techniques. Some studies (Guzov et al. 2021; Yi et al. 2022; Wang et al. 2023a) explore human-scene interaction using various sensors, but further research is needed to advance these methods in dynamic environments. Recent work by (Zhou et al. 2024c) employs multiple sensors for human activity recognition, incorporating sensor data in the feature selection process. Such approach could enhance human interaction in dynamic environments.

3 Single-person 3D-HPE methods

This section provides an overview of current techniques in single-person 3D-HPE, focusing on predicting the 3D coordinates of key body joints, including the head, shoulders, elbows, and knees. These methods are categorized based on their use of monocular or multi-view images and videos, and further classified by the deep learning algorithms they employ, such as CNNs, GCNs, Transformers, and hybrid models. The general procedure for single-person 3D-HPE is summarized in Fig. 4.

3.1 3D-HPE from monocular image

3D-HPE from monocular images involves the use of RGB images from a single camera which doesn't require calibration of camera parameters. It has gained significant attention as an essential component in numerous human-centric applications. Despite this interest, the task remains highly challenging due to depth ambiguity, and occlusion, necessitating large volumes of in-the-wild datasets for effective generalization. Depth ambiguity can be defined as the difficulty of accurately determining the depth (distance from the camera) of body joint keypoints. This occurs because different 3D poses can produce similar or identical 2D projections, making it hard to infer the correct depth from a single camera view. Similarly, occlusion refers to the situation where parts of the body such as arms or legs are blocked from view by other parts of the body or external objects, making it difficult to estimate the full pose. Methods must be robust to occlusions and infer the hidden body joints based on the visible joints. Various methods have been developed to address these challenges, which are discussed in the following sub-sections.

3.1.1 CNN-based methods

Several research efforts have introduced end-to-end approaches to tackle the need for extensive 3D pose annotations, aiming to improve generalization and effectively address challenges such as depth ambiguity and occlusion. Umar et al. (2020) utilize 2D pose-annotated images to train their model, incorporating a heatmap loss and a multi-view consistency loss. Mitra et al. (2020) use synchronized videos from multiple views to learn pose embed-

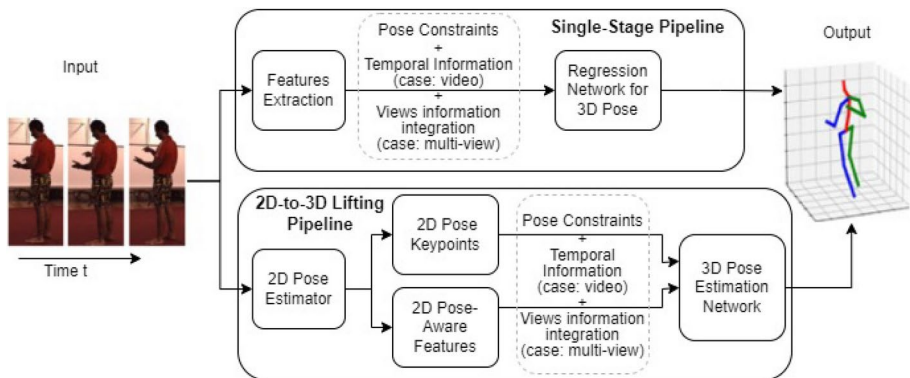


Fig. 4 Single-person 3D-HPE pipelines in single stage and 2D-to-3D lifting

dings, using a limited set of 3D pose labels to ensure that similar poses are mapped closely in the embedding space. The research work (Kundu et al. 2020b) utilizes unlabeled monocular video samples and designs a kinematic structure preservation pipeline incorporating hierarchical limb connectivity. The research work (Luvizon et al. 2023) proposes Scalable Sequential Pyramid Networks to address scalability issues by regressing poses at multiple scales and generating heatmaps and depth maps used for 3D pose regression.

Some research works utilize various body parts geometry to handle depth ambiguity and occlusion. Fang et al. (2018); Xu et al. (2022) present methods to leverage pose grammar such as kinematics, symmetry, and motor coordination. These architectures feature a hierarchy of Bi-directional RNNs, which efficiently capture the bidirectional relationships within the dependency grammars. Liu et al. (2020a) leverage the relationships between various body parts using a graphical convolutional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997). The research work (Kundu et al. 2020a) introduces a differentiable part-based representation that includes a single part-based 2D puppet model, constraints on human pose articulation, and a collection of unpaired 3D poses. Wu and Xiao (2020) uses depth values of human joints and limb depth maps, created by densely interpolating depth values along the limbs, assuming the limbs are rigid structures.

Researchers have also introduced various learning and optimization strategies to address the problem of data scarcity, generalization, and occlusion. Li et al. (2020b) present a self-supervised learning approach leveraging geometric principles to create a supervision signal from multi-view frames. To handle the impact of noise from self-occlusions, the network incorporates the confidence levels of the 2D joints. Wandt et al. (2021) design CanonPose to learn monocular 3D-HPE using multi-view self-supervision from uncalibrated cameras. The key concept is that 2D detections from one viewpoint can be projected onto another viewpoint through a canonical pose space. Zhang et al. (2020) introduce an inference stage optimization method to enhance model generalization by conducting self-supervised learning on unlabeled target images during the inference phase. Kundu et al. (2022) combine model-free joint localization and model-based parametric regression networks. It incorporates negative samples, including backgrounds and simulated synthetic joint-level occlusions, to enhance the effectiveness of the proposed pose and joint uncertainties. CNN-based methods are straightforward to integrate with other networks. For example, (Gao et al. 2023) utilize a CNN-based body pose estimation sub-network to predict the coordinates of hand joints, significantly improving the overall performance of their network.

3.1.2 GCN-based methods

GCNs have been used in two-stage methods as the detected 2D keypoints can be represented as a graph structure, allowing the relationships between body joints to be effectively captured and processed. The body joints relationship and limb length constraints are utilized to handle depth ambiguity and occlusion. Ma et al. (2021) introduce ContextPose which integrates a pictorial structure model imposing limb length constraints within a graph neural network. Xu and Takano (2021) propose Graph Stacked Hourglass Network by integrating multi-scale and multi-level features with prior knowledge of human skeletons. The research work (Xu et al. 2021) proposes a method using a teacher-student approach. The teacher network employs pose-dictionary-based regularization and the student network uses a GCN to estimate 3D coordinates exploiting geometry consistency. Zou and Tang (2021) introduce a

weight-modulated GCN to separate the feature transformations of different nodes and learn a variety of relational patterns between body joints. Ci et al. (2022) capture dependencies among graph nodes by decomposing the Laplacian operator into a product of a predefined structure matrix and a learnable weight matrix. Azizi et al. (2022) present a spectral GCN architecture utilizing the Möbius transformation on the eigenvalue matrix of the graph Laplacian to explicitly encode the pose through inter-segmental angles and joint translations. Zhang et al. (2023c) propose the Parallel Hop-Aware Graph Attention Network to learn correlations of skeleton joints while maintaining spatially precise representations of the human graph and the semantic correlation of multi-hop nodes.

3.1.3 Transformer-based methods

Some studies suggest using pose-specific self-attention mechanisms to exploit the structure of the human skeleton. Li et al. (2023a) propose a pose-oriented adaptation of the transformer architecture which computes the relative distance between pairs of joints and encodes this information as an attention bias within the self-attention process. This mechanism enhances the ability to capture dependencies within the human skeleton. Zhou et al. (2024a) propose to leverage image information by selecting valuable features from various image patches in their network.

3.1.4 Hybrid/other methods

GCN+transformer: Some researchers have combined GCNs and Transformers in their networks to leverage the strengths of both architectures. Zhao et al. (2022) propose GraFormer, comprising GraAttention and ChebGConv components. The GraAttention integrates graph convolution layers in place of MLP following the multi-head attention to capture global information from all nodes. The ChebGConv block facilitates information exchange between non-adjacent nodes, thereby achieving a broader receptive field. Chen et al. (2024) introduce a dynamic graph transformer, consisting of a Transformer Encoder (TE), an immobile GCN, and a dynamic GCN. The TE module utilizes a self-attention mechanism to understand complex global relationships among skeleton joints. The immobile GCN focuses on identifying the local physical connections between human joints. Meanwhile, the dynamic GCN is dedicated to learning sparse dynamic K-nearest neighbor interactions based on various action poses, effectively addressing the pose-dependent correlations of joints.

Other methods: Li and Pun (2023) combine GAN-based pose generation and MLP-based estimation networks through co-training. The GAN-based network increases diversity in the generated poses and a joint-angle loss to ensure poses remain within reasonable limits. An MLP-based body-parts grouping strategy is proposed for pose estimation. Chai et al. (2023) propose an unsupervised method based on GAN to tackle the problem of 3D human pose domain adaptation. The global position alignment is achieved through a geometric constraint and the local root-relative poses are augmented to increase the diversity of local structures in the 3D poses. Nie et al. (2023) introduce an AutoEncoder (AE) based method to learn a 3D human body model by addressing the 2D-3D ambiguity through domain adaptation in regression. Kang et al. (2023) propose an enhancement to convolution operations called GridConv by transforming the pose representation from a graph structure

to a weave-like grid pose using Semantic Grid Transformation. An attention module is integrated into the convolutional kernel, which enhances the capability of GridConv to encode contextual information.

3.2 3D-HPE from monocular video

Estimating 3D Human Pose from videos captured by a single RGB camera is crucial for various applications, such as surveillance systems, due to the low cost and ease of deployment of RGB cameras. Monocular video-based methods can leverage temporal information to resolve depth ambiguity and occlusion challenges. However, these methods encounter several difficulties, including the high-dimensional variability and nonlinearity of human dynamics, along with issues like motion blur and variations in movement speed relative to the training data. Various methods have been proposed to address these challenges, which are examined in the following sub-sections.

3.2.1 CNN-based method

Most of the methods utilize temporal information to address depth ambiguity and other challenges. Artacho and Savakis (2022) present an end-to-end unified framework, Uni-Pose+, leveraging a linear sequential LSTM for temporal information. Pavllo et al. (2019) employ dilated temporal convolutions to capture long-term dependencies, allowing for the simultaneous processing of multiple frames for solving depth ambiguity. Wang et al. (2020) propose a method to learn feature representations for both images and spatial context using CNNs and LSTM to capture temporal dependencies for smooth prediction. Liu et al. (2020b) integrate an attention mechanism to utilize temporal context through a temporal convolutional network (TCN). Xu et al. (2020) integrate kinematic analysis into deep models to leverage human prior knowledge by adhering to the constraints of perspective projection to refine noisy 2D inputs and 3D pose refinement is achieved through trajectory completion. Cheng et al. (2020) propose multi-scale features both spatially and temporally to deal with people at varying distances and speeds. Liu et al. (2021) leverage an attention-based model and a multi-scale structure of dilated convolutions to learn implicit dependencies within large temporal receptive fields without increasing network layers. Cho et al. (2021) proposed a method to adapt to camera distortion scenarios in monocular video by utilizing model-agnostic meta-learning. Lee et al. (2023) propose Temporal Propagating LSTM Networks using spatio-temporal posture correlations to minimize posture entropy, addressing joint interdependency, temporal consistency, and human perception.

3.2.2 GCN-based methods

GCN-based methods utilize spatial and temporal representations from multiple neighboring nodes to get richer information about the keypoints. Zeng et al. (2021) propose a dynamic hop-aware hierarchical channel-squeezing fusion layer, which updates node features by extracting relevant information from neighboring nodes and minimizing unwanted noise. The research work (Yu et al. 2023) proposes a global–local learning architecture that leverages both global spatio-temporal representation and local joint representation within a GCN-based model utilizing ground truth 2D pose data.

3.2.3 Transformer-based methods

The self-attention mechanism in Transformers enables the distinct capture of global correlations across extended input sequences. Consequently, several transformer-based approaches have been developed to process long video sequences, thereby acquiring more comprehensive pose information to tackle various challenges. Most of the methods (Zheng et al. 2021; Li et al. 2022; Zhao et al. 2023b; Kim et al. 2024) utilize spatial transformers to capture local relationships between 2D joints within each frame through spatial self-attention layers and considering the positional information of the 2D joints and temporal transformer module to examine the global dependencies among the spatial feature representations across different frames.

Zhang et al. (2022) introduce MixSTE (Mixed Spatio-Temporal Encoder) to model the temporal motion of each joint individually and a spatial transformer to capture inter-joint spatial correlations. Unlike approaches that focus on the central frame, MixSTE processes entire frames from the input video, enhancing the coherence between the input and output sequences. Tang et al. (2023) introduce a spatio-temporal criss-cross attention (STC) mechanism that models spatial and temporal information in parallel. Zhou et al. (2024b) propose a SlowFastFormer that includes a parallel encoding stage with slow and fast branches followed by a blending stage to handle lower/higher rate motion. Cai et al. (2024a) propose a diffusion-based method that employs a disentanglement strategy in the forward diffusion process, leveraging hierarchical information to explicitly model pose priors. The model includes a denoiser composed of a hierarchical-related spatial transformer and a temporal transformer, which enhances the connections among hierarchical joints by increasing the attention weight of adjacent joints during the reverse diffusion process.

3.2.4 Hybrid methods

GCN+transformers: Zhu et al. (2021) present PoseGTAC with a Graph Atrous Convolution (GAC) to expand receptive fields and learn multi-scale pose context, alongside a Graph Transformer Layer to capture global long-range relationships. Li et al. (2023b) propose an extension of previous work (Li et al. 2022) by replacing standard transformer encoders with graph transformer encoders to capture both global and local information of body joints to better constrain the physical structure of the estimated 3D poses. Gong et al. (2023) introduce a diffusion-based architecture integrating a GCN-based diffusion model and a context encoder aiming to address indeterminacy and uncertainty. Zhai et al. (2023) propose a Hop-wise GraphFormer with Intragroup Joint Refinement to investigate the potential correlations among human joints. The Hop-wise GraphFormer component clusters joints based on their k-hop neighbors and employs a transformer-like attention mechanism to these clusters, aiming to uncover hidden joint synergies. Chen et al. (2023) introduce the High-order Directed Transformer (HDFormer) to address the challenges of overlapping keypoints, rapid pose changes, and varying scale changes in the depth estimation of 3D keypoints. HDFormer leverages joint-joint, bone-joint, and hyperbone-joint relationships, incorporating self-attention and high-order attention to create a multi-order attention module. This approach integrates GCN within the transformer-based framework to achieve a balance between performance and efficiency.

3.3 3D-HPE from multi-view images

The primary objective of estimating single-person 3D pose from multiple camera images is to determine the absolute 3D locations of body joints within a world coordinate system. Utilizing multi-view information can effectively restore depth information and occluded parts. However, applying the multi-view approach in practice presents several challenges, as it requires synchronized and calibrated cameras to aggregate information from multiple views. Various methods employing different deep learning algorithms have been proposed to address these challenges, as discussed in the following sub-sections.

3.3.1 CNN-based methods

Remelli et al. (2020) proposed a lightweight architecture to obtain feature maps in a canonical frame of reference which are disentangled from the camera poses. It predicts 2D joint locations from a learned camera-independent representation of 3D pose and lifts 2D pose to 3D via an efficient formulation of differentiable triangulation. AdaFuse (Zhang et al. 2021) fuses the heatmaps from each view using epipolar geometry and camera parameters to find corresponding points and a SoftMax operator is applied to reduce noise. The research work (Luvizon et al. 2022) addresses the task of predicting 3D human poses in absolute coordinates separating the prediction of absolute depth and the estimation of relative joint depths by utilizing view frustum space pose estimation. Wan et al. (2023) propose a Multi-View Fusion module that utilizes keypoints detected in source views to generate pseudo heatmaps. It represents the probability distribution of keypoints localized in the reference view through feature matching and Holistic Triangulation incorporates anatomy priors to maintain pose coherence and plausibility inferring the entire pose as a unified entity. Jiang et al. (2023) introduce a probabilistic triangulation module for estimating uncalibrated 3D human poses from multiple views. Usman et al. (2022) propose a Metapose without 3D joint annotations or camera parameters during training or inference. Metapose combines pose predictions and uncertainty estimates from multiple views by fusing the outputs of single-view modules using a simple view-equivariant neural network.

3.4 3D-HPE from multi-view videos

Estimating the 3D pose of a single person using videos from multiple viewpoints can efficiently address issues related to depth ambiguity and occlusion. Nevertheless, this method requires substantial computational power and an expensive setup. As a result, there are relatively few research studies in this area and not suitable for real-time processing.

Shuai et al. (2023) propose a Multi-view Fusion Transformer for handling variable multi-view sequences without the need for camera calibration. It incorporates the inter-view relationships into its relative-attention block and reconstructs features from multiple views. Additionally, the method employs a random mask mechanism within the multi-view fusion and the temporal fusing transformer to enhance robustness against variations in the number of views and the length of the video. Cai et al. (2024b) propose FusionFormer which also does not rely on camera parameters and fuses multi-view and multi-frame features within a unified scheme. FusionFormer encodes 2D pose estimation results into pose features and

then uses a transformer encoder to jointly merge multi-view and multi-frame features into a global feature.

Zhang et al. (2024) propose a multi-view 3D human pose estimation method based on a deep semantic graph transformer. This network dynamically learns deep semantic features and their relationships, incorporating the position, spatial structure, and skeletal edges of all human joints. By progressively fusing crucial spatial-temporal information from multiple viewpoints, it effectively models the long-term dependencies between frames.

4 3D multi-person pose estimation methods

3D multi-person pose estimation (3D-MPPE) is highly suitable for addressing various real-world problems, including sports performance analysis, human interaction, and video surveillance. Despite its potential, research in 3D-HPE for multiple individuals remains limited and challenging. This is primarily due to the lack of sufficient in-the-wild datasets, issues with intra-person occlusion, and varying input image/video quality. Additionally, it is necessary to estimate the 3D pose in an absolute coordinate system to accurately reflect each person's position relative to others, mirroring their actual locations in real-world scenes.

4.1 3D-MPPE from monocular image

Estimating the 3D pose of multiple people from monocular images is particularly challenging because it requires calculating the poses of multiple individuals from the features of a single image. Various methods have been proposed to address this problem, utilizing different deep learning algorithms.

4.1.1 CNN-based methods

Rogez et al. (2020); Benzine et al. (2020) propose anchor-based top-down methods to estimate multi-person 3D poses. The PandaNet Benzine et al. (2020) enables the use of lower resolution outputs compared to heatmaps, as a single output pixel can represent the entire pose of a subject beneficial for managing occlusion. Guo et al. (2021) proposed a method that utilizes a decoupling dual-branch structure to reconstruct human height and determine the depth and position of the 3D human root. An attention mechanism improves relative pose estimation accuracy, and synthetic occlusions are added to training samples to handle bounding box shrinkage due to occlusions.

Jin et al. (2022) introduce a single-stage pipeline for multi-person 3D pose estimation by presenting the Decoupled Regression Model. This approach combines 2D pose regression with person depth estimation to perform 3D pose regression jointly through a position query operation and adaptively understand the scale information of instances. Wang et al. (2022) propose a Distribution-Aware Single-stage model designed to localize the 3D positions of individuals and their body joints within the camera coordinate space simultaneously.

4.1.2 GCN-based methods

Han et al. (2022) propose a top-down method to predict 3D human joint locations and their associated uncertainties simultaneously. Visual joints typically have low uncertainties and occluded joints have high uncertainties. Based on the learned uncertainties and the relationships among joints, the uncertainty-aware GCN refines the initial 3D joint estimates by ensuring that those with low uncertainties regulate joints with high uncertainties. Chen et al. (2022) developed a part-aware top-down model to address the heterogeneous nature of human body parts, unlike existing methods that use a uniform decoder architecture for all parts.

4.2 3D-MPPE from monocular video

Existing methods for 3D multi-person pose estimation from monocular videos are examined in this subsection. These methods aim to address challenges such as intra/inter-person occlusion and depth ambiguity by effectively utilizing temporal information.

4.2.1 GCN-based methods

Cheng et al. (2021a) propose a top-down method which uses directed graph-based joint-GCN and human-bone GCN, leveraging heatmap confidence scores and part affinity field scores to handle occlusion, partial out-of-frame scenarios, and inaccurate human detection. They also introduce joint-TCN for person-centric 3D pose across frames, velocity-TCN for pose consistency in consecutive frames, and root-TCN for estimating 3D pose in camera coordinates using the weak perspective projection without the need for camera parameters. The research work (Cheng et al. 2021b) introduces an integration network that combines a top-down branch, which predicts 3D poses for all individuals within the bounding box, and a bottom-up branch, which addresses scale variations using joint heatmaps from the top-down branch. Additionally, this research proposes a two-person pose discriminator to tackle issues related to inter-person occlusion and close interactions in monocular video for multi-person pose estimation. The authors extend the research work (Cheng et al. 2021b) in their paper (Cheng et al. 2023) by introducing test-time optimization to address the discrepancies between training and testing data. This approach integrates high-order temporal constraints, reprojection loss, and bone-length regularization.

4.2.2 Transformer-based method

Park et al. (2023) propose a top-down method with a sequence-to-sequence 2D-to-3D lifting model powered by a geometry-aware data augmentation for solving the problem of robustness, occlusion, and jitter in the output. The Single Person Spatial Transformer captures spatial correlations of joints belonging to each person within a frame. Then, the Inter-Person Spatial Transformer learns the relationships among multiple individuals in the same frame. Finally, the Single Joint Temporal Transformer provides temporal context to each token in the sequence.

4.3 3D-MPPE from multi-view images

Estimating 3D poses of multiple individuals from multi-view images requires identifying correspondences across views. Current approaches for multi-person scenarios in multi-view settings can be categorized into reconstruction-based methods and volumetric representation-based methods, as depicted in Fig. 5. Reconstruction-based approaches (Dong et al. 2022) and volumetric methods (Zhang et al. 2022) use multi-step processes, while (Wang et al. 2021b) proposes a solution that eliminates the need for intermediate stages. The subsequent sections explore recent techniques that leverage these strategies to enhance cross-view correspondence for improved performance in multi-person pose estimation.

4.3.1 CNN-based methods

Lin and Lee (2021) propose a top-down method utilizing the plane sweep algorithm to implicitly enforce cross-view consistency constraints through multiple reference camera views, enhancing the accuracy of depth regression. Their method conducts depth regression for each joint in each 2D pose within a target camera view using a coarse-to-fine approach. This begins with regressing the person-level depth and then estimates the relative depth at the joint level for each individual. The 3D poses are derived by back-projecting these estimated depths.

Tu et al. (2020) propose VoxelPose, a bottom-up method for estimating multi-person poses from multiple camera views by operating directly in 3D space. Rather than relying on cross-view correspondence, VoxelPose combines features from all camera views within the 3D voxel space. The Cuboid Proposal Network, which utilizes 3D convolution, is used to localize all individuals in the scene. Then, the Pose Regression Network is designed to estimate a detailed 3D pose for each detected individual.

4.3.2 GCN-based method

The research work (Wu et al. 2021) proposes a top-down method that employs a coarse-to-fine strategy comprising three modules. Initially, the Multi-view Matching Graph Module identifies coarse human center candidates by matching people across different views using visual and geometric cues. Then, the Center Refinement Graph Module refines these human

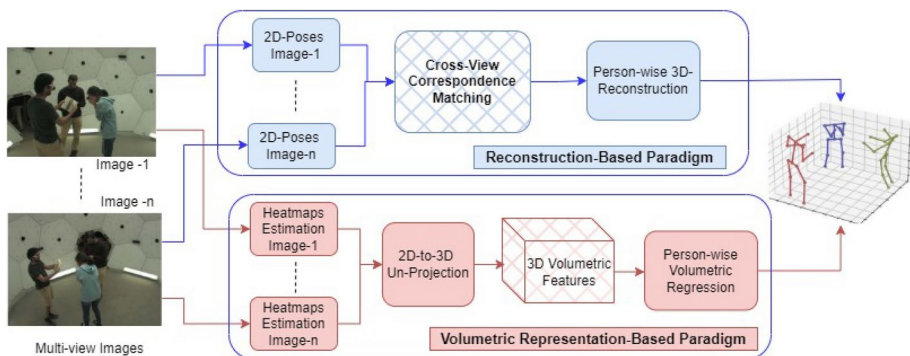


Fig. 5 General paradigms for multi-view 3D-MPPE

center locations by aggregating multi-view features and adaptively sampling points to enhance efficiency and accuracy in localization. Finally, the Pose Regression Graph refines the 3D poses by leveraging the spatial relationships between body joints and the geometric relationships across multiple views.

4.3.3 Transformer-based method

Wang et al. (2021b) propose a Multi-view Pose Transformer (MvP) model that uses a direct regression method to estimate the 3D poses of multiple people. MvP features a hierarchical approach to represent query embeddings for multiple skeleton joints and introduces an adaptive query method based on the input. Additionally, MvP incorporates a geometrically guided attention mechanism known as projective attention to merge cross-view information for each joint. Furthermore, MvP introduces a RayConv operation to embed view-dependent camera geometry into feature representations, enhancing projective attention.

4.4 3D-MPPE from multi-view videos

Multi-view multi-person 3D pose estimation focuses on accurately locating the 3D skeleton joints for each individual in a scene using inputs from multiple camera views. The primary challenges include ambiguity in cross-view associations and the vast state space. The existing approaches to address these issues are discussed in the following subsections.

4.4.1 CNN-based methods

Chen et al. (2020) propose a method to process video inputs from multiple cameras sequentially instead of simultaneously using the temporal consistency in videos to align 2D poses from each camera view directly with 3D poses in 3D space. Reddy et al. (2021) present TesseTrack, which employs a spatio-temporal formulation that does not depend on a fixed number of camera views. A 3D CNN is utilized to locate each person within the voxel space and 4D CNN processes a defined spatio-temporal volume around each detected individual to produce short-term, person-specific representations. Dong et al. (2022) propose a multi-way matching algorithm to identify cycle-consistent correspondences of detected 2D poses across different views by integrating geometric and appearance cues. Then, the 3D pictorial structure model is used to reconstruct the 3D pose for each person from the corresponding bounding boxes and keypoints. Zhang et al. (2023c) propose a multi-branch network to directly estimate and track 3D poses from a 3D voxel-based representation constructed from multi-view images. Specifically, noisy and incomplete information from all camera views is warped to a common 3D space, forming a comprehensive feature volume. The method estimates 3D poses from the voxel representation by predicting whether each voxel contains a particular body joint.

5 Learning techniques

Deep learning methods demand substantial amounts of data to ensure the robustness and generalizability of models. Current approaches to 3D-HPE predominantly follow a fully supervised paradigm, requiring ground truth 3D data as input. However, obtaining 3D pose data is both costly and time-consuming, often necessitating a multi-view setup or a motion capture system, making it impractical for uncontrolled scenarios. To address the scarcity of in-the-wild 3D pose data, several learning techniques such as weakly-/semi-/ self-supervised, and unsupervised learning have been employed which are summarized in Fig. 6.

5.1 Supervised learning

Supervised learning uses labeled datasets to train the model. Most of the 3D-HPE methods described in section 3 and 4 such as (Zhang et al. 2023b; Luvizon et al. 2023; Zhang et al. 2023c; Chen et al. 2024; Zhan et al. 2022; Chen et al. 2022; Luvizon et al. 2022; Jin et al. 2022; Dong et al. 2022; Tang et al. 2023; Gong et al. 2023; Holmquist and Wandt 2023; Yu et al. 2023; Azizi et al. 2022; Zhao et al. 2022; Li and Pun 2023; Kang et al. 2023; Zhang et al. 2022; Cai et al. 2024a; Zhai et al. 2023; Chen et al. 2023; Park et al. 2023) leverage supervised learning models. These methods perform highly for the given dataset, however,

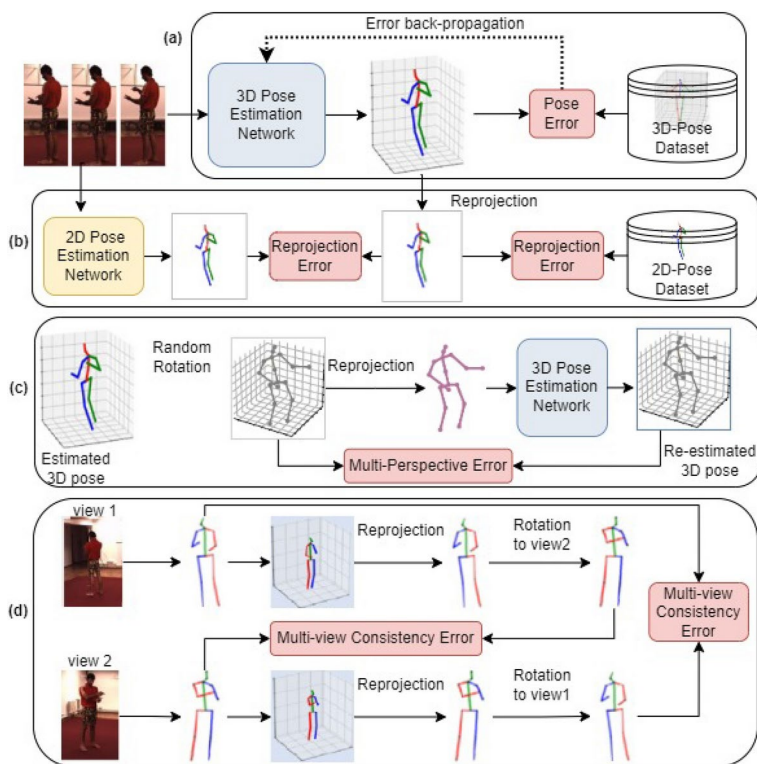


Fig. 6 Summary of learning techniques: **a** supervised learning, **b–d** self-supervised learning with various strategies

lack generalizability in uncontrolled environments due to the limited availability of labeled in-the-wild 3D pose datasets. Hence, various learning strategies beyond supervised learning have been proposed to handle the data scarcity and diversity problem.

5.2 Semi-supervised learning

Semi-supervised learning aims to improve learning performance by utilizing unlabeled data along with the labeled data. It addresses the challenge of scarce labeled data by exploiting unlabeled data. Pavllo et al. (2019) utilize unlabeled video data for semi-supervised training, drawing inspiration from cycle consistency. Their approach involves predicting 2D keypoints for an unlabeled video using a standard 2D keypoints detector, estimating 3D poses from these keypoints, and mapping the estimated 3D poses back to 2D space. Cheng et al. (2020) also incorporate 2D video data to introduce a semi-supervised learning approach. Mitra et al. (2020) introduce a multi-view consistent metric-learning based semi-supervised learning approach that eliminates the need for 2D pose labels to train a network for 3D human pose estimation from single images. Additionally, it incorporates limited 3D supervision to promote the learning of pose-specific features within the network.

The research works (Cheng et al. 2021b, 2023) aim to enhance network performance through semi-supervised learning incorporating reprojection error and multi-perspective. The reprojection error measures the difference between the projection of generated 3D poses and the detected 2D poses. In contrast, the multi-perspective error assesses the consistency of the predicted 3D poses from various viewing angles. Nie et al. (2023) present a learning architecture based on domain adaptation, which integrates supervised and semi-supervised modes in a unified manner without the need for additional information such as multi-views or temporal consistency. This approach leverages abundant 2D poses to enhance the accuracy of 3D pose estimation.

5.3 Weakly-supervised learning

Weakly supervised learning involves working with imperfectly labeled data, where labels may be noisy, incomplete, or inexact. It tackles the issue of poor quality labels by developing models robust to label imperfections. Umar et al. (2020) utilizes unlabeled multi-view images and an independent collection of images annotated with 2D poses. The 3D supervision is derived through a multi-view consistency loss, which enforces the 3D pose obtained from multi-view images to remain consistent under rigid transformations. Chen et al. (2021) propose Deductive Weakly-Supervised Learning (DWSL) for estimating 3D poses from monocular images using uncalibrated cameras and multi-view 2D human poses. DWSL learns latent representations of depth and camera pose for 3D pose reconstruction, and incorporates deductive reasoning to infer human poses from different views. The method uses reconstruction loss to ensure the reliability of what the model learns and infers. Qiu et al. (2023) introduce Weakly-Supervised Pre-Training (WSP) method to extract weak 3D information from monocular images without 3D annotations. WSP is initially pre-trained on 2D pose datasets and then fine-tuned on 3D pose datasets.

5.4 Unsupervised learning

Unsupervised learning involves training a model on a dataset without any labeled data aiming to discover hidden patterns within the data. The research work (Kundu et al. 2020b) proposes to estimate 3D pose from monocular images from unlabeled in-the-wild monocular video utilizing kinematic constraints. Xu and Takano (2021) propose an unsupervised teacher-student network for 3D-HPE from 2D poses. The teacher network ensures rotation invariance by aligning 2D poses from different views to the same 3D pose leveraging a cycle-consistent architecture. The student network also follows the cycle-consistent architecture to ensure pose estimation consistency in the input view with rotation-equivariance, enhancing its training through geometric self-supervision. Yu et al. (2021) proposed a method to address scale ambiguity and pose ambiguity in unsupervised learning for 3D pose estimation from monocular video. A random projection of the generated 3D pose undergoes the lifting network, inverse transformation, and re-projection processes. This enables the network to self-supervise the training process by leveraging temporal geometric consistency.

Wandt et al. (2022) proposed ElePose for estimating 3D human poses from monocular images using only 2D annotations by predicting the camera elevation in parallel with a lifting network. The 3D pose is randomly rotated and projected back to 2D to compute a 2D loss. Honari et al. (2023) explore the concept of contrastive self-supervised (CSS) learning, which involves maximizing the similarity between a given sample and a positive counterpart while minimizing the similarity with a negative one. The authors separate each latent vector into a time-variant component and a time-invariant component, applying the contrastive loss solely to the time-variant component. Chai et al. (2023) propose an unsupervised domain adaptation framework that requires non-sequence 2D poses and camera intrinsic parameters from the target dataset, along with a large-scale 3D annotated human pose dataset serving as the source dataset.

5.5 Self-supervised learning

Self-supervised learning is a type of unsupervised learning in which the model generates pseudo-labels from the data and uses a self-generated supervision mechanism to learn useful representations. Wang et al. (2020) introduced a self-supervised correction mechanism involving two complementary learning tasks: converting 2D poses to 3D poses and projecting 3D poses back to 2D ensuring geometric consistency. The work (Kundu et al. 2020a) proposes to use a pair of video frames of the same person with different poses, a part-based 2D puppet model, human pose articulation constraints, and unpaired 3D pose for learning the model. Li et al. (2020b) propose a self-supervised learning method for estimating 3D poses from monocular images which leverages geometric knowledge to construct supervision signals by using multi-view images during the training process. Zhang et al. (2020) introduce a method that incorporates self-supervised learning during the model's inference stage, distinguishing it from other methods that apply self-supervised learning during training.

CanonPose (Wandt et al. 2021) works with uncalibrated, but temporally synchronized, multi-view cameras. The method uses an off-the-shelf 2D pose estimator and outputs a 3D pose in a canonical rotation, allowing projection to any camera setup. Gong et al. (2022) propose an approach that explicitly generates 2D-3D pose pairs, enabling full supervision

and significantly enhancing the self-learning process. This approach incorporates a reinforcement-learning-based imitator, that is trained alongside a pose estimator and a pose generator. Kundu et al. (2022) propose a domain adaptation model with labeled synthetic samples and uncertainty estimation for monocular 3D human pose estimation. Kim et al. (2024) propose a self-supervised training approach for estimating 3D poses from monocular video that handles multiple hypotheses from in-the-wild videos with uncalibrated cameras. The network is trained using multi-view data, leveraging multi-view consistency for supervision.

6 Datasets, evaluation metrics, and augmentation

This section presents the commonly used datasets in recent 3D-HPE methods in subsection 6.1, while subsection 6.2 covers the evaluation metrics applied across various datasets. To mitigate issues related to data scarcity and enhance diversity, subsection 6.3 introduces data augmentation techniques alongside the existing datasets. Table 2 provides an overview of commonly used datasets and their respective evaluation metrics in recent technologies.

6.1 Datasets

Human3.6M (Ionescu et al. 2014) is the largest and most widely utilized dataset for 3D human pose estimation, containing 3.6 million video frames captured by four synchronized cameras operating at 50 Hz from four distinct viewpoints. The dataset was generated by recording professional actors (six males and five females) performing 15 different daily activities, including eating, posing, walking, talking on the phone, and sitting, in an indoor environment. Of the 11 subjects, seven have been annotated with 3D poses. The training set consists of data from five subjects (S1, S5, S6, S7, and S8), while the test set includes two subjects (S9 and S11).

The Human3.6M dataset is commonly used to evaluate 3D-HPE models, with several protocols established to ensure standardized comparisons across methods. The most frequently used are Protocol 1, Protocol 2, and, in some cases, Protocol 3. Protocol 1 calculates the mean per joint position error (MPJPE) in millimeters by measuring the average Euclidean distance between predicted and ground truth 3D joints, aligning the root joints of both poses. Protocol 2 reports the error after applying Procrustes alignment, which adjusts for translation, rotation, and scale, and is commonly referred to as PA-MPJPE. Protocol 3 normalizes the poses by aligning the scale between the predicted and ground truth poses before computing MPJPE, often termed N-MPJPE.

MPI-INF3DHP (Mehta et al. 2017) dataset features real human subjects with ground truth 3D annotations obtained from a markerless motion capture system in both indoor and outdoor environments. It includes data of individuals in everyday clothing, interacting with objects, and performing a wide range of motions from various camera angles. The dataset comprises 8 actors engaging in 8 different activities, such as walking, sitting, and performing complex exercise poses. The 3D-HPE models on this dataset are primarily evaluated using the Percentage of Correct Keypoints (PCK), which assesses the percentage of predicted joints that fall within a specified distance from the ground truth joints. Additionally,

Table 2 Summary of benchmark datasets with evaluation metrics

Dataset	Env	NC	Subj	TF	NP	Eval. Metrics	Remarks
Human3.6m	Indoor	4	11	3.6M	Single	MPJPE;N-/PA-MPJPE	Largest motion capture dataset
HumanEva-I	Indoor	7	4	12 seq	Single	MPJPE; PA-MPJPE	Motion capture; 4 gray scale + 3 color cameras
Total Capture	Indoor	8	5	1.892M	Multi	MPJPE; PA-MPJPE	MoCap; 13 IMUs
MPI-INF-3DHP	Indoor	14	8	1.3M	Single	PCK; AUC; MPJPE	Motion capture; data augmentation
Shelf	Indoor	3	5	-	Multi	PCK	Heavy occlusion
Campus	Outdoor	3	3	-	Multi	PCK	Outdoor Environment
CMU Panoptic	Indoor	521	8	1.5M	Multi	AP; MPJPE	Large social interaction
MuPoTS-3D	Indoor + outdoor	N	8	>8K	Multi	PCKrel; PCKabs	From MPI-INF-3DHP; 5 indoor; 15 outdoor
3DPW	Outdoor	1	7	60 seq	Multi	MPJPE; PA-MPJPE	in-the-wild; IMUs; phone cameras
SURREAL	Lab	N	145	6 M	Single	MPJPE; PA-MPJPE	Rendered from mocap data(H3.6M)
JTA	Synthetic	N	>21	500k	Multi	PCK; F1	Urban scenarios

NC no. of cameras, *NS* no. of subjects, *TF* total frames, *NP* no. of persons, *I/V* image/video, *seq* sequences, 3DHP & 3DPW are useful for cross-dataset evaluation

the Area Under the Curve (AUC) is calculated by plotting the PCK across various threshold distances, and MPJPE is also used for evaluation.

HumanEva (Sigal et al. 2010) comprises two datasets: HumanEva-I and HumanEva-II. HumanEva-I is larger, while HumanEva-II features higher-quality motion capture data using a more advanced hardware system. HumanEva-I includes data from 4 subjects performing 6 distinct actions, such as walking, boxing, jogging, throw/catch, and gesturing, recorded from three camera angles at 60 Hz. In contrast, HumanEva-II involves only 2 subjects executing an extended sequence of actions, referred to as 'Combo.' For instance, a subject may start walking along an elliptical path and then transition to jogging in the same direction. HumanEva-I provides training and validation data intended to be used across both datasets, while the test data is derived from HumanEva-II. 3D-HPE methods typically use MPJPE and PA-MPJPE as evaluation metrics for the HumanEva dataset.

MuCo-3DHP and MuPoTS-3D (Mehta et al. 2018) datasets are designed for multi-person 3D human pose estimation. The MuCo-3DHP dataset serves as a training set and is created by compositing the MPI-INF-3DHP dataset (Mehta et al. 2017). In contrast, the

MuPoTS-3D dataset is employed as the test set, encompassing 20 scenes captured in natural, real-world environments. Evaluation for these datasets is performed using the 3D Percentage of Correct Keypoints (3DPCK) metric.

KTH multiview football dataset II (Kazemi et al. 2013) features images of football players, including 2D and 3D poses centered on the root joint as ground truth. While the dataset offers partial camera parameters for projecting these 3D poses into three different views, specific intrinsic and extrinsic parameters are not explicitly defined. The camera configuration used to capture the dataset differs significantly from those used in training scenarios for Human3.6M and MPI-INF-3DHP, making this dataset notably challenging by comparison.

The total capture dataset (Trumble et al. 2017) offers synchronized video, IMU (Inertial Measurement Unit), and Vicon labeling across a substantial dataset of approximately 1.9 million frames captured by eight cameras. It includes recordings of four male and one female subjects, each performing five distinct activities: Range of Motion (ROM), Walking, Acting, Running, and Freestyle. Each activity is repeated three times. According to the dataset's structure, the training set comprises ROM-1,2,3, Freestyle-1,2, Walking-1,3, Acting-1,2, and Running-1 performed by subjects 1, 2, and 3. The testing set consists of Freestyle3 (FS3), Acting3 (A3), and Walking2 (W2) performed by subjects 1 to 5. The standard evaluation metrics for this dataset are MPJPE and PA-MPJPE.

The CMU panoptic (Joo et al. 2015) is a large-scale real-world indoor 3D human pose dataset featuring 65 video sequences depicting everyday activities. Captured within the Panoptic studio, this dataset provides detailed 3D pose annotations for multiple individuals engaged in various social interactions and daily routines. The standard evaluation metric for this dataset is the Average Precision (AP) of 3D human root location, especially in multi-person scenarios. AP evaluates how effectively the model localizes the root of each individual in 3D space.

The occlusion-person dataset (Zhang et al. 2021) provides a synthetic dataset with occlusion labels for the joints. It includes thirteen human models dressed in various clothing styles placed across nine different scenes such as living rooms, bedrooms, and offices. Objects such as sofas and desks are strategically positioned to partially obscure certain body joints. Each scene is equipped with eight cameras capturing multi-view images, and depth maps provide precise 3D coordinates for 15 joints as ground truth. Occlusion labels for each joint in an image are determined by comparing its depth value to the depth of the corresponding 3D joint in the camera's coordinate system. Approximately 20% of the body joints in the dataset are occluded under these conditions.

JTA (Fabbri et al. 2018) is a synthesized dataset generated from scenes in the Grand Theft Auto V (GTA-V) game. This dataset features multiple persons, with up to 32 individuals appearing in a single frame, exhibiting diverse illumination, viewpoints, and occlusion scenarios. Notably, the dataset adjusts the size of person images based on the camera's proximity to each individual, resulting in significant variations in person size as crowds span from close to far distances relative to the camera across different scenes. This characteristic makes JTA valuable for research focused on camera-centric applications, despite its synthetic nature. The JTA dataset typically employs PCK and F1 Score for evaluating methods. The F1 Score provides a balance between the accuracy of detected keypoints and the model's ability to minimize incorrect predictions.

Campus dataset and shelf dataset (Belagiannis et al. 2014): The Campus dataset records interactions among three individuals in an outdoor setting, captured using three cameras. In contrast, the Shelf dataset documents four people dismantling a shelf, utilizing five cameras for recording. Notably, the Shelf dataset features a higher degree of occlusion compared to the Campus dataset.

3DPW (Marcard et al. 2018) comprises in-the-wild images captured using IMUs, encompassing both indoor and outdoor scenes. This dataset offers calibrated camera parameters and precise SMPL parameters. However, it includes only a limited number of occluded samples, which do not adequately represent the more intricate occlusion scenarios typically encountered in real-world environments.

SURREAL (Varol et al. 2017) is a synthetic dataset that includes 6 million frames from 1,964 video sequences featuring 115 subjects. The dataset comprises photo-realistic renderings depicting people under extensive variations in shape, texture, viewpoint, and pose.

In addition to the aforementioned datasets, researchers have introduced other 3D pose datasets to enrich the diversity of existing resources. For example, SLOPER4D (Dai et al. 2023) offers a scene-aware dataset situated in an urban environment, providing contextually relevant data. Another dataset, mRI (An et al. 2022), focuses on multi-modal 3D pose data, catering to scenarios involving multiple sensory inputs.

The generalization and robustness of deep learning-based approaches largely depend on the availability of large and diverse datasets. Researchers employ various strategies to maximize the effectiveness of these datasets. A common approach is using multiple benchmark datasets such as Human3.6M and MPI-INF3DHP, where models are typically trained on Human3.6M and tested on MPI-INF3DHP to evaluate their generalization capability. Additionally, various online and offline pose augmentation techniques, as outlined in Sect. 6.3, have been explored to enhance the training data by introducing more variations and complexities. To address the challenge of limited annotated data, recent methods increasingly adopt self-supervised learning strategies as explained in Sect. 5. To bridge the domain gap, some approaches incorporate inference-time optimization during testing. Depending on the application context and the availability of datasets, implementing one or more of these strategies can improve model performance.

6.2 Evaluation metrics

Mean per joint position error (MPJPE) It measures the average Euclidean distance between the ground truth positions of joints and their corresponding predicted positions. A lower MPJPE value indicates better performance of a pose estimation model. Mathematically, MPJPE is calculated as follows:

$$E_{MPJPE}(f, S) = \frac{1}{N_S} \sum_{i=1}^{N_S} \|P_{f,S}^{(f)}(i) - P_{gt,S}^{(f)}(i)\|_2 \quad (1)$$

where S denotes the corresponding skeleton in the frame f , and N_s is the number of considered joints. The estimated position of joint i^{th} is given by $P_{f,S}^{(f)}(i)$ whereas $P_{gt,S}^{(f)}(i)$ is the corresponding ground truth of i^{th} joint. MPJPE measures the Euclidean distance between ground truth and predicted joint positions calculated frame by frame, and then averaged.

N-MPJPE normalizes MPJPE to a standard scale, while PA-MPJPE uses Procrustes analysis to align predictions to ground truth through translation, rotation, and scaling.

Percentage of correct key-points (PCK) measures how accurately keypoints are localized within a specified threshold distance, often set at 150 mm in many studies. Higher PCK values indicate better model performance. PCK_{abs} evaluates keypoint accuracy based on absolute camera positions, while PCK_{rel} assesses accuracy relative to person-centric joint positions. Area Under Curve (AUC) is computed across various PCK thresholds, providing a comprehensive assessment of model performance where higher AUC values indicate superior performance.

Percentage of correct parts (PCP) is a metric used to assess the accuracy of 3D limb detection. It determines whether a body part is correctly estimated based on the distance between predicted joints and their corresponding ground truth positions, with the criterion that this distance should be less than half the length of the limb. This specific variant is often referred to as PCP@0.5. One significant limitation of PCP@0.5 is its tendency to penalize shorter limbs more severely which can affect the overall evaluation of the model's performance in datasets with varying limb lengths.

3D percentage of correct key-points absolute ($3DPCK_{abs}$) (Moon et al. 2019) measures the accuracy of keypoint localization in absolute camera-centered coordinates without root alignment. It evaluates how well a model can predict the positions of keypoints relative to the camera without adjusting for the root joint. To assess the localization accuracy of the absolute 3D human root, the Average Precision of 3D Human Root Location (AP_{25}^{root}) is utilized. AProot25 considers a prediction correct if the Euclidean distance between the predicted and ground-truth root joint coordinates is smaller than 25 cm.

Others Recent advancements in 3D human pose estimation have introduced new metrics and methodologies to improve accuracy and training efficiency. The Pose SIMilarity (PSIM) Metric (Lee et al. 2023) is based on perceptual mechanisms that enhance evaluation by incorporating structural, positional, and intrinsic similarities, capturing complex spatial and temporal dependencies in human postures. Meanwhile, the Triangulation Residual loss (TR loss) (Zhao et al. 2023a) adopts global multiview geometric consistency for self-supervised learning, surpassing traditional methods that rely on local pairwise consistency and expensive epipolar constraints. These innovations aim to elevate performance and reduce computational costs in 3D pose estimation research.

6.3 Pose data augmentation

Deep learning models require large datasets and face significant challenges in 3D pose estimation due to the high cost and time required for accurate 3D pose annotations. The training data is biased towards indoor environments and common activities which results in poor generalization in diverse, unconstrained environments (Li et al. 2020a; Gong et al. 2021). The limited number of actors and actions in existing datasets make 3D pose estimators tend to overfit, which can be seen from the performance degradation of the algorithm on cross-datasets, especially for rare and complex poses (Li and Pun 2023). Pose data augmentation methods are proposed to solve these bottleneck problems and broadly categorized into offline and online techniques based on the phase of data generation and training as shown in Fig. 7.

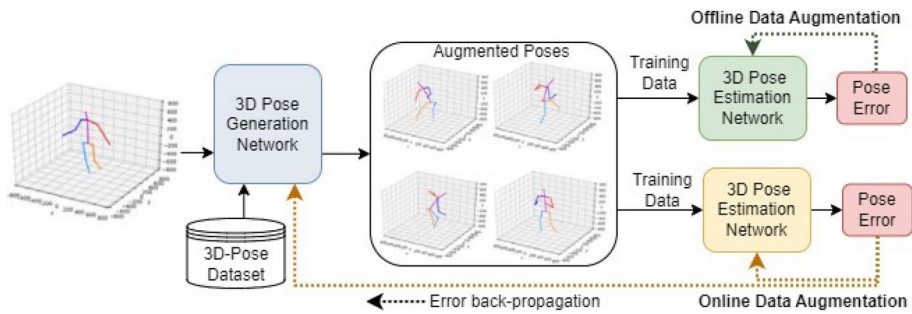


Fig. 7 Working flow of offline vs online pose data augmentation

In offline technique, pose data is generated first and the 3D-HPE model uses it for training, while online techniques train the pose generation and human pose estimation network simultaneously. The differentiable capability of online methods allows the augmentor to be co-optimized with the 3D pose estimator, using estimation errors as feedback to generate more varied poses continuously. The pose data augmented from offline technique are reusable by similar task models while online pose augmentation takes advantage of training both the network in parallel with mutual correction.

6.3.1 Offline pose data augmentation methods

LCR-Net++ (Rogez et al. 2020) generates pseudo ground-truth 3D joint locations from real-world images with 2D annotations. A large corpus of MoCap 3D poses is projected orthographically onto multiple random virtual views to create an extensive set of 2D poses with corresponding oriented 3D poses. The closest recovered pose is utilized to define the pseudo ground-truth full-body 3D pose, which is then used for training the network. Li et al. (2020a) introduce a method to explore the 3D human pose space using evolutionary operators such as crossover and mutation guided by pose prior knowledge. Cho et al. (2021) propose a method for generating synthetic distorted data from undistorted 2D keypoints, facilitating cross-scenario research on camera distortion.

Guan et al. (2023) propose a 3D pose generator named PoseGU which uses a minimal number of seed samples from a ground truth dataset. The method incorporates Counterfactual Risk Minimization to ensure unbiased learning in 3D pose estimation by correcting the distribution mismatch between the ground truth and the generated 3D pose datasets. Chai et al. (2023) propose a method with an augmentation generator creating the parameters for transformations to bone angle, bone length, and rotation from the source dataset, and an anchor discriminator ensures the realism and quality of the transformed pose pairs. Park et al. (2023) propose a geometry-aware data augmentation strategy by employing techniques such as person translation, person rotation, ground plane translation, and ground plane rotation. The method introduces noise or removes occluded keypoints during augmentation using a person volumetric model to simulate occlusion.

6.3.2 Online pose data augmentation methods

Cheng et al. (2020) implement a method for making certain keypoints or frames by setting the corresponding heatmaps to zero for a temporal convolution network to address different occlusion scenarios. Guo et al. (2021) utilize techniques such as rotation and horizontal flipping during training. They crop input images and adjust the bounding boxes to address inter-person occlusion in multi-person pose estimation. Gong et al. (2021) introduce PoseAug to modify geometric factors such as posture, body size, viewpoint, and position using differential operations. PoseAug employs a part-aware Kinematic Chain Space in the discriminator to ensure the validity of the augmented data. Xu et al. (2022) propose a pose simulator that augments training samples with virtual 2D poses by projecting each 3D pose into different 2D poses from various camera angles, which aids in learning geometric features. Han et al. (2022) incorporate augmentation techniques during training such as horizontal flip, rotation, synthetic occlusion, and color jittering.

Zhang et al. (2023a) propose a model that integrates the augmentor, the estimator, and the discriminator. The differentiable augmentor adjusts major geometric aspects of the 2D-3D pose pairs to better align with the estimator, using training error as feedback. The part-aware 3D discriminator broadens the range of feasible augmented poses through localized supervision, ensuring the plausibility and diversity of joint angles. Moreover, body part properties, such as leg characteristics (Wang et al. 2023), which play a key role in human locomotion, can be incorporated to enhance pose diversity. Li and Pun (2023) propose a generator to broaden the distribution of each joint angle within the existing dataset while constraining them to a plausible range using an adaptive interpolation discriminator. Du et al. (2024) combine evolutionary data augmentation with an anatomy-aware global-local representation for 3D human pose estimation. Gholami et al. (2024) proposes a NeRF-based (Mildenhall et al. 2020) model called PoseGen for pose data augmentation from images. PoseGen includes a generator that produces 3D human poses and camera viewpoints which are evaluated by a discriminator to ensure their plausibility. The 3D poses and camera viewpoints are then input into a NeRF model to render corresponding human images.

7 Performance of 3D-HPE methods

This section outlines the evaluation of recent advancements in 3D-HPE across several benchmark datasets. Human3.6M stands out as the most commonly used dataset, serving as the foundation for state-of-the-art models. These models leverage multiple datasets, including Human3.6M itself, MPI-INF-3DHP, MuPoTS-3D, CMU Panoptic, Total Capture, Campus, Shelf, and JTA. Single-person methods primarily rely on Human3.6M for training and evaluate their performance using metrics such as MPJPE, PA-MPJPE, and N-MPJPE. To ensure robustness, models trained on Human3.6M are validated against MPI-INF-3DHP using metrics such as 3DPCK, AUC, and MPJPE.

Despite its focus on single-person scenarios, Human3.6M is also used as a benchmark for evaluating multi-person 3D HPE models against state-of-the-art methods. MuPoTS-3D serves as a critical benchmark for multi-person monocular 3D pose estimation, offering datasets for assessing performance using person-centric (PCK_{rel}) and camera-centric (PCK_{abs}) metrics. CMU Panoptic is essential for evaluating models in complex multi-

view, multi-person scenarios. Additionally, smaller datasets like Campus, Shelf, and JTA are utilized to test models across various settings in multi-person 3D human pose estimation. Overall, these datasets provide comprehensive benchmarks that enable researchers to evaluate and compare the effectiveness of their 3D-HPE models across different challenges and scenarios.

The performance of recent works on benchmark datasets is summarized in tables, with some comparisons also illustrated through graphs to provide a clear comparative analysis of the models. These methods are classified based on their deep learning algorithms for different scenarios, including monocular/multi-view approaches for single/multi-person 3D-HPE, using either image or video inputs. Most methods demonstrate improvements when using ground truth 2D poses as input. However, results based on 2D poses estimated by pose detectors are considered for practical purposes.

Table 3 illustrates the performance results of monocular image and video-based single-person 3D human pose estimation (HPE) methods on the Human3.6M dataset and cross-validation on the MPI-INF-3DHP dataset. Figure 8 visually compares the performance of image-based methods for single-person 3D pose estimation using a chart. This comparison shows that GCN-based methods achieve the best performance on the Human3.6M dataset, while GAN-based methods perform best during cross-validation on MPI-INF-3DHP. Similarly, Fig. 9 illustrates the performance of video-based methods, showing that a hybrid approach combining GCN and Transformer delivers the best results on both datasets.

Table 4 presents a comparison of the performance of Multi-view Single-Person 3D Human Pose Estimation (HPE) methods across the Human3.6m, Total Capture, and CMU Panoptic datasets. The CNN-based method, Adafuse Zhang et al. (2021), demonstrates strong performance with multi-view image inputs, while the transformer-based method Cai et al. (2024b) excels with multi-view video inputs.

Table 5 outlines the performance of monocular 3D multi-person pose estimation methods on the MuPoTS-3D and JTA datasets, covering both image-based and video-based inputs. A graphical chart for the MuPoTS-3D dataset is shown in Fig. 10. The comparative analysis reveals that CNN-based methods continue to perform well for monocular image-based approaches, while GCN-based methods demonstrate strong performance with video input techniques.

Similarly, Table 6 highlights the performance of Multi-view Multi-person 3D Human Pose Estimation (HPE) methods on the CMU Panoptic, Campus, and Shelf datasets, with a visual representation provided in Fig. 11. The comparison shows that transformer-based methods using a bottom-up approach perform comparably to GCN-based methods that utilize a top-down approach for multi-view image inputs. For multi-view video-based methods, the top-down approach in voxel space performs best for multi-person scenarios. The top results are highlighted in **bold** for clearer visibility.

8 Challenges and future directions

8.1 Challenges

Researchers have dedicated significant effort to addressing the challenges in 3D-HPE and have achieved considerable success to some extent. Given the diverse nature of 3D-HPE, it

Table 3 Performance of monocular single-person 3D-HPE methods on **Human3.6M** & **MPI-INF-3DHP** datasets

Algorithm	Method	Human3.6M		MPI-INF-3DHP			Remarks
		MPE↓	P-MPJPE↓	PCK↑	AUC↑	MPE↓	
CNN	Luvizon et al. (2023)	55.5	-	83.2	44.3	96.8	End-to-end; supervised
	Zhang et al. (2020)	-	-	83.6	48.2	92.2	ISO; self-supervised
	Wandt et al. (2021)	81.9	53.0	69.0	-	119.2	Self-supervised
	Xu et al. (2022)	58.1	43.8	-	-	-	Online data aug
	Kundu et al. (2022)	103.2 S->H	88.9 S->H	-	-	96.2 (92.9 ISO)	S->H: trained on SURREAL & tested on H3.6M
GCN	Xu et al. (2021)	85.3	59.8	68.2	35.2	-	Unsupervised
	Xu and Takano (2021)	51.9	-	80.1	45.8	-	2D-to-3D; supervised
	Zou and Tang (2021)	49.4	39.1	86.1	53.7	-	2D-to-3D; supervised
	Ci et al. (2022)	44.2	37.3	74.0	36.7	-	2D-to-3D; supervised
	Zhang et al. (2023b)	49.1	34.9	86.9	55.0	-	2D-to-3D; supervised
Transformer	Li et al. (2023a)	50.5	-	84.1	53.7	-	Uncertainty learning
	Zhou et al. (2024a)	46.4	-	88.2	59.3	68.9	2D-to-3D; supervised
GCN+ Transformer	Zhao et al. (2022)	51.8	-	79.0	43.8	-	2D-to-3D; supervised
	Chen et al. (2024)	49.8	38.9	85.5	53.6	80.4	2D-to-3D; supervised
Autoencoder	Nie et al. (2023)	50.3	37.3	83.5	45.9	83.03	Semi-supervised
GAN	Chai et al. (2023)	49.9	34.2	92.1	62.5	61.3	Unsupervised; offline data aug
	Li and Pun (2023)	47.3	36.8	89.9	58.2	69.7	Supervised; online data aug
Grid Conv	Kang et al. (2023)	46.3	37.6	89.2	57.6	72.1	Grid Convolution

Monocular video-based single-person 3D-HPE methods

Table 3 (continued)

Algorithm	Method	Human3.6M		MPI-INF-3DHP			Remarks
		MHE↓	P-MPJPE↓	PCK↑	AUC↑	MHE↓	
CNN	Liu et al. (2021)	44.8	35.6	-	-	-	2D-to-3D;supervised
	Lee et al. (2023)	40.5	31.8	82.6	48.8	81.3	Probabilistic
GCN	Zeng et al. (2021)	47.9	-	82.1	46.2	-	2D-to-3D;supervised
	Yu et al. (2023)	44.4	-	98.53	79.12	27.76	2D-to-3D;supervised
Transformer	Honari et al. (2023)	100.3	74.12	-	-	93.0	End-to-end; CSS
	Tang et al. (2023)	40.5	31.8	98.7	83.9	23.1	2D-to-3D;supervised
	Zhao et al. (2023b)	45.2	35.6	97.9	78.8	27.8	Frequency domain
	Kim et al. (2024)	69.0	50.3	-	-	102.1	Probabilistic;Self-supervised
	Cai et al. (2024a)	39.7	-	98.5	78.1	29.2	Diffusion-based
	Zhou et al. (2024b)	42.6	34.2	98.2	75.7	36.4	Hierarchical supervision
GCN+ Transformer	Li et al. (2023b)	42.5	34.2	94.8	65.8	54.0	Probabilistic
	Gong et al. (2023)	36.9	-	98.0	75.9	29.1	Diffusion-based
	Chen et al. (2023)	40.3	31.7	98.7	72.9	37.2	2D-to-3D;supervised
	Zhai et al. (2023)	41.1	32.5	87.2	57.0	-	2D-to-3D;supervised

ISO inference stage optimization, CSS contrastive self-supervised learning, MPJPE in mm, PCK/AUC in %, **bold**: best result

encounters various challenges across different scenarios. For instance, the challenges associated with monocular input techniques differ from those related to multi-view techniques. Likewise, multi-person methods encounter more challenges than single-person methods. These challenges can be broadly categorized into common and scenario-specific challenges as shown in Fig. 12.

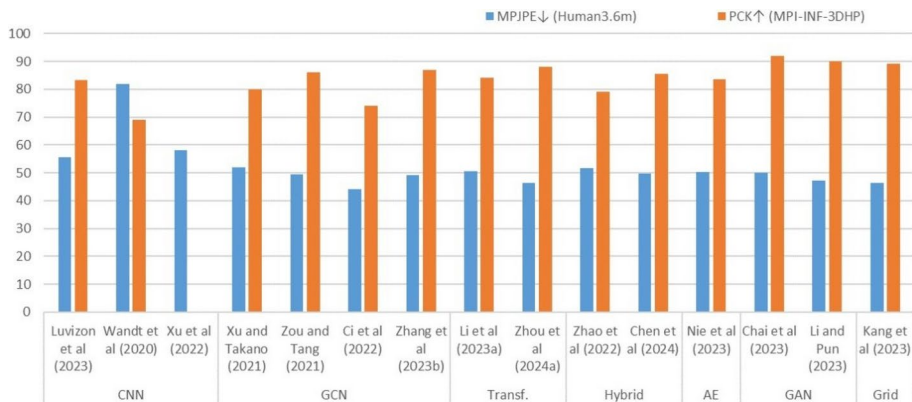


Fig. 8 Performance of monocular image-based methods for single-person 3D-HPE

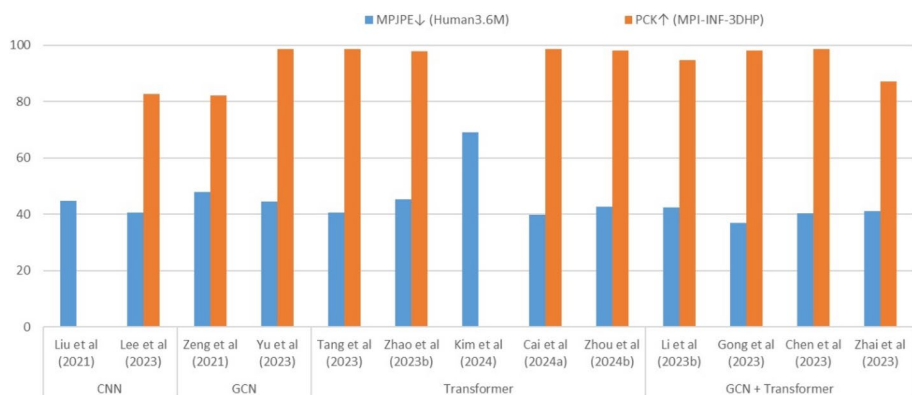


Fig. 9 Performance of monocular video-based methods for single-person 3D-HPE

8.1.1 Common challenges

Articulated human body: The human body's structure and pose are highly variable and complex, primarily due to the large number of degrees of freedom involved in its mechanics. Each joint in the body allows movement along multiple axes, leading to a vast range of possible configurations and poses. This variability poses significant challenges for accurately estimating 3D-HPE. Moreover, differences in body shape, size, and proportions among different individuals add to the complexity, requiring models to generalize across a wide range of body types.

Dense ambiguity: Image projections from different 3D poses can appear very similar because the transition from 3D to 2D inherently loses depth information. As a result, multiple distinct configurations of 3D keypoints can produce identical 2D keypoints projections on an image plane. Consequently, simply training a model to convert 2D-pixel positions to 3D world coordinates is likely to be unsuccessful. Monocular pose estimation in outdoor environments offers flexibility without needing camera calibration, but it encounters the

Table 4 Performance of multi-view single-person 3D HPE methods on **Human3.6m**, **total capture** & **CMU panoptic** datasets

Algorithm	Method	H3.6m	Total	Panoptic	Remarks
		$MPE \downarrow$	$MPE \downarrow$	$MPJPE \downarrow$	
CNN	Remelli et al. (2020)	30.2	27.5	-	Supervised; 2D-to-3D
	Zhang et al. (2021)	19.5	19.2	14.7	Panoptic(4-views); supervised; 2D-to-3D
	Luvi-zon et al. (2022)	39.0	-	-	Supervised; 2D-to-3D
	Wan et al. (2023)	21.1	23.0	-	Supervised; 2D-to-3D
	Jiang et al. (2023)	27.8	-	24.2	Panoptic(4-views); probabilistic Triangulation
<i>For video-based input</i>					
Transformer	Shuai et al. (2023)	26.2	24.6	-	T=27; no camera params; CPN (H3.6m) Res101 (TC)
	Cai et al. (2024b)	25.4	15.0	-	T=27; no camera params; CPN; ViTPose (15.1); Res101 (TC)
Hybrid	Zhang et al. (2024)	27.6	-	-	T=27; no camera params; CPN; Graph Transformer

T=27 means 27 frames, CPN Res101 as 2D backbone, MPJPE in mm, **bold**: best result

severe challenge of an ill-posed problem as it lacks additional data unlike video or multi-view setups that benefit from temporal or multiple viewpoint information.

Limited 3D annotations: Deep learning techniques require substantial amounts of data to perform optimally. However, the scarcity of 3D ground-truth data poses a significant challenge to the success of 3D human pose estimation (HPE). Most widely used datasets, such as Human3.6M and HumanEva, are collected in controlled indoor environments using motion capture systems. As a result, models trained on these datasets tend to struggle with generalization, despite achieving high performance within the confines of the datasets themselves.

Limited diversity datasets: Many paired 3D pose datasets exhibit limited diversity in terms of pose variety, appearances (including background and lighting conditions), and out-fits. Consequently, existing methods for 3D human pose estimation frequently struggle with inferior generalization performance when faced with out-of-distribution (OOD) samples. This challenge largely stems from the restricted diversity of 2D-3D pose pairs available in the training data.

Table 5 Performance of monocular multi-person 3D HPE methods on **MuPoTS-3D** & **JTA** datasets

Algorithm	Method	MuPoTS-3D t=15 cm		JTA (t in m) F1 $\uparrow$			Remarks
		PCKrel $\uparrow$	PCKabs $\uparrow$	t=0.4	t=0.8	t=1.2	
CNN	Rogez et al. (2020)	70.6	-	-	-	-	Topdown; Offline DataAug; Probabilistic
	Benzine et al. (2020)	72.0	-	3DPCK=83.2 (all)			Supervised; Topdown; No F1 for JTA
	Jin et al. (2022)	80.9	39.3	-	-	-	Supervised; Single-stage
	Wang et al. (2022)	82.7	39.2	-	-	-	Supervised; Single-stage
GCN	Han et al. (2022)	79.7	-	-	-	-	Uncertainty; Topdown; Data Aug(online)
	Chen et al. (2022)	77.8	-	-	-	-	Supervised; Top-down
Autoencoder	Fabbri et al. (2020)	-	-	50.82	64.76	70.44	Supervised; Bottomup; voxel
<i>Monocular video-based methods for multi-person</i>							
GCN	Cheng et al. (2021a)	87.5	45.7	-	-	-	Supervised; Top-down
	Cheng et al. (2021b)	89.6	48.0	57.22	68.51	72.86	SemiSupervised; Topdown +bottomup
	Cheng et al. (2023)	89.6	48.1	58.15	69.32	74.19	SemiSupervised; ISO; Hybrid
Transformer	Park et al. (2023)	83.7	50.9	-	-	-	Volumetric DataAug. (offline)

PCK/F1 in %, t denotes threshold for F1 as the point is true-positive when the distance from the target point is less than t

Objects and self-occlusion: Some parts of the person may be occluded due to opaque objects so it is difficult to estimate the pose correctly. In some specific poses, the person might have self-occluded. One part of the body may hide into another part of the body which makes it difficult to calculate the pose of the person.

Computational complexity: The performance of existing models has improved significantly, but this often comes at the expense of increased computational resources. With the widespread adoption of handheld devices such as mobile phones and tablets, which are typically resource-constrained, maintaining the performance of 3D-HPE models becomes highly challenging.

Environmental factors: Variations in lighting, background clutter, clothing, and camera angles can affect the accuracy of pose estimation models. It is difficult to incorporate outdoor environments where lighting, background conditions, camera viewpoints as well as the behavior of human subjects are diverse and unconstrained.

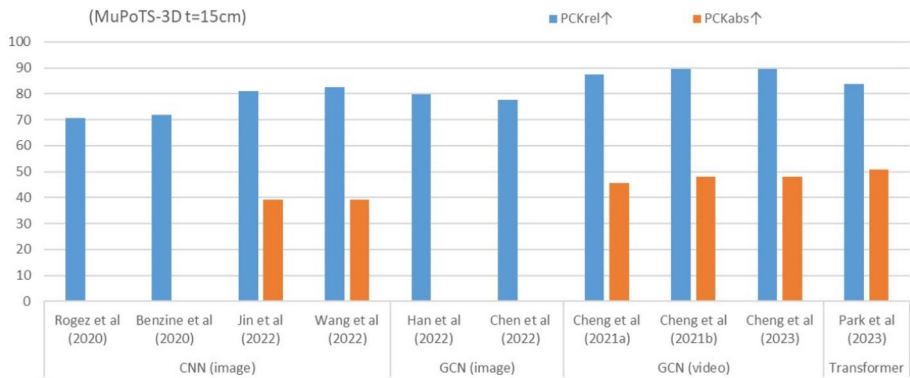


Fig. 10 Performance of monocular methods for 3D multi-person pose estimation

Table 6 Performance of multi-view multi-person 3D-HPE methods on **CMU panoptic, Campus & Shelf** datasets

Algorithm	Method	Panoptic			Campus	Shelf	Remarks
		AP25↑	mAP↑	MPJPE↓	PCK↑	PCK↑	
CNN	Tu et al. (2020)	83.59	95.39	17.68	96.7	97.0	Supervised;Bottomup;Voxel
	Lin and Lee (2021)	92.12	97.68	16.75	97.0	97.9	Supervised;Topdown
GCN	Wu et al. (2021)	-	98.18	15.84	-	97.7	Supervised: Topdown
Trasformer	Wang et al. (2021b)	92.30	96.02	15.80	96.6	97.4	Supervised;Bottomup
<i>Multi-view video-based methods for multi-person</i>							
CNN	Chen et al. (2020)	-	-	-	96.6	96.8	Supervised;Topdown
	Reddy et al. (2021)	-	-	7.3	97.4	98.2	Supervised;Topdown; voxel
	Dong et al. (2022)	Qualitative analysis			96.3	96.9	Supervised;Topdown
	Zhang et al. (2023c)	-	-	16.97	96.7	97.1	Supervised; Bottomup;Voxel

MPJPE in mm, PCK/AP in %, mAP: mean value of AP, **bold**: best result

8.1.2 Challenges in monocular setup

Generalizability: Monocular image methods are constrained by their inability to utilize additional cues such as multi-view or temporal information. Consequently, they rely heavily on large and diverse datasets for training, which are often limited in availability. This reliance on restricted diversity within existing 3D pose datasets significantly hampers the generalizability of supervised 3D pose estimation models.

Motion blur: Fast movements can induce motion blur in images or video, making it challenging to accurately detect and estimate body parts. Many existing approaches treat the

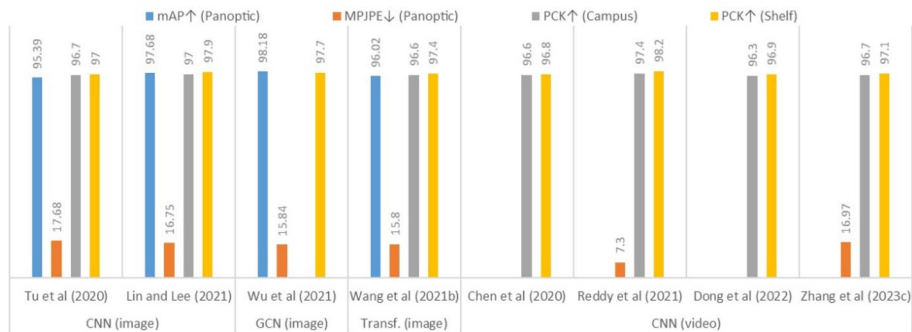


Fig. 11 Performance of multi-view methods for 3D multi-person pose estimation

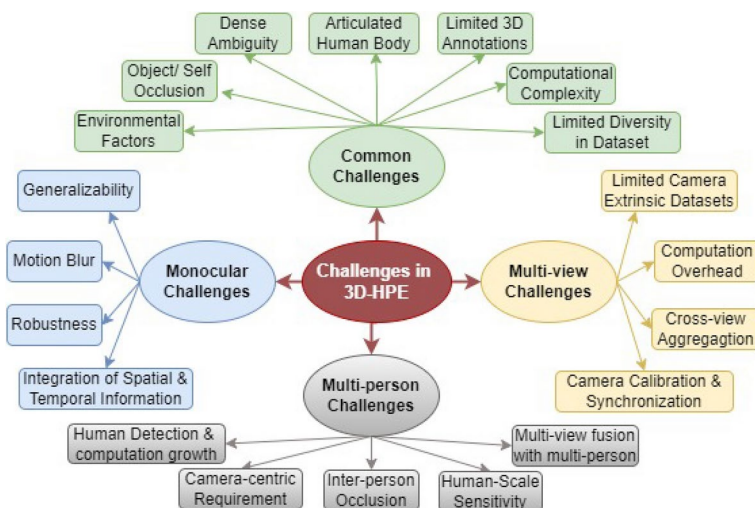


Fig. 12 Common and specific challenges in 3D-HPE

3D HPE problem simply as a coordinate regression issue, neglecting the inherent kinematic structure of the human body. This oversight frequently results in inaccurate outcomes.

Integration of spatial and temporal information: Utilizing temporal information from video sequences is essential for addressing depth ambiguity and occlusion in human pose estimation. While videos provide sequential variations beneficial for understanding body structure, integrating spatial and temporal information effectively remains challenging. Current methods primarily focus on frame-level feature correlations and often overlook the relationships between joints across frames. Future research needs to explore these interdependencies to improve the accuracy of pose estimation models.

Robustness: Existing methods often experience decreased performance when the target person is significantly small or large, or when their motion is too fast or slow compared to the scale and speed of the training data. Addressing these challenges requires adaptive extraction of temporal features from videos of varying lengths and ensuring compatibility with single-image input for versatility in different contexts.

8.1.3 Challenges in multi-view setup

Camera synchronization and calibration: Multi-view methods for 3D human pose estimation necessitate synchronized and calibrated cameras, which adds complexity to the process. A fundamental challenge in this approach is depth uncertainty, particularly when camera parameters are unknown. Many multi-view methods rely on precise camera calibration to effectively utilize multi-view information. However, accurately calibrating cameras synchronously in dynamic capture scenarios is unrealistic. While incorporating camera parameters can provide strong prior knowledge of networks, achieving accurate calibration in dynamic environments remains difficult.

Limited camera extrinsic datasets: Current datasets used for multi-view human pose estimation suffer from a lack of diversity in camera external parameters. Consequently, models trained on these datasets often struggle to generalize effectively to unseen camera configurations. This limitation necessitates reannotation and retraining of models whenever there are changes in camera setups. Moreover, existing models are not robust enough to handle unseen views, such as uncommon camera angles or distances. Due to their training on limited data, these models typically excel only with test examples that closely resemble their training views, leading to significant performance degradation when confronted with unfamiliar perspectives.

Cross-view aggregation: Multi-view methods for 3D human pose estimation require effective fusion of multi-view and/or multi-frame features to aggregate information efficiently. Methods like Triangulation or pictorial structure models, which reconstruct 3D poses from 2D views, can be prone to errors in crowded or complex scenes. Volumetric approaches use discretized 3D representations to gather and process multi-view features directly in 3D space, but they face challenges due to the cubic increase in computational complexity with the size of the volumetric space. Real-time estimation of 3D poses for multiple humans is particularly challenging due to the complexity of associating 2D poses across different views and the immense state space involved, especially in crowded environments where interactions between multiple individuals further complicate cross-view associations.

Increased computation overhead: Using multiple cameras for data capture introduces complexity, as it involves synchronizing and calibrating the cameras to obtain accurate camera parameters and integrate information effectively. As the number of people and cameras increases, the computational complexity escalates significantly.

8.1.4 Challenges in multi-person

Inter-person occlusion: Multi-person human pose estimation models face the challenge of handling occlusions between individuals due to interactions among people and with physical objects in the environment. Besides self-occlusions where limbs can obstruct other body parts, occlusions from other individuals or objects add further difficulty to pose estimation. The proximity and interactions among multiple people make it unreliable to consistently group human joints.

Camera-centric requirements: Single-person 3D human pose estimation methods typically use person-centric coordinates relative to the target person's center, which cannot be directly applied to multi-person 3D human pose estimation requiring camera coordinates.

The main challenge arises because using person-centric coordinates obscures the location of each individual within the scene, making it unclear where each person should be positioned. Absolute 3D human pose estimation presents a more complex problem where predictions are made relative to a fixed reference camera position, ensuring that all estimated skeletons are correctly placed in a unified 3D space.

Human detection error and linear computation growth: Identifying multiple individuals in images is difficult due to the variability in human sizes, crowded scenes, and close interactions between people. Using a top-down approach to estimate the pose of each person individually results in computational costs that increase proportionally with the number of people in the image.

Human-scale sensitivity and redundancy: Bottom-up approaches estimate poses for all individuals in an image simultaneously and subsequently group keypoints belonging to each person, which introduces redundant computational complexity.

Multi-view fusion with multi-person: Integrating information from multiple-view images with multiple individuals is particularly challenging because it requires identifying the identities of each person across all cameras. Additionally, features from different views that correspond to the same individual need to be recognized and grouped for 3D pose estimation.

8.2 Future directions

Notable emerging trends and needs, including LLM-based 3D-HPE, real-time processing, and scene-aware 3D pose estimation in dynamic environments, have been addressed in section 2.6. Despite progress in recent techniques, there are still numerous areas where improvements can be made. This section highlights the most promising directions for future research and development.

8.2.1 Occlusion handling

Occlusion is widely regarded as a significant challenge in 3D-HPE. Estimating 3D human poses becomes particularly difficult due to various types of occlusion, such as inter-person and intra-person occlusion, as well as occlusion caused by objects. Combining the following strategies could provide an effective solution for addressing occlusion challenges.

- Datasets that incorporate occlusion scenarios, such as The Occlusion-Person Dataset and the data augmentation techniques to train the model with occluded datasets, can be utilized for making the model robust to the occlusion.
- Additionally, the diffusion-based model can be used to learn the underlying structure of human poses and generate plausible estimates for occluded body parts.
- NeRF can generate novel views that account for occluded body parts, enabling more accurate pose estimation by utilizing complementary information from novel viewpoints.

8.2.2 Strategies for addressing data inadequacy

Creating in-the-wild scenario datasets for 3D human pose annotation is highly expensive. Most datasets are created in lab environments, with only a few available for in-the-wild scenarios, which are insufficient for deep learning models. This scarcity results in decreased model performance in uncontrolled environments or complex poses. Consequently, it is crucial to create innovative strategies to address data limitations, and our recommendations are as follows:

- By employing techniques such as contrastive learning, pose consistency, or reconstruction-based objectives, self-supervised learning can improve generalization and performance, even in scenarios where annotated data is scarce or incomplete.
- Pose Data Augmentation techniques can be utilized to generate diverse pose pairs which can be further improved by introducing negative samples, such as backgrounds and joint-level occlusions.
- Inference Time Optimization techniques can refine estimated 3D poses without permanently updating the network, thereby improving the generalizability of 3D pose models across different pose distributions in source and target data.

8.2.3 3D multi-person pose estimation

Many real-world applications involve scenes with multiple persons, yet the field of 3D multi-person pose estimation remains immature. Enhanced methods are needed to address challenges in 3D-MPPE, including person interactions, variations in human size, and considerations related to camera-centric perspectives.

- Bottom-up approaches can efficiently process crowds when combined with adaptive scale-variation techniques to tackle the issue of varying human sizes.

8.2.4 3D-HPE from multi-view images

Existing multi-view pose estimation methods are more accurate than monocular methods but are hindered by the challenge of precise camera calibration and cross-view aggregation strategy, which is difficult to achieve synchronously in dynamic capture scenarios.

- Current methods for cross-view aggregation primarily rely on reconstruction-based and volumetric approaches. These methods can be further enhanced by incorporating NeRF-based technologies to improve the aggregation of information from multiple views.

9 Conclusion

This survey has provided a comprehensive overview of recent advancements and strategies in 3D human pose estimation. The landscape of 3D-HPE is evolving rapidly, driven by a diverse array of technologies including Convolutional Neural Networks, Graph Convolutional Networks, transformers, and their combinations. These technologies which are applied across various scenarios such as monocular/multi-view camera setups, single/multi-person scenarios, and image/video inputs, are reviewed thoroughly. Beyond traditional supervised learning, the survey explored emerging techniques such as weak, semi, and self-supervised learning methods, highlighting their contributions to improving accuracy and robustness in pose estimation. It explored online and offline pose data augmentation strategies to increase the generalizability of 3D-HPE models. It briefly discussed the importance of absolute keypoint coordinates, problem-solving spaces, probabilistic/ diffusion-based approaches, benchmark datasets, evaluation metrics, and the performance of recent methods. The categorization of challenges into common and scenario-specific issues offers clarity on current obstacles in the field. Moreover, the identification of future research directions with actionable recommendations highlights the ongoing opportunities for innovation and advancement in 3D-HPE.

Acknowledgements This work is supported by the Beijing Natural Science Foundation, China (Grant Nos. 4222037, L181010). We want to express our sincere gratitude to all those who contributed to this research.

Author Contributions Rama Bastola Neupane: Conceptualization, Data Curation, Investigation, Formal Analysis, Visualization, Writing-Original Draft; Manuscript Revision Kan Li: Conceptualization, Supervision, Resources; Tesfaye Fenta Boka: Writing-review and editing, Data Curation;

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Andriluka M, Pishchulin L, Gehler P, et al (2014) 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3686–3693
- An S, Li Y, Ogras U (2022) mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. In: Advances in neural information processing systems, pp 27414–27426

- Artacho B, Savakis A (2022) Unipose+: a unified framework for 2d and 3d human pose estimation in images and videos. *IEEE Trans Pattern Anal Mach Intell* 44:9641–9653. <https://doi.org/10.1109/TPAMI.2021.3124736>
- Azizi N, Possegger H, Rodolà E, et al (2022) 3d human pose estimation using möbius graph convolutional networks. In: European conference on computer vision, pp 160–178
- Bartol K, Bojanic D, Pribanic T (2022) Generalizable human pose triangulation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2022–June. IEEE Computer Society, pp 11018–11027, <https://doi.org/10.1109/CVPR52688.2022.01075>
- Belagiannis V, Amin S, Andriluka M, et al (2014) 3d pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1669–1676
- Benzine A, Chabot F, Luvison B, et al (2020) Pandanet : Anchor-based single-shot multi-person 3d pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6856–6865
- Benzine A, Luvison B, Pham QC et al (2021) Single-shot 3d multi-person pose estimation in complex images. *Pattern Recognit*. <https://doi.org/10.1016/j.patcog.2020.107534>
- Bertoni L, Kreiss S, Alahi A (2019) Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6861–6871
- Cai Q, Hu X, Hou S, et al (2024a) Disentangled diffusion-based 3d human pose estimation with hierarchical spatial and temporal denoiser. In: Proceedings of the AAAI conference on artificial intelligence, pp 882–890
- Cai Y, Zhang W, Wu Y, et al (2024b) Fusionformer: A concise unified feature fusion transformer for 3d pose estimation. In: AAAI conference on artificial intelligence, pp 900–908
- Chai W, Jiang Z, Hwang JN, et al (2023) Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. In: Proceedings of the IEEE international conference on computer vision. Institute of Electrical and Electronics Engineers Inc., pp 14609–14619, <https://doi.org/10.1109/ICCV51070.2023.01347>
- Chen Z, Huang Y, Yu H et al (2022) Learning a robust part-aware monocular 3d human pose estimator via neural architecture search. *Int J Comput Vision* 130:56–75. <https://doi.org/10.1007/s11263-021-01525-0>
- Chen L, Ai H, Chen R, et al (2020) 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, pp 3279–3288
- Chen Z, Dai J, Bai J et al (2024) Dgformer: dynamic graph transformer for 3d human pose estimation. *Pattern Recognit*. <https://doi.org/10.1016/j.patcog.2024.110446>
- Cheng Y, Wang B, Tan RT (2023) Dual networks based 3d multi-person pose estimation from monocular video. *IEEE Trans Pattern Anal Mach Intell* 45:1636–1651. <https://doi.org/10.1109/TPAMI.2022.3170353>
- Cheng Y, Wang B, Yang B, et al (2021a) Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. In: AAAI conference on artificial intelligence. AAAI
- Cheng Y, Wang B, Yang B, et al (2021b) Monocular 3d multi-person pose estimation by integrating top-down and bottom-up networks. In: IEEE conference on computer vision and pattern recognition. IEEE
- Cheng Y, Yang B, Wang B, et al (2020) 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In: Proceedings of the AAAI conference on artificial intelligence, pp 10631–10638
- Chen H, He JY, Xiang W, et al (2023) Hdformer: High-order directed transformer for 3d human pose estimation. In: Proceedings of the thirty-second international joint conference on artificial intelligence, pp 581–589
- Chen Y, Wang Z, Peng Y, et al (2018) Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7103–7112
- Chen X, Wei P, Lin L (2021) Deductive learning for weakly-supervised 3d human pose estimation via uncalibrated cameras. In: AAAI conference on artificial intelligence, pp 1089–1096
- Cho H, Cho Y, Yu J, et al (2021) Camera distortion-aware 3d human pose estimation in video with optimization-based meta-learning. In: Proceedings of the IEEE international conference on computer vision. Institute of Electrical and Electronics Engineers Inc., pp 11149–11158, <https://doi.org/10.1109/ICCV48922.2021.01098>
- Choi S, Choi S, Kim C (2021b) Mobilehumanpose: Toward real-time 3d human pose estimation in mobile devices. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2328–2338
- Choi H, Moon G, Chang JY, et al (2021a) Beyond static features for temporally consistent 3d human pose and shape from a video. In: IEEE conference on computer vision and pattern recognition. IEEE
- Ci H, Ma X, Wang C et al (2022) Locally connected network for monocular 3d human pose estimation. *IEEE Trans Pattern Anal Mach Intell* 44:1429–1442. <https://doi.org/10.1109/TPAMI.2020.3019139>

- Dai Y, Lin Y, Lin X, et al (2023) Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2023-June. IEEE Computer Society, pp 682–692, <https://doi.org/10.1109/CVPR.2023.00073>
- Desmarais Y, Mottet D, Slangen P et al (2021) A review of 3d human pose estimation algorithms for markerless motion capture. *Comput Vision Image Underst*. <https://doi.org/10.1016/j.cviu.2021.103275>
- Dong J, Fang Q, Jiang W et al (2022) Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE Trans Pattern Anal Mach Intell* 44:6981–6992. <https://doi.org/10.1109/TPAMI.2021.3098052>
- Du S, Yuan Z, Lai P et al (2024) Joypose: jointly learning evolutionary data augmentation and anatomy-aware global-local representation for 3d human pose estimation. *Pattern Recognit*. <https://doi.org/10.1016/j.patcog.2023.110116>
- Fabbri M, Lanzi F, Calderara S, et al (2018) Learning to detect and track visible and occluded body joints in a virtual world. In: Proceedings of the European conference on computer vision (ECCV), pp 430–446
- Fabbri M, Lanzi F, Calderara S, et al (2020) Compressed volumetric heatmaps for multi-person 3d pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7204–7213
- Fang HS, Xu Y, Wang W, et al (2018) Learning pose grammar to encode human body configuration for 3d pose estimation. In: Proceedings of the AAAI conference on artificial intelligence
- Feng Y, Lin J, Dwivedi SK, et al (2024) Chatpose: Chatting about 3d human pose. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2093–2103
- Gamra MB, Akhloufi MA (2021) A review of deep learning techniques for 2d and 3d human pose estimation. *Image Vision Comput*. <https://doi.org/10.1016/j.imavis.2021.104282>
- Gao Q, Deng Z, Ju Z et al (2023) Dual-hand motion capture by using biological inspiration for bionic bimanual robot teleoperation. *Cyborg Bionic Syst* 4:0052
- Gholami M, Wandt B, Rhodin H, et al (2022) Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2022-June. IEEE Computer Society, pp 13065–13075, <https://doi.org/10.1109/CVPR52688.2022.01273>
- Gholami M, Ward R, Wang ZJ (2024) Posegen: learning to generate 3d human pose dataset with nerf. In: AAAI conference on artificial intelligence, pp 1905–1913
- Gong J, Foo LG, Fan Z, et al (2023) Diffpose: toward more reliable 3d pose estimation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2023-June. IEEE Computer Society, pp 13041–13051, <https://doi.org/10.1109/CVPR52729.2023.01253>
- Gong K, Li B, Zhang J, et al (2022) Posetriplet: co-evolving 3d human pose estimation, imitation, and hallucination under self-supervision. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2022-June. IEEE Computer Society, pp 11007–11017, <https://doi.org/10.1109/CVPR52688.2022.01074>
- Gong K, Zhang J, Feng J (2021) Poseaug: a differentiable pose augmentation framework for 3d human pose estimation. In: IEEE conference on computer vision and pattern recognition. IEEE
- Guan S, Lu H, Zhu L et al (2023) Posegu: 3d human pose estimation with novel human pose generator and unbiased learning. *Comput Vision Image Underst*. <https://doi.org/10.1016/j.cviu.2023.103715>
- Guo Y, Ma L, Li Z et al (2021) Monocular 3d multi-person pose estimation via predicting factorized correction factors. *Comput Vision Image Underst*. <https://doi.org/10.1016/j.cviu.2021.103278>
- Guzov V, Mir A, Sattler T, et al (2021) Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4318–4329
- Han C, Yu X, Gao C et al (2022) Single image based 3d human pose estimation via uncertainty learning. *Pattern Recognit*. <https://doi.org/10.1016/j.patcog.2022.108934>
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9:1735–1780
- Holmquist K, Wandt B (2023) Diffpose: Multi-hypothesis human pose estimation using diffusion models. In: Proceedings of the IEEE international conference on computer vision. Institute of Electrical and Electronics Engineers Inc., pp 15931–15941, <https://doi.org/10.1109/ICCV51070.2023.01464>
- Honari S, Constantin V, Rhodin H et al (2023) Temporal representation learning on monocular videos for 3d human pose estimation. *IEEE Trans Pattern Anal Mach Intell* 45:6415–6427. <https://doi.org/10.1109/TPAMI.2022.3215307>
- Huo R, Gao Q, Qi J, et al (2023) 3d human pose estimation in video for human-computer/robot interaction. In: Yang H, Liu H, Zou J, et al (eds) International conference on intelligent robotics and applications, vol 14273. Springer Nature Singapore, pp 176–187, <https://doi.org/10.1007/978-981-99-6498-7>

- Ionescu C, Papava D, Olaru V et al (2014) Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans Pattern Anal Mach Intell* 36:1325–1339. <https://doi.org/10.1109/TPAMI.2013.248>
- Jiang B, Hu L, Xia S (2023) Probabilistic triangulation for uncalibrated multi-view 3d human pose estimation. In: Proceedings of the IEEE international conference on computer vision. Institute of Electrical and Electronics Engineers Inc., pp 14804–14814. <https://doi.org/10.1109/ICCV51070.2023.01364>
- Jin L, Xu C, Wang X, et al (2022) Single-stage is enough: Multi-person absolute 3d pose estimation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2022-June. IEEE Computer Society, pp 13076–13085. <https://doi.org/10.1109/CVPR52688.2022.01274>
- Joo H, Liu H, Tan L, et al. (2015) Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE international conference on computer vision, pp 3334–3342, URL <http://www.cs.cmu.edu/>
- Kang Y, Liu Y, Yao A, et al (2023) 3d human pose lifting with grid convolution. In: Proceedings of the AAAI conference on artificial intelligence, pp 1105–1113
- Kazemi V, Burenies M, Azizpour H, et al (2013) Multi-view body part recognition with random forests. In: 2013 24th British machine vision conference, BMVC2013. British Machine Vision Association
- Kim HW, Lee GH, Nam WJ et al (2024) Mhcanonnet: multi-hypothesis canonical lifting network for self-supervised 3d human pose estimation in the wild video. *Pattern Recognit.* <https://doi.org/10.1016/j.patcog.2023.109908>
- Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: International conference on learning representations (ICLR'2017)
- Kundu JN, Seth S, Jamkhandi A, et al (2021) Non-local latent relation distillation for self-adaptive 3d human pose estimation. In: Annual conference on neural information processing systems. MIT Press
- Kundu JN, Seth S, Jampani V, et al (2020a) 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, pp 6152–6162
- Kundu JN, Seth S, V RM, et al (2020b) Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. In: Proceedings of the AAAI conference on artificial intelligence, pp 11312–11319
- Kundu JN, Seth S, Ym P, et al (2022) Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2022-June. IEEE Computer Society, pp 20416–20427. <https://doi.org/10.1109/CVPR52688.2022.01980>
- Lamas A, Tabik S, Montes AC et al (2022) Human pose estimation for mitigating false negatives in weapon detection in video-surveillance. *Neurocomputing* 489:488–503. <https://doi.org/10.1016/j.neucom.2021.12.059>
- Lee K, Kim W, Lee S (2023) From human pose similarity metric to 3d human pose estimator: temporal propagating lstm networks. *IEEE Trans Pattern Anal Mach Intell* 45:1781–1797. <https://doi.org/10.1109/TPAMI.2022.3164344>
- Li S, Ke L, Pratama K, et al (2020a) Cascaded deep monocular 3d human pose estimation with evolutionary training data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6173–6183
- Li Y, Li K, Jiang S, et al (2020b) Geometry-driven self-supervised method for 3d human pose estimation. In: Proceedings of the AAAI conference on artificial intelligence, pp 11442–11449
- Li W, Liu H, Tang H et al (2023) Multi-hypothesis representation learning for transformer-based 3d human pose estimation. *Pattern Recognit.* <https://doi.org/10.1016/j.patcog.2023.109631>
- Li W, Liu H, Tang H, et al (2022) Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2022-June. IEEE Computer Society, pp 13137–13146. <https://doi.org/10.1109/CVPR52688.2022.01280>
- Lin TY, Maire M, Belongie S, et al (2014) Microsoft coco: common objects in context. In: Computer vision-ECCV 2014: 13th European conference, pp 740–755
- Lin J, Lee GH (2021) Multi-view multi-person 3d pose estimation with plane sweep stereo. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11886–11895
- Li H, Pun CM (2023) Cee-net: Complementary end-to-end network for 3d human pose generation and estimation. In: Proceedings of the AAAI conference on artificial intelligence, pp 1305–1313
- Li H, Shi B, Dai W, et al (2023a) Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In: Proceedings of the AAAI conference on artificial intelligence, pp 1296–1304
- Liu J, Ding H, Shahroudy A et al (2020) Feature boosting network for 3d pose estimation. *IEEE Trans Pattern Anal Mach Intell* 42:494–501. <https://doi.org/10.1109/TPAMI.2019.2894422>

- Liu R, Shen J, Wang H et al (2021) Enhanced 3d human pose estimation from videos by using attention-based neural network with dilated convolutions. *Int J Comput Vision* 129:1596–1615. <https://doi.org/10.1007/s11263-021-01436-0>
- Liu W, Bao Q, Sun Y et al (2022) Recent advances in monocular 2d and 3d human pose estimation: a deep learning perspective. *ACM Comput Surv* 55:1–41
- Liu W, Lin X, Chen X et al (2023) Vision-based estimation of mds-updrs scores for quantifying Parkinson's disease tremor severity. *Med Image Anal*. <https://doi.org/10.1016/j.media.2023.102754>
- Liu R, Shen J, Wang H, et al (2020b) Attention mechanism exploits temporal contexts: real-time 3d human pose reconstruction. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 5064–5073
- Loper M, Mahmood N, Romero J et al (2015) Smpl: a skinned multi-person linear model. *ACM Trans Graph (TOG)* 34:1–16
- Luvizon DC, Tabia H, Picard D (2023) Ssp-net: scalable sequential pyramid networks for real-time 3d human pose regression. *Pattern Recognit*. <https://doi.org/10.1016/j.patcog.2023.109714>
- Luvizon DC, Picard D, Tabia H (2022) Consensus-based optimization for 3d human pose estimation in camera coordinates. *Int J Comput Vision* 130:869–882. <https://doi.org/10.1007/s11263-021-01570-9>
- Marcad TV, Henschel R, Black MJ, et al (2018) Recovering accurate 3d human pose in the wild using imus and a moving camera. In: *Proceedings of the European conference on computer vision (ECCV)*, pp 601–617, URL <http://virtualhumans.mpi-inf.mpg.de/3DPW>
- Ma X, Su J, Wang C, et al (2021) Context modeling in 3d human pose estimation: A unified perspective. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 6238–6247
- Matsukawa T, Okabe T, Suzuki E et al (2020) Hierarchical gaussian descriptors with application to person re-identification. *IEEE Trans Pattern Anal Mach Intell* 42:2179–2194. <https://doi.org/10.1109/TPAMI.2019.2914686>
- Mehta D, Rhodin H, Casas D, et al (2017) Monocular 3d human pose estimation in the wild using improved cnn supervision. In: *International conference on 3D vision (3DV)*, IEEE, pp 506–516
- Mehta D, Sotnychenko O, Mueller F, et al (2018) Single-shot multi-person 3d pose estimation from monocular rgb. In: *International conference on 3D vision (3DV)*, pp 120–130
- Mildenhall B, Srinivasan PP, Tancik M, et al. (2020) Nerf: Representing scenes as neural radiance fields for view synthesis. In: *European conference on computer vision (ECCV)*, pp 405–421
- Mitra R, Gundavarapu NB, Sharma A, et al. (2020) Multiview-consistent semi-supervised learning for 3d human pose estimation. In: *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp 6907–6916
- Moon G, Chang JY, Lee KM (2019) 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*, pp 10133–10142
- Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. *Computer Vision-ECCV 2016: 14th European conference. The Netherlands, Amsterdam*, pp 483–499
- Nie Q, Liu Z, Liu Y (2023) Lifting 2d human pose to 3d with domain adapted 3d body concept. *Int J Comput Vision* 131:1250–1268. <https://doi.org/10.1007/s11263-023-01749-2>
- Park S, You E, Lee I, et al (2023) Towards robust and smooth 3d multi-person pose estimation from monocular videos in the wild. In: *Proceedings of the IEEE international conference on computer vision. Institute of Electrical and Electronics Engineers Inc.*, pp 14726–14736, <https://doi.org/10.1109/ICCV51070.2023.01357>
- Pavlakos G, Zhou X, Derpanis KG, et al (2017) Coarse-to-fine volumetric prediction for single-image 3d human pose. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 7025–7034
- Pavlo D, Feichtenhofer C, Grangier D, et al (2019) 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*, pp 7753–7762
- Peng S, Geng C, Zhang Y et al (2023) Implicit neural representations with structured latent codes for human body modeling. *IEEE Trans Pattern Anal Mach Intell* 45:9895–9907. <https://doi.org/10.1109/TPAMI.2023.3245815>
- Peng S, Zhang Y, Xu Y, et al (2021) 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*, pp 9054–9063
- Qiu Z, Qiu K, Fu J et al (2023) Weakly-supervised pre-training for 3d human pose estimation via perspective knowledge. *Pattern Recognit*. <https://doi.org/10.1016/j.patcog.2023.109497>

- Rajasegaran J, Pavlakos G, Kanazawa A, et al (2023) On the benefits of 3d pose and tracking for human action recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2023-June. IEEE Computer Society, pp 640–649, <https://doi.org/10.1109/CVPR.2023.00069>
- Reddy ND, Guigues L, Pishchulin L, et al (2021) Tesseract: end-to-end learnable multi-person articulated 3d pose tracking. In: IEEE conference on computer vision and pattern recognition. IEEE
- Remelli E, Han S, Honari S, et al (2020) Lightweight multi-view 3d pose estimation through camera-disentangled representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6040–6049
- Rogez G, Weinzaepfel P, Schmid C (2020) Lcr-net++: multi-person 2d and 3d pose detection in natural images. *IEEE Trans Pattern Anal Mach Intell* 42:1146–1161. <https://doi.org/10.1109/TPAMI.2019.2892985>
- Sarafianos N, Boteanu B, Ionescu B et al (2016) 3d human pose estimation: a review of the literature and analysis of covariates. *Comput Vis Image Underst* 152:1–20. <https://doi.org/10.1016/j.cviu.2016.09.002>
- Shan W, Liu Z, Zhang X, et al (2023) Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In: Proceedings of the IEEE international conference on computer vision. Institute of Electrical and Electronics Engineers Inc., pp 14715–14725, <https://doi.org/10.1109/ICCV51070.2023.01356>
- Shuai H, Wu L, Liu Q (2023) Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Trans Pattern Anal Mach Intell* 45:4122–4135. <https://doi.org/10.1109/TPAMI.2022.3188716>
- Sigal L, Balan AO, Black MJ (2010) Humaneva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int J Comput Vision* 87:4–27. <https://doi.org/10.1007/s11263-009-0273-6>
- Su SY, Yu F, Zollhöfer M, et al (2021) A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In: Advances in neural information processing systems, pp 12278–12291
- Sun K, Xiao B, Liu D, et al (2019) Deep high-resolution representation learning for human pose estimation. In: Proceeding of the IEEE/CVF conference on computer vision and pattern recognition, pp 5693–5703
- Sun J, Zhou L, Geng B, et al (2024) Leg state estimation for quadruped robot by using probabilistic model with proprioceptive feedback. *IEEE/ASME transactions on mechatronics*
- Tang Z, Qiu Z, Hao Y, et al (2023) 3d human pose estimation with spatio-temporal criss-cross attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition., Institute of Electrical and Electronics Engineers (IEEE), pp 4790–4799, <https://doi.org/10.1109/cvpr52729.2023.00464>
- Tian Y, Zhang H, Liu Y et al (2023) Recovering 3d human mesh from monocular images: a survey. *IEEE transactions on pattern analysis and machine intelligence*. <https://doi.org/10.1109/TPAMI.2023.3298850>
- Trumble M, Gilbert A, Malleon C, et al (2017) Total capture: 3d human pose estimation fusing video and inertial sensors. In: Proceedings of 28th British machine vision conference, URL <http://cvssp.org/data/totalcapture/>
- Tu H, Wang C, Zeng W (2020) Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: In computer vision-ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, pp 197–212
- Tuyls K, Omidshafiei S, Muller P, et al (2021) Game plan: what ai can do for football, and what football can do for ai
- Umar I, Pavlo M, Jan K (2020) Weakly-supervised 3d human pose learning via multi-view images in the wild. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp 5243–5252
- Usman B, Tagliasacchi A, Saenko K, et al (2022) Metapose: Fast 3d pose from multiple views without 3d supervision. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2022-June. IEEE Computer Society, pp 6749–6760, <https://doi.org/10.1109/CVPR52.688.2022.00664>
- Varol G, Romero J, Martin X, et al (2017) Learning from synthetic humans. In: Proceeding of the IEEE conference on computer vision and pattern recognition, pp 109–117
- Vaswani A, Brain G, Shazeer N, et al (2017) Attention is all you need. *Advances in neural information processing systems* 30
- Wan X, Chen Z, Zhao X (2023) View consistency aware holistic triangulation for 3d human pose estimation. *Comput Vis Image Underst*. <https://doi.org/10.1016/j.cviu.2023.103830>
- Wandt B, Little JJ, Rhodin H (2022) Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2022-June. IEEE Computer Society, pp 6625–6635, <https://doi.org/10.1109/CVPR52688.2022.00652>
- Wandt B, Rudolph M, Zell P, et al (2021) Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13294–13304
- Wang K, Lin L, Jiang C et al (2020) 3d human pose machines with self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 42:1069–1082. <https://doi.org/10.1109/TPAMI.2019.2892452>

- Wang K, Boonpratotong A, Chen W et al (2023) The fundamental property of human leg during walking: linearity and nonlinearity. *IEEE Trans Neural Syst Rehabil Eng* 31:4871–4881
- Wang J, Luvizon D, Xu W, et al (2023a) Scene-aware egocentric 3d human pose estimation. In: *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp 13031–13040
- Wang Z, Nie X, Qu X, et al (2022) Distribution-aware single-stage models for multi-person 3d pose estimation. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2022-June. IEEE computer society, pp 13086–13095, <https://doi.org/10.1109/CVPR52688.2022.01275>
- Wang J, Tan S, Zhen X et al (2021) Deep 3d human pose estimation: a review. *Comput Vis Image Underst*. <https://doi.org/10.1016/j.cviu.2021.103225>
- Wang D, Xuan S, Zhang S (2024) Locllm: Exploiting generalizable human keypoint localization via large language model. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 614–623
- Wang T, Zhang J, Cai Y, et al (2021b) Direct multi-view multi-person 3d pose estimation. In: *Advances in neural information processing systems*, pp 13153–13164
- Wehrbein T, Rudolph M, Rosenhahn B, et al (2021) Probabilistic monocular 3d human pose estimation with normalizing flows. In: *IEEE/CVF International conference on computer vision*. Institute of Electrical and Electronics Engineers (IEEE), pp 11179–11188, <https://doi.org/10.1109/iccv48922.2021.01101>
- Wei WL, Lin JC, Liu TL, et al (2022) Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2022-June. IEEE Computer Society, pp 13201–13210, <https://doi.org/10.1109/CVPR52688.2022.01286>
- Wu S, Jin S, Liu W, et al (2021) Graph-based 3d multi-person pose estimation using multi-view images. In: *IEEE/CVF international conference on computer vision*. Institute of Electrical and Electronics Engineers (IEEE), pp 11128–11137, <https://doi.org/10.1109/iccv48922.2021.01096>
- Wu H, Xiao B (2020) 3d human pose estimation via explicit compositional depth maps. In: *Proceedings of the AAAI conference on artificial intelligence*, pp 12378–12385
- Xu Y, Wang W, Liu T et al (2022) Monocular 3d pose estimation via pose grammar and data augmentation. *IEEE Trans Pattern Anal Mach Intell* 44:6327–6344. <https://doi.org/10.1109/TPAMI.2021.3087695>
- Xu C, Chen S, Li M, et al (2021) Invariant teacher and equivariant student for unsupervised 3d human pose estimation. In: *AAAI conference on artificial intelligence*, pp 3013–3021
- Xu T, Takano W (2021) Graph stacked hourglass networks for 3d human pose estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 16105–16114
- Xu J, Yu Z, Ni B, et al (2020) Deep kinematics analysis for monocular 3d human pose estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 899–908
- Yan J, Zhou M, Pan J, et al (2022) Recent advances in 3d human pose estimation: From optimization to implementation and beyond. *International Journal of Pattern Recognition and Artificial Intelligence* 36
- Ye H, Zhu W, Wang C, et al (2022) Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In: *European conference on computer vision*, Springer, pp 142–159
- Yi X, Zhou Y, Habermann M, et al (2022) Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 13167–13178
- Yu BX, Zhang Z, Liu Y, et al (2023) Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In: *Proceedings of the IEEE international conference on computer vision*. Institute of Electrical and Electronics Engineers Inc., pp 8784–8795, <https://doi.org/10.1109/ICCV51070.2023.00810>
- Yu Z, Ni B, Xu J, et al (2021) Towards alleviating the modeling ambiguity of unsupervised monocular 3d human pose estimation. In: *Proceedings of the IEEE international conference on computer vision*. Institute of Electrical and Electronics Engineers Inc., pp 8631–8640, <https://doi.org/10.1109/ICCV48922.2021.00853>
- Zeng A, Sun X, Yang L, et al (2021) Learning skeletal graph neural networks for hard 3d pose estimation. In: *Proceedings of the IEEE international conference on computer vision*. Institute of Electrical and Electronics Engineers Inc., pp 11416–11425, <https://doi.org/10.1109/ICCV48922.2021.01124>
- Zhai K, Nie Q, Ouyang B, et al (2023) Hopfir: Hop-wise graphformer with intragroup joint refinement for 3d human pose estimation. In: *Proceedings of the IEEE international conference on computer vision*. Institute of Electrical and Electronics Engineers Inc., pp 14939–14949, <https://doi.org/10.1109/ICCV51070.2023.01376>
- Zhang Z (2012) Microsoft kinect sensor and its effect. *IEEE Multimedia* 19:4–10. URL www.microsoft.com/
- Zhang Z, Wang C, Qiu W et al (2021) Adafuse: adaptive multiview fusion for accurate human pose estimation in the wild. *Int J Comput Vis* 129:703–718. <https://doi.org/10.1007/s11263-020-01398-9>

- Zhang J, Gong K, Wang X et al (2023) Learning to augment poses for 3d human pose estimation in images and videos. *IEEE Trans Pattern Anal Mach Intell* 45:10012–10026. <https://doi.org/10.1109/TPAMI.2023.3243400>
- Zhang S, Wang C, Nie L et al (2023) Learning enriched hop-aware correlation for robust 3d human pose estimation. *Int J Comput Vision* 131:1566–1583. <https://doi.org/10.1007/s11263-023-01770-5>
- Zhang Y, Wang C, Wang X et al (2023) Voxeltrack: multi-person 3d human pose estimation and tracking in the wild. *IEEE Trans Pattern Anal Mach Intell* 45:2613–2626. <https://doi.org/10.1109/TPAMI.2022.3163709>
- Zhang J, Nie X, Feng J (2020) Inference stage optimization for cross-scenario 3d human pose estimation. In: *Advances in neural information processing systems*, pp 2408–2419
- Zhang J, Tu Z, Yang J, et al (2022) Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2022-June. IEEE Computer Society, pp 13222–13232, <https://doi.org/10.1109/CVPR52688.2022.01288>
- Zhang L, Zhou K, Lu F, et al (2024) Deep semantic graph transformer for multi-view 3d human pose estimation. In: *AAAI conference on artificial intelligence*, pp 7205–7214
- Zhan Y, Li F, Weng R, et al (2022) Ray3d: Ray-based 3d human pose estimation for monocular absolute 3d localization. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2022-June. IEEE Computer Society, pp 13106–13115, <https://doi.org/10.1109/CVPR52688.2022.01277>
- Zhao W, Wang W, Tian Y (2022) Graformer: Graph-oriented transformer for 3d pose estimation. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2022-June. IEEE Computer Society, pp 20406–20415, <https://doi.org/10.1109/CVPR52688.2022.01979>
- Zhao J, Yu T, An L, et al (2023a) Triangulation residual loss for data-efficient 3d pose estimation. In: *Advances in neural information processing systems*
- Zhao Q, Zheng C, Liu M, et al. (2023b) Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2023-June. IEEE Computer Society, pp 8877–8886, <https://doi.org/10.1109/CVPR52729.2023.00857>
- Zheng C, Wu W, Chen C et al (2023) Deep learning-based human pose estimation: a survey. *ACM Comput Surv* 56:1–37
- Zheng C, Zhu S, Mendieta M, et al. (2021) 3d human pose estimation with spatial and temporal transformers. In: *Proceedings of the IEEE international conference on computer vision*. Institute of Electrical and Electronics Engineers Inc., pp 11636–11645, <https://doi.org/10.1109/ICCV48922.2021.01145>
- Zhou T, Yang Y, Wang W (2023) Differentiable multi-granularity human parsing. *IEEE Trans Pattern Anal Mach Intell* 45(7):8296–8310
- Zhou L, Chen Y, Wang J (2024) Slowfastformer for 3d human pose estimation. *Comput Vis Image Underst*. <https://doi.org/10.1016/j.cviu.2024.103992>
- Zhou X, Huang Q, Sun X, et al (2017) Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: *Proceedings of the IEEE international conference on computer vision*, pp 398–407
- Zhou T, Wang W, Liu S, et al (2021) Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 1622–1631
- Zhou Y, Xie J, Zhang X, et al. (2024c) Energy-efficient and interpretable multisensor human activity recognition via deep fused lasso net. *IEEE transactions on emerging topics in computational intelligence*
- Zhou F, Yin J, Li P (2024a) Lifting by image - leveraging image cues for accurate 3d human pose estimation. In: *AAAI conference on artificial intelligence*, pp 7632–7640
- Zhu Y, Xu X, Shen F, et al (2021) Posegtac: Graph transformer encoder-decoder with atrous convolution for 3d human pose estimation. In: *international joint conference on artificial intelligence*. Morgan Kaufmann
- Zou Z, Tang W (2021) Modulated graph convolutional network for 3d human pose estimation. In: *Proceedings of the IEEE international conference on computer vision*. Institute of Electrical and Electronics Engineers Inc., pp 11457–11467, <https://doi.org/10.1109/ICCV48922.2021.01128>